A Qualitative Evaluation Criterion for Human-Robot Interaction System in Achieving Collective Tasks

Jamil Abou Saleh, Fakhreddine Karray, and Michael Morckos
Center of Pattern Analysis and Machine Intelligence, University of Waterloo, Canada
{jaabousa, karray, mmorckos}@uwaterloo.ca

Abstract—This work intends to identify common performance metrics for task-oriented human-robot interaction. We present a methodology to assess the system performance of a human-robot team in achievement of collective tasks. We propose a systematic approach that addresses the performance of both the human user and the robotic agent as a team. Toward this end, we attempt to determine the true time that an operator has to dedicate to a robot in action. We define the robot attention demand (RAD) as a function of both direct interaction time (DIT) and indirect interaction time (IIT), where the IIT is a direct consequence of the human trust in automation. We propose a two-level fuzzy temporal model to evaluate the human trust in automation while interacting with robots. Another fuzzy temporal model is presented to evaluate the human reliability during interaction time. The model is then generalized to accommodate multi-robot scenarios. Sequential and parallel robot cooperation schemes with varying levels of task dependency are considered. The fuzzy knowledge bases are further updated by implementing an application robotic platform where robots and users interact naturally to complete tasks with varying levels of complexity. User feedback is noted and used to tune the knowledge base rules where needed, to better represent a human expert’s knowledge.

Keywords: performance metrics, fuzzy logic, finite automata

I. INTRODUCTION

With a multitude of developing scenarios of how humans and robots can simultaneously collaborate as a team, it becomes instrumental to assess the performance of such teams. Toward this end, performance evaluation metrics are used as important measures to achieve these goals. Mission effectiveness is one of the most popular and widely used metrics to evaluate the performance of human-robot teams. It is a measure of how well a task is actually performed [2]. Although successful deployment of this performance metric can be found in many scenarios and applications, such measures face several crucial problems. A robot might appear to be effective at a certain time, but might be making negative progress on the overall goal [10]. Situation awareness also finds itself as another emerging metric used in the literature to assess the system performance [3]; however, it requires task-specific metrics to be designed in order to assess the performance of each of its objectives. Schreckenghost [18] introduced the concept of a work efficiency index (WEI), which represents the ratio of productive time to overhead time that occurs during interaction time. Kannan [8] defined a metric for calculating how useful fault tolerance is for a multi-robot team by calculating the redundancy of a system. Shah [19] examined the productivity of a human-robot team (HRT) through the mean time between interventions (MTBI), mean time completing an intervention (MTCI), and the probability that an intervention would be needed.

The idea of developing a common toolkit of performance metrics has also been discussed by many researchers. Olsen and Goodrich discuss six interrelated performance metrics that can lead the design of human-robot interaction systems [10], [6]. The most instrumental ones are the robot attention demand (RAD), and the fan-out (FO). Robot attention demand is a measure of the fraction of total task time that a user must spend interacting with a robot. RAD is defined as shown in equation 1, where IE represents the interaction effort and NT the neglect tolerance.

\[
RAD = \frac{IE}{IE + NT} 
\]  

(1)

Fan-out, on the other hand, is a measure of how many robots a user can interact with simultaneously and effectively. FO is defined as shown in equation 2.

\[
FO = \frac{1.0}{RAD} = \frac{IE + NT}{IE} 
\]  

(2)

Doing so, Olsen suggests to approximate the value of the neglect tolerance by measuring the time between the human instruction and either a drop of the robot performance below the effectiveness threshold, or the intervention of the human with another instruction. Olsen, however, states that in this scenario, this metric is no more independent from the user, and hence the operator’s trust in the robot’s autonomous abilities becomes a critical issue - a factor that has been extensively discussed in the literature by many researchers, especially in the fields of human-machine interaction [11], [9], [5], [7]. Olsen and Goodrich also suggest to experimentally measure FO, and use that value to approximate the IE measure. However, there are several limitations that make it difficult to measure this fan-out limit. The first limitation is caused by task saturation. In this scenario, the task space becomes either too crowded, or saturated, making it hard to measure this FO factor experimentally. The second critical limitation is based on the human cognitive and physical limitations. When the task space becomes crowded, human reliability becomes a serious issue. Several studies show that human factors are responsible for 20% to 90% of the failures in many systems [4], [17], [21], [1].

Therefore, in other related work, we attempt to address these limitations. We propose a framework that aims to determine the true time that an operator has to dedicate to the robot. An alternative definition of the RAD as a function of both
some simulations and experimental results are presented in mathematical models. Experimental system setup along with II. Section III briefly describes our proposed generalized multi-robot human interaction systems [16].

reliability fuzzy temporal models are described in section the framework to propose a generic performance metric for multi-robot human interaction systems [16].

\[ RAD = DIT + IIT = DIT + NT \times (1 - Tr) \]  

(3)

We also propose an alternative definition of the FO metric as shown in equation 4, where HR (a value between 0 and 1) represents the human operator’s ability to manage the increased task complexity [15]. Then, we further augment the framework to propose a generic performance metric for multi-robot human interaction systems [16].

\[ FO = \frac{1.0}{RAD} \times HR \]  

(4)

In this work, we further extend our previous findings by implementing an application robotic platform where robots and users interact via natural language to complete tasks with varying levels of complexity and success. Users feedback is noted and used to tune some rules in the proposed fuzzy knowledge base. The remainder of the paper is organized as follows: the proposed human trust in automation and human reliability fuzzy temporal models are described in section II. Section III briefly describes our proposed generalized mathematical models. Experimental system setup along with some simulations and experimental results are presented in Section IV. Finally, section V concludes this paper.

II. PROPOSED FUZZY TEMPORAL MODELS

A. Human Trust in Automation Fuzzy Temporal Model

![Fig. 1. Overall Architecture of the Proposed Trust Model](image)

Previous discussions on trust and its evolution, and the attempts to quantify this instrumental phenomenon in human-robot interaction teams, make it clear that coming with a unique mathematical formula that governs its temporal behaviour is going to be unrealistic, domain and application specific. Trust phenomenon is fuzzy by nature, and determining it depends on many factors, including time and previous state. Modelling this phenomenon, however, using only fuzzy logic might not give the anticipated results, and therefore, a state machine is required. A finite fuzzy state machine (FFSM) that combines both advantages of fuzzy logic and state machine is proposed. The larger the number of inputs to this model, however, the more complicated the knowledge base will be. Consequently, identifying the major factors which contribute largely to building up trust becomes crucial. Thus, we propose a two-level framework, shown in figure 1, to identify and estimate the level of trust. Level II is implemented using a finite fuzzy state machine that takes three inputs (the second-order perceptions): fault size, productivity, and awareness, and outputs the corresponding level of trust based on the previous state of trust and the current perceived inputs (inferred from level I). Fault size, productivity, and awareness are modelled using three membership functions: low, medium, and high. Five states are used to model trust (very low, low, medium, high, and very high); thus, 3*3*3*5 = 135 rules are needed to represent its knowledge base. The output function maps the trust states (very low, low, medium, high, very high) to the following zero-order consequences \( (0.1, 0.3, 0.5, 0.7, 0.9) \). The crisp overall consequence is then generated by aggregating the qualified crisp output of each rule using the weighted average method.

Each of the second-order perceptions is fuzzy in nature. Those factors are not temporal, thus no FFSM is needed. As a result, a fuzzy Mamdani inferencing model is used for this purpose. Three membership functions are chosen to represent each of those factors: low, medium, and high. Fault size is thought to be highly determined by three lower-order factors: fault frequency (FF), fault cruciality (FC), and the ability of the robot to recover from its faults (FR). Each is represented by three membership functions; therefore, 27 rules are needed. The same applies to awareness. Three inputs mainly determine the value of this factor: machine awareness of its capabilities (MA), context awareness of the task (CA), and machine awareness of the human operator's availability and cognitive and physical abilities and limitations (HA). Another 27 rules are needed to represent this model. As for productivity, two inputs mainly determine the value of this factor: task/goal successfulness and completion (TC), and task complexity and sophistication (TS); 9 rules are used to represent the knowledge base of this model.

The two-level architecture to estimate the trust factor is very important as it dramatically decreases the complexity of the system and the size of the knowledge base. If only a FFSM were used to model this trust factor, \( 5 \times 3^5 = 32,805 \) rules would have been needed, which is impractical and almost impossible to accurately implement. Further details can be found in [14].

B. Human Reliability Fuzzy Temporal Model

Previous discussions on human reliability (HR) demonstrate the crucial importance of modelling this factor. With respect to this work, we believe that modelling human reliability should be also fuzzy. Since this factor is highly related to time and to its previous state, a FFSM is then
required. The same description presented for level II in trust modelling is used for modelling the human reliability. Many factors affect humans and their performance; some are psychological, physiological, or sociological, among others. However, we only focus in this work on those that arise from interacting directly with the robot, and future work will put more focus toward better modelling of this somewhat complex phenomenon. In this context, three input variables are considered: (1) number of sub-tasks being simultaneously perceived (NS), (2) mental workload required during task completion (MW), and (3) external/internal burden (EB).

Three membership functions are used to model each variable (low, medium, and high), and human reliability is modelled using 5 states (very low, low, medium, high, very high). However, an important factor that one shall keep in mind is time. Although the above three variables play remarkable role in determining the human reliability, interaction time (or time while being active) is also crucial, since the human performance degrades with time even when the task is simple and does not impose much physical and cognitive load. Therefore, the states (very low, low, medium, high, very high) are mapped to the following consequences 

\[ 0.1 - a \times t, 0.3 - a \times t, 0.5 - a \times t, 0.7 - a \times t, 0.9 - a \times t \]

where \( a \) is a subjective constant that varies from one person to another and represents the natural human degradation factor, and \( t \) represents the time factor. Weighted average method is then used to generate the crisp final consequence. Note that we chose to model the natural degradation linearly; however, exponential modelling can be also used, but since a small value of \( a \) is usually the case, both linear and exponential representations would behave similarly, and hence the linear form is chosen for its simplicity. Further details can be found in [15].

III. METRICS GENERALIZATION MODELS FOR MULTI-ROBOT SYSTEMS

The fan-out metric, as introduced so far, is a performance criterion to assess the performance of the human-robot system, where one robotic agent is involved. However, an important issue ignored by Olsen and Goodrich, among other researchers, is a scenario where the robotic cognitive system is composed of more than one robot. Hence, a special consideration of multi-robot systems should be also addressed, where the performance metric should be an indication of how well the overall multi-system is performing. Several cases, however, emerge from this fact, as these robots can have sequential or parallel ways of executing their tasks, and with different levels of dependency; therefore, each case is considered individually, and a proper generalization model of our proposed metric is addressed. In the following, we briefly discuss this manner, and a more detailed description can be found in [16].

A. Sequential Execution of Tasks

In this scenario, \( N \) robots are cooperating with the guidance of a human user to finish a specific task. In doing so, only one robot is active at one point, while the other robots are idle. Therefore, the system FO is that of the active robot as shown in equation 5.

\[ \text{SystemFO} = \text{FO}_1 \lor \text{FO}_2 \lor \ldots \lor \text{FO}_N \]  

where \( \text{FO}(_{\text{idlerobot}}) = 0 \).

B. Parallel Execution of Tasks

1) Independent Execution of Tasks: In this scenario, the robots are active at the same time executing independent tasks in parallel, thus dependency-related issues can be ignored. Usually, in most cases the average of the \( N \) robots FO is indicative of the system FO, but since robots do not contribute equally to the overall task completion, a weighted average method is applied as shown in equation 6, where \( W_i \) is the percent contribution of the \( i_{\text{th}} \) robot toward the final goal completion.

\[ \text{SystemFO} = \sum_{i=1}^{N} W_i \text{FO}_i, \text{where} \sum_{i=1}^{N} W_i = 1 \]  

2) Dependent Execution of Tasks: This scenario is the most complicated. First, we consider two robots that are active at the same time executing dependent tasks in parallel. When a total (100\%) task dependency is encountered, the system FO is forced to follow the smallest robot FO (the one with the poorest performance), and the less dependency exists, the closer the system FO will be to the weighted average value (the task independent scenario). Therefore, the system FO is modelled as shown in equation 7, where \( d \), a value between 0 and 1, represents the task dependency between robots one and two.

\[ \text{SystemFO} = d \times \min(\text{FO}_1, \text{FO}_2) + (1 - d) \times \sum_{i=1}^{N} W_i \text{FO}_i \]  

Generalizing this conclusion for N-robots, inter-task dependencies should be considered. For example, when three robots are involved, there exist task dependencies between robots 1 and 2, robots 1 and 3, and robots 2 and 3. The system FO is shown in equation 8, where \( d_{i,j} \) is the task dependency between robots \( i \) and \( j \), and \( \text{FO}_i \) is the fan-out of the \( i_{\text{th}} \) robot. Equation 8 consists of two parts: \( \sum_{i,j:i<j} W_i W_j / N(N-1) \) and \( (\text{DEP}_{i,j} + \text{IND}_{i,j}) \). The second part represents the relative weighted average FO of the \( i_{\text{th}} \) and \( j_{\text{th}} \) robots, while the first part weights how much contribution both robots make toward the final goal. However, when dealing with pairwise dependencies, \( \binom{N}{2} \) possibilities arise, where the sum of all contribution weights corresponding to all possibilities is equal to \( (N-1) \); hence explaining the division by \( (N-1) \).

\[ \text{SysFO} = \sum_{i,j,i<j} \frac{W_i + W_j}{N-1} \times (\text{DEP}_{i,j} + \text{IND}_{i,j}), \text{where} \]

\[ \text{DEP}_{i,j} = d_{i,j} \times \min(\text{FO}_i, \text{FO}_j) \]

\[ \text{IND}_{i,j} = (1 - d_{i,j}) \times \left( \frac{W_i}{W_i + W_j} \times \text{FO}_i + \frac{W_j}{W_i + W_j} \times \text{FO}_j \right) \]

Inter-robot task dependencies, however, can highly affect other dependencies in the system. In other words, the dependency between robot 1 and robot 2, might also have some
additional cost on the dependency between robot 2 and robot 3, and so on. Consequently, the abovementioned system FO is regarded as an upper bound for the practical system FO which is lower bounded by the smallest robot FO. This is shown in equation 9. Further details can be found in [16].

\[
\min(FO_1, FO_2, \ldots, FO_N) < \text{RealSystemFO} < \text{SysFO}
\] (9)

IV. EXPERIMENTAL SYSTEM

The knowledge bases presented in this work are based on a human expert’s knowledge, and the most recent work in the area of cognitive human-machine interaction and performance evaluation metrics. However, and in order to show the strength of our proposed system, an application robotic platform that enables man-machine interaction is implemented, and users’ feedback while interacting with the system was noted. The knowledge base was then fine-tuned to better reflect the user’s knowledge.

In the following, we briefly discuss the main software and hardware components of the robotic system, as well as the proposed platform design and implementation, followed by some experimental results.

A. Experimental Setup

The set of experiments conducted in this work involves two PeopleBot robots, working singularly or together towards independent tasks, and a human operator. PeopleBot [12] is a differential-drive mobile robot that is well known for human-robot interaction projects. PeopleBot is equipped with infrared table sensors and a gripper with sensors which allow the robot to pick up an object from one location and place it at another. PeopleBot features come with a laser navigation package for autonomous robotic navigation and localization. It also comes with pan/tilt/zoom camera that can be used for object and people recognition, color tracking, or other robot vision tasks. PeopleBot SDK package also provides some tools for creating maps of a robot’s operating environment. Verbal Interaction with PeopleBot is enabled using Vestec’s automatic speech recognition engine (VASRE) [20]. VASRE supports multiple languages, large vocabulary, and continuous speech recognition. Its acoustic models were trained based on continuous hidden Markov modelling. It is equipped with noise reduction techniques and voice detection algorithms to ensure more accurate speech recognition.

The application robotic platform supports a distributed architecture for reliable and scalable operation of clients and robot servers. The proposed distributed architecture comprises three components: the robot server, client, and the resource manager RM. The RM is the control tower of the distributed architecture. It manages one or more robot servers and coordinates communication sessions between servers and clients. Under the idle state, the robot server periodically communicates with the RM to report its status. The RM balances server loads over different machines. For example, if a robot server A is loaded with several queued commands, the RM guides the next client request to robot server B that has similar capabilities, if available. The RM keeps a queue of instructions for each robot. This allows multiple robots to be controlled simultaneously. The task queue contains sequences of tasks that need to be executed. It consists of a list of parallel tasks, where each parallel task is a collection of tasks that need to be executed in series.

B. Experimental Results

The purpose of the proposed set of experiments is to support the correctness and validity of the proposed fuzzy knowledge base, and tune rules where needed to best accommodate and represent the human expert’s knowledge. However, we only focus on the two-level human trust in automation model, along with its proposed knowledge base. The human reliability factor is not further addressed throughout this work as it requires future in-depth studies in other disciplines. Nine users were chosen for this purpose, and each was exposed to a set of five to six scenarios where the robot attempts to complete a set of different tasks, with varying levels of success, under the command and operation of the human user. Human trust in automation, along with other first- and second-order perceptions, are marked at different time units and compared to those obtained/inferred using our proposed framework.

The tasks vary from simple to more complex. In some scenarios, the robot is instructed to perform a series of simple tasks of moving a certain distance forward or backward, turning left or right at a certain angle, and/or controlling its gripper. More complex tasks require the robot to pick an object from a certain location and place it at a goal location. In doing so, the robot has to navigate the environment, avoiding static and dynamic obstacles, etc. Users’ perceptions were helpful to enrich the expert’s knowledge base. Several rules were tuned after receiving feedback from users. Some rules belonging to the productivity knowledge base were tuned when feedback showed that users tend to give more weight toward task completion than task complexity and sophistication. Figure 2 shows a sample simulation for one user, and the implication of this rule tuning on the overall percent estimation error between the practical feedback (received from users) and

![Fig. 2. Percent Productivity Estimation Error: Before and After Rules Tuning (Sample User)](image-url)
Fig. 3. Subject #4 - Levels I and II Inferences
Fig. 4. Subject #5 - Levels I and II Inferences

(a) Fault Size Inference

(b) Awareness Inference

(c) Productivity Inference

(d) Human Trust in Automation

(e) Robot Attention Demand

(f) Fan Out
initial human trust in automation at user is asked to start interacting (starting with the same good robot performance, and successful task completion. The user’s trust in the system automation with a different set of scenarios, which focused on poor robot attention demand (RAD). Direct interaction time (DIT) performance. The user’s trust in the system automation increases, the indirect interaction time spent monitoring the robot and interfering when needed, decreases, and vice versa. Figure 3(f) shows the same variations of the human trust in automation with respect to time, along with its corresponding FO metric value. In this scenario (assuming a DIT of 25%), the ideal FO would be 4. However, since FO is not independent of the human trust in automation, when the trust is high, the practical FO is close to its ideal value, while when the trust is low and the corresponding IIT is high, FO is far below its ideal value, and closer to 1; thus, the human user is assumed to have too little time to interact with other robots.

Similar results are also shown in figure 4 where five continuous scenarios took place with varying levels of success and completion. User feedback was noted and compared to the inferred values. Table III shows the results for all the nine users selected in this work. The table shows the approximation errors for all inferred values, starting from fault size, to awareness, productivity, and finally human trust in automation. The results are very encouraging for the correctness of the knowledge base. Future work will further include more users to take part in this work, interacting with different types of other robots. Future work will also address the human reliability factor.

<table>
<thead>
<tr>
<th>TABLE I</th>
<th>PRODUCTIVITY % ERROR REDUCTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject</td>
<td>Old Rules</td>
</tr>
<tr>
<td>Subject#1</td>
<td>2.28</td>
</tr>
<tr>
<td>Subject#2</td>
<td>11.40</td>
</tr>
<tr>
<td>Subject#3</td>
<td>5.05</td>
</tr>
<tr>
<td>Subject#4</td>
<td>6.03</td>
</tr>
<tr>
<td>Subject#5</td>
<td>9.92</td>
</tr>
<tr>
<td>Subject#6</td>
<td>11.59</td>
</tr>
<tr>
<td>Subject#7</td>
<td>5.02</td>
</tr>
<tr>
<td>Subject#8</td>
<td>8.14</td>
</tr>
<tr>
<td>Subject#9</td>
<td>9.40</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE II</th>
<th>TRUST % ERROR REDUCTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Subject</td>
<td>Old Rules</td>
</tr>
<tr>
<td>Subject #1</td>
<td>6.43</td>
</tr>
<tr>
<td>Subject #2</td>
<td>6.38</td>
</tr>
<tr>
<td>Subject #3</td>
<td>7.83</td>
</tr>
<tr>
<td>Subject #4</td>
<td>5.33</td>
</tr>
<tr>
<td>Subject #5</td>
<td>6.40</td>
</tr>
<tr>
<td>Subject #6</td>
<td>10.60</td>
</tr>
<tr>
<td>Subject #7</td>
<td>8.86</td>
</tr>
<tr>
<td>Subject #8</td>
<td>5.29</td>
</tr>
<tr>
<td>Subject #9</td>
<td>15.14</td>
</tr>
</tbody>
</table>

obtained using our proposed two-level trust evaluation framework, and the practical ones obtained from one sample user. Two independent sets of three scenarios each took place. The sets are independent and separate which explains the discontinuity at time stamp $t = 3$, which represents time stamp $t = 0$ for the second set. The first set focused on good robot performance, and successful task completion. The user’s trust evolution was noted. In the second set, the user is asked to start interacting (starting with the same initial human trust in automation at $t = 0$) with the robot with a different set of scenarios, which focused on poor robot performance. The user’s trust in the system automation was also noted. Figure 3(a) shows the user’s first-order perceptions of fault frequency, fault cruciality, and fault recovery, along with the overall fault size. The latter value is compared to that obtained using our fault size fuzzy inference model. Figures 3(b) and 3(c) address the same manner for both the awareness and the productivity factors. Figure 3(d) compares the trust value as noted from the user and generated using our proposed fuzzy level II. Results show accurate trust approximation and good inferences in both levels I and II, which reflects proper and representative knowledge base design.

Figure 3(e) shows the implication of the human trust in automation factor on both indirect interaction time (IIT) and robot attention demand (RAD). Direct interaction time (DIT) is assumed to be 25% of the overall task time, during which the human user is to instruct and inform the robot about the task to be completed. Results show that when the human trust in automation increases, the indirect interaction time spent monitoring the robot and interfering when needed, decreases, and vice versa. Figure 3(f) shows the same variations of the human trust in automation with respect to time, along with its corresponding FO metric value. In this scenario (assuming a DIT of 25%), the ideal FO would be 4. However, since FO is not independent of the human trust in automation, when the trust is high, the practical FO is close to its ideal value, while when the trust is low and the corresponding IIT is high, FO is far below its ideal value, and closer to 1; thus, the human user is assumed to have too little time to interact with other robots.

Similar results are also shown in figure 4 where five continuous scenarios took place with varying levels of success and completion. User feedback was noted and compared to the inferred values. Table III shows the results for all the nine users selected in this work. The table shows the approximation errors for all inferred values, starting from fault size, to awareness, productivity, and finally human trust in automation. The results are very encouraging for the correctness of the knowledge base. Future work will further include more users to take part in this work, interacting with different types of other robots. Future work will also address the human reliability factor.

<table>
<thead>
<tr>
<th>TABLE III</th>
<th>INFERENCE % APPROXIMATION ERRORS</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Error</td>
<td>Fault Size</td>
</tr>
<tr>
<td>-----------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Subject#1</td>
<td>8.13</td>
</tr>
<tr>
<td>Subject#2</td>
<td>9.32</td>
</tr>
<tr>
<td>Subject#3</td>
<td>6.50</td>
</tr>
<tr>
<td>Subject#4</td>
<td>3.50</td>
</tr>
<tr>
<td>Subject#5</td>
<td>9.73</td>
</tr>
<tr>
<td>Subject#6</td>
<td>4.61</td>
</tr>
<tr>
<td>Subject#7</td>
<td>10.72</td>
</tr>
<tr>
<td>Subject#8</td>
<td>7.48</td>
</tr>
<tr>
<td>Subject#9</td>
<td>9.18</td>
</tr>
<tr>
<td>Avg Error</td>
<td>7.09</td>
</tr>
<tr>
<td>Std Dev</td>
<td>2.28</td>
</tr>
</tbody>
</table>

C. Computing with Words
So far in this work, the user’s first-order perceptions were received in forms of numerical values on a scale of 10. This method of expressing perceptions, however, seems to less likely to occur in the real world. The user is more
likely to express their perceptions using words, and say that the productivity was high or very high as opposed to the fact that it was 7 or 9. Humans think in relative ranges; our linguistic statements about the world include adjectives, adverbs, intensifiers and other modifiers. This method of computing is addressed by Zadeh as computing with words (CWW) [22], in which the objects of computation are words and propositions drawn from a natural language. CWW attempts to merge the fields of computational science, fuzzy logic, and natural language processing.

Supporting this theory, cognitive psychologist Eleanor Rosch, the mother of prototype theory [13], showed in a series of experiments that English speakers tend to numerically evaluate modifiers within a particular range very consistently. Such findings are often used to support the mapping of words onto values, and also to show that human cognition and categorization are based on physical human perception. Therefore, our future augmented framework suggests addressing careful design and implementation of the word interpreter (WI) module. The WI accepts linguistic first-order perceptions and maps them to numerical values that are fed to level I for fuzzification. Computing with words, however, is not an easy task, and a special consideration for linguistic hedges should be carefully taken into account. A user, for instance, may choose to describe fault recovery by somehow very high, or task completion by somewhere between low and medium or not very low, etc. Therefore, future work will focus on careful implementation of such a module to take full advantage of the power of computing with words. One can also note that human users’ perceptions can be subjective and highly related to factors such as experience, confidence, and level of expertise. This observation further motivated us to address interval fuzzy type II sets in our future related work to accommodate for such fuzziness in the proposed fuzzy sets.

V. Conclusion

The main goal of any human-robot interaction system is to increase the effectiveness of the team in accomplishing some task. Therefore, designing a performance metric that can assess this effectiveness is crucial. We believe that such evaluation criteria should focus equally on human and machine performance. We propose a two-level trust evaluation model which estimates the human trust in automation. This model combines the advantages of fuzzy logic and finite state machines to best model this phenomenon. The model reduces the system complexity and the size of the knowledge base by grouping perception into first- and second-order perceptions. A time-based human reliability assessment model that uses a fuzzy finite state machine to estimate the human reliability state is also proposed. Then, the model is extended to accommodate multi-robot scenarios. The fuzzy knowledge bases are further updated by implementing a robotic platform where robots and users interact via natural language to complete tasks with varying levels of complexity and success. User feedback is recorded and used to tune the knowledge base where needed.

REFERENCES