Evaluation of the Quality of Interaction from the robot point of view in Human-Robot Interactions

Amandine Mayima, Aurélie Clodic, Rachid Alami

Evaluation of the Quality of Interaction from the robot point of view in Human-Robot Interactions *

Amandine Mayima[1], Aurélie Clodic[1], and Rachid Alami[1]

LAAS-CNRS, Université de Toulouse, CNRS, INSA, Toulouse, France
firstname.name@laas.fr

Abstract. It is important for a service robot to be able to evaluate if an interaction with a human is going well or not. This paper presents a novel way to evaluate human-robot interactions, in the context of social interactions and collaborative tasks between a human and a robot. We propose a model allowing the robot to measure in real-time the quality of its interactions. This new information will improve its decision-making process.

Keywords: Quality of Interaction · Evaluation · HRI · Interaction session · Collaborative tasks

1 Introduction

As humans we are able to tell, to a certain extent, if a social interaction is going well or not, when chatting or executing a task. This knowledge allows us to adapt our behavior. Therefore, we want the robot to embody this kind of behavior. To this effect, it needs to measure the quality of the interaction so it can know if it is going well or not. Then, if it is not, the robot needs to identify the reason of the problem and from there, to try to find a solution to improve the interaction. Although the concept of quality of interaction is quite abstract, [12] shows that when it is measured by human observers, the inter-observer reliability of the concept is quite high.

Multiple ways to evaluate robots and interactions exist but in these methods, the evaluation is realized by humans after the interactions with the robot. Only a few frameworks try to make the robot evaluate the quality of the ongoing interaction. To endow the robot with such ability can improve its decision making process. In order to fill this gap, as a first step, we built a model and created metrics integrated in a robotic framework in order to evaluate the quality of interaction at three different levels, from the larger to the smaller: the interaction session level, the task level and the action level.

* The research leading to these results has received funding from the European Unions H2020 programme under grant agreement No. 688147, MuMMER http://mummer-project.eu/

We first present in Section 3 our representation of a human-robot social interaction in a three levels decomposition. Then we introduce two novel metrics as tools to evaluate the quality of interaction in Section 4. Finally, we give elements that need to be taken into account for this evaluation in Section 5 and conclude.

2 Related work

It is interesting to take a look at the concept of usability, used in Human-Computer Interaction (HCI). It is defined by the ISO 9241 standard as "The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use". To evaluate if those three usability issues are covered for a given system, different evaluation methods exist that are compared in [9]. These methods are based on questionnaires, interviews and discussions with users and experts that have tested/used the system. According to the used method, design, coding, testing or release of the application can be evaluated.

Then, the field of Human-Robot Interaction (HRI) grew and tried to define its own methods to evaluate robotic systems and human interactions with the robot, there are three main issues: what to evaluate, how to evaluate and why to evaluate. [20] proposes a theoretical framework inspired from HCI and user experiences. [1] lists metrics and benchmarks for human-robot interaction (some of them will be discussed below) and offer an analysis of the evaluation field. [7] proposed a set of benchmarks divided up in three categories: the robot evaluation based on safety (i.e., level of robot safety for its human users during interaction with the environment) and scalability (i.e., ability of the robot system to be used by different user populations) benchmarks, the social interaction evaluation based on autonomy (i.e., need of human input during task execution), imitation (i.e., influence of robot similarities with humans on task performances) and privacy (i.e., influence of user's perceived sense of privacy on robot performance as an assistive presence) benchmarks and finally task performance evaluation based on social success (i.e., robot success to fulfill its social identity) and understanding of domain (i.e., influence of the robot's understanding of human behavior on the task performance) benchmarks. They mention works related to the different benchmarks but do not specify a way to evaluate them.

Two types of metrics for HRI or robot performances in HRI can be distinguished: subjective and objective metrics. Subjective measures are most often questionnaires to users after their interaction with the robot as in [16] where users were asked to evaluate individual system components (personal attention system, speech processing, dialog manager, robot internal status) and then to give their preferences concerning the robot. They can also be completed by observations as in [5]. [10] combines subjective and objective measures as well to evaluate the fluency in human-robot collaboration. The authors admit that the notion of fluency is not well defined and somewhat vague but claim it can be assessed and recognized when compared to non-fluent scenario. They build a questionnaire to assess how people perceive the collaboration with the robot. They also propose a list of objective metrics, only based on time measurement, designed to be quite general: robot idle time (i.e., robot waiting time for additional inputs from the human in order to take a decision), human idle time (i.e., human waiting time for the robot to complete an assigned task), concurrent activity (i.e., active time of both the robot and the human), functional delay (i.e., time difference between the end of one agent's task, either the human or the robot, and the beginning of the other agent's task). [18] claims the chosen metrics should be application and task dependent. They give sets of choosable metrics to evaluate navigation functions, perception abilities, robot's management by a human operator, manipulation tasks and social functionalities. Some of those metrics can be found in [13]. Then, they propose metrics that can be used to assess the system performance and that are not task dependent: the quantitative performance of a task execution based on the effectiveness (i.e., the percentage of the mission that was accomplished with the designed autonomy) and the efficiency (i.e., the time required to complete a task), the quality of the effort rated by the stakeholders in a subjective way and the appropriate utilization of mixed-initiative by measuring the percentage of requests for assistance made by robot, the percentage of requests for assistance made by operator and the number of interruptions of operator rated as non-critical. Finally, they propose metrics to evaluate the robot performance based on its self-awareness, its human awareness and its autonomy.

Some papers assess human-robot interactions with objective metrics based on the measurement of human states. In [19], they propose a framework allowing the robot to perceive (with face detection) and evaluate in real-time the affective state (i.e. anger, happiness, sadness, surprise, etc) and the engagement state (i.e. whether the person is interested or bored in the interaction) of the persons with which it's interacting. In [11], they developed a bio-instrumentation system evaluating human stress in real-time by measuring physiological parameters such as respiration, heart rate, perspiration, pulse wave and arm motion.

Only a few evaluate the interaction to make the robot adapt its behavior when the result is negative. In [19], they use the affective and engagement states evaluation of the human to improve the robot process of decision making. In [11], the robot executes a motion designed to decrease human stress when assessed too high.

There is a lack of robotic frameworks allowing the robot to evaluate in real-time the on-going interaction, based on multiple metrics. For us, the modeling of the estimated quality of interaction in the robot system is important as it gives an additional information to the decision-making system and allows more readability of the robot internal state.

3 Human-Robot Social Interaction Levels

There are different kinds of robots, with different kinds of purposes. We consider the ones that adopt a social behavior when interacting with humans and that got to execute and accomplish tasks in collaboration with them. An interaction can be seen under the scope of three levels. From the smallest to the biggest: actions which come within collaborative tasks which, in turn, are found within an interaction session.

3.1 Interaction Sessions

We define an interaction session as the period during which the robot and a human interact together and are engaged (for simplification purpose, we stay in the context of the interaction between one robot and one human). It is divided in three parts, based on the structure of an interaction defined in [14] and the engagement model of [17]. It starts by initiating an interaction with another agent, then it lasts as long as the interactants are maintaining the interaction through conversation and collaborative tasks and, finally it ends when at least one of the interactant is disengaged, either by abruptly ending the interaction, by closing the interaction as in [15][4] or by the completion of the goals of the interaction. During an interaction session, both the robot and the human need to remain engaged. There is no unique definition of what it means to be engaged. We chose one which is frequently used and that comes from [17]: "Engagement is the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake". The robot has to be able to demonstrate engagement and disengagement processes but to recognize them as well. The engagement behavior varies depending on the state of the interaction: conversation, collaborative task or idle interactant(s). There is a need to define what behavior the robot has to exhibit and what behavior it should expect from the human for each state. Therefore, in our model, the robot looks at the other interactant's face (i.e. keeps its head oriented toward the interactant's face) to demonstrate engagement in conversation and idle contexts. The robot will expect the same kind of behavior from the human in those. As for collaborative tasks, exhibited and expected behaviors have to be part of the task management.

3.2 Collaborative tasks

The robot and the human it is interacting with may have to perform tasks together. Each task is executed by agents in order to achieve a goal and this execution follows a plan. The following formalisms are inspired from [6].

Goal A goal g is defined as:

$$g = \langle name, AGC, O \rangle$$

where *name* is used to identify the goal, O is a set of facts representing the desired world state and AGC is a set of agents involved in the goal.

Plan There are many ways to generate a plan. But no matter the way, automated planning (plan computed before execution), reactive planning (action selection at execution time) or another planning type, a plan is a sequence of primitive tasks that are also called actions.

A plan p is defined as:

$$p = \langle id, g, ACP, L \rangle$$

where id is used to identify the plan and g is the goal that the plan allows to achieve. ACP is a set of actions (described in 3.3) that composes the plan, and L is a set of links defining actions or plans order and causal links. A link l is defined as $l = \langle previous, after \rangle$ where $previous$ is the id of the action or the plan which needs to be achieved before the action or the plan with the id $after$ is performed.

Agents An agent is a robot or a human that will take part in the task execution. An agent is defined as:

$$ag = \langle name, type, CAP, ms, AG \rangle$$

where $name$ is used to identify the agent, $type$ represents if the agent is a human or a robot, CAP the set of high level action names, representing the actions that the agent is able to perform, MS is the mental state of the agent (described in the next paragraph) and AG is a set of agents containing all agents excepting $ag : AG = AGC \setminus ag$. These agents are defined in the same way as ag and model how the given agent represents them.

The mental state MS of ag is defined as:

$$ms = \langle WS, gs, ps, ACS \rangle$$

where WS is a set of facts representing the current world state from the agent point of view. gs represents the state of the goal from the agent point of view. It can be either *progress*, *done* or *aborted*. ps represents state of the plan from the agent point of view. It can be either *progress*, *done*, *aborted* or *unknown* if the agent is not aware of the plan. Finally, ACS represents the set of states of the actions from the agent point of view. The state of an action ac is represented as $acs = \langle id, state \rangle$ where $state$ can be either *progress*, *done*, *failed*, *asked* (an agent asked for the action to be done), *planned* (need to be done later according to the current plan), *needed* (need to be done now according to the current plan but not possible), *ready* (need to be done now according to the current plan and possible) or *unknown*. The evolution of the state of an action ac is described in Fig. 1.

3.3 Actions

To achieve a goal, agents will have to execute a plan which is a sequence of actions, the smallest level of the interaction. What we call action is a primitive

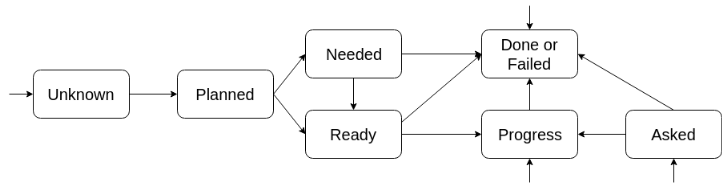

Fig. 1. Evolution of the state of an action ac

task, a task that cannot be decomposed at the level of the component supervising the execution of the goal. Let's ACT be the set of all actions. An action $ac \in ACT$ is defined as:

$$ac = \langle id, name, AGC, EN, PRE, EFF, EXP, dl \rangle$$

where id is the action identifier and $name$ represents its name. AGC is a set of the names of the agents needed for the action execution, EN a set of entities (objects or agents) which allows to define precisely the action. PRE and EFF are sets of facts representing respectively the action preconditions and effects. EXP is a set of expected reactions from the other agents to the executed action. dl is the deadline, the latest time by which the action has to be executed.

4 Introduction of new metrics at the task level

We presented above the three levels defining a human-robot social interaction. We now describe two novel metrics that will be taken into account to compute the evaluation of the quality of interaction.

4.1 Distance to the goal

A task is executed to complete a goal. However, during the execution, a lot of obstacles and unexpected events can happen (i.e. a physical obstacle is on the way of the path during a navigation task, the human leaves the task for some time, etc.). To have a hint about how well a task is going, we define a new metric called distance to the goal d_g. At a frequency defined by the programmer, the system computes if, at time t the goal is closer to be achieved than at time $t-1$. As the metrics "task effectiveness" of [13] and "efficiency" of [18], it tries to measure if the agents have been efficient but allows to have more details about the proceedings of the task and to react in real time. It is quite obvious how to measure the distance to the goal for geometric actions (e.g. the robot is 2 meters away from the goal of a navigation task) but it is less for other kinds of tasks (e.g. in a route guidance task as in [8], there is no obvious metric to measure the distance to the goal). Therefore, to generalize and to abstract from units of measurement, we suggest to use the goal distance variation Δd_g and not an absolute value of d_g. We give four methods to measure the distance to the goal variation, that can be chosen according to the type of actions or the type of task planning:

- Distance to the goal variation based on the action state in MS_{robot}, knowing that $(state_t = successor(state_{t-1})) \vee (state_t = state_{t-1})$

$$\begin{cases} (acs_t \neq acs_{t-1}) \wedge (state_t \neq failed) \Rightarrow \Delta d_g < 0 \\ (acs_t \neq acs_{t-1}) \wedge (state_t = failed) \Rightarrow \Delta d_g > 0 \\ acs_t = acs_{t-1} \Rightarrow \Delta d_g = 0 \end{cases}$$

- Distance to the goal variation based on action repetitions with n the nth time the action is being executed:

$$ag.perform(ac)_{n,t} \ \wedge \ acs_{t-1} = \langle id_{ac}, failed \rangle \ \wedge \ n > 1 \Rightarrow \Delta d_g > 0$$

- Distance to the goal variation based on the geometric distance for an action involving spatial moves of $en \in EN_{ac}$ with $p_{en,c}$ the current position of en and $p_{en,p}$ the planned position:

$$\begin{cases} d(p_{en,c}, p_{en,p})_t < d(p_{en,c}, p_{en,p})_{t-1} \Rightarrow \Delta d_g < 0 \\ d(p_{en,c}, p_{en,p})_t > d(p_{en,c}, p_{en,p})_{t-1} \Rightarrow \Delta d_g > 0 \\ p_{en,c,t} = p_{en,c,t-1} \Rightarrow \Delta d_g = 0 \end{cases}$$

- If using a planner: Distance to the goal variation based on the variation of the cost or the variation of the number of steps to realize before to reach the goal:

$$\begin{cases} cost_t < cost_{t-1} \Rightarrow \Delta d_g < 0 \\ cost_t > cost_{t-1} \Rightarrow \Delta d_g > 0 \end{cases}$$

$$\begin{cases} n_t < n_{t-1} \Rightarrow \Delta d_g < 0 \\ n_t > n_{t-1} \Rightarrow \Delta d_g > 0 \end{cases}$$

4.2 Human contribution to the goal

During the task execution, some actions have to be realized by the human agent. We argue that it is important to evaluate if the human actually performs the actions the robot expects her to, if she is committed to the task as defined in [3]. So, the robot can adapt, even disengage from the interaction if, for example, she is considered not at all committed for a while. To evaluate the human contribution to the goal allows to have information about the human commitment to the task. We define ACH, the set of actions for which the human is an actor and that the robot expects to be realized at time t, as $ACH \subset ACP$ with $\forall ac \in ACH$ and $human \in AGC$. We propose a representation of the human contribution for a given action ac in Fig. 2, $ca_{t,ac} = \{good, undefined, bad\}$.

We also propose a representation of the human contribution for a given task, $cg_t = f(ca_{t,ac}) = \{good, undefined, bad\}$, described in Fig. 3.

These two models are quite simple but are a first draft that can be enriched in the future.

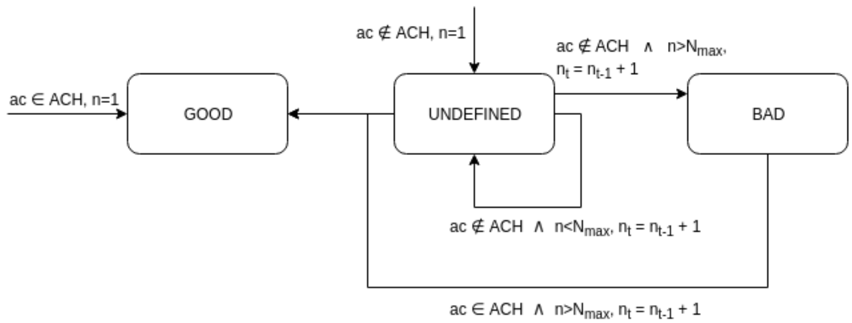

Fig. 2. Evolution of the human contribution for a given action ac, with ACH the set of actions for which the human is an actor, n the nth time the action is being executed and N_{max} the defined threshold to go from a undefined contribution to a bad one.

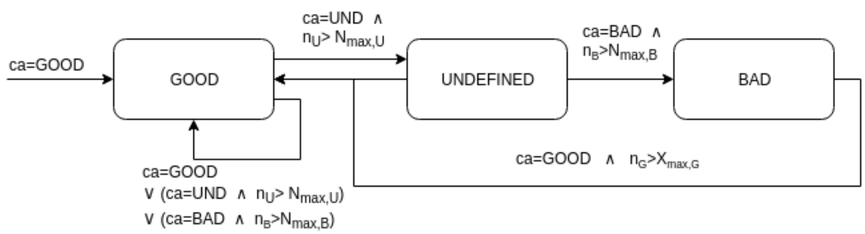

Fig. 3. Evolution of the human contribution for a given task. n_G is the number of time ca output *good*, n_U is the number of time ca output $undefined$, n_B is the number of time ca output *bad*. The $N_{max,U}, N_{max,B}, N_{max,G}$ are respectively the corresponding thresholds: the number of times ca can be $undefined$ before assessing cg as $undefined$, the number of times ca can be bad before assessing cg as *bad* and the number of times ca has to be good before assessing cg as *good*.

5 Evaluation of the Quality of Interaction

Until now, only a few works took interest in the quality of the interaction from the robot point of view. Therefore we decided to investigate the topic. Having this ability would enhance, in future work, its decision making process. There are three main reasons for measuring the interaction quality:

- At task execution time, the robot can make the choice to change the modalities of the different elements it uses, for example the language in which it communicates with the human, the volume of its speakers, or the parameters of its planner.
- Still at task execution time, a very bad performance all along a task could allow to recognize humans not engaged in the interaction, not committed to tasks or trying to play the robot. Then, it would leave the interaction.
- Off-line, using interaction session logs, it allows the developer to improve the system if too many interactions are measured as bad and she can adjust her

design according to the agents' performances. As well, it allows to identify blocking parts of interactions/tasks.

We define the quality of interaction at three different levels, each time as a function of multiple metrics as in [2]. They can be used and weighted according to the designer, to what is important to measure and to know during a specific session or a specific task.

5.1 At the interaction session level

The quality of interaction assessment purpose at the interaction session level is to be able to say if globally the session went well. We suggest to take into account:

- the way the session has been terminated (i.e., ended with goodbyes, left in the middle of a task, etc)
- the number of executed tasks with the human during the session, NE
- the number of succeeded tasks with the human during the session, NS
- the efficiency of the interaction, the percentage of successive tasks, $NP = \frac{NS}{NE}$
- the success rate of a given task (i.e. pick and place task always fail)
- the average of the human contributions along the tasks

5.2 At the task level

The purpose of the assessment of the quality of interaction at the task level is to give the robot an additional information on the ongoing task so it can adapt, taking into account:

- the success of a given task at the end of the execution
- the distance to goal
- the agent performance $AP = \frac{succeeded(AC)}{expected(AC)}$
- the human contribution to the goal
- the human commitment based on commitment generation and the repairing models of [3]

5.3 At the action level

Evaluating the quality of interaction at the action level as well allows the robot to adapt in a different way than at the task level.

- the fulfilling by the robot of the human expectations to her executed actions (i.e., the robot understood the human sentence but knows or not any appropriate answer)
- the human executing the expected action
- the state of the action executed

6 Conclusion

We propose a novel way to evaluate the human-robot interaction. Until now, most of the works presented a robot evaluation from the human point of view. The few works evaluating the interaction from the robot point of view do it based on only one metric. Our evaluation is based on multiple metrics of whom two that we detailed, the distance to the goal and the human contribution to the goal. We defined a model to evaluate the quality of interaction at three different levels, from the larger to the smaller: the interaction session level, the task level and the action level. The robot can use this data in real-time to adapt or to react to its human partner. In future work, we are preparing an implementation of a system that gives this ability to the robot.

References

1. Aly, A., Griffiths, S., Stramandinoli, F.: Metrics and benchmarks in human-robot interaction: Recent advances in cognitive robotics. Cognitive Systems Research **43**, 313–323 (Jun 2017). https://doi.org/10.1016/j.cogsys.2016.06.002, http://www.sciencedirect.com/science/article/pii/S1389041716300912
2. Bensch., S., Jevtić., A., Hellström., T.: On interaction quality in human-robot interaction. In: Proceedings of the 9th International Conference on Agents and Artificial Intelligence - Volume 1: ICAART,. pp. 182–189. INSTICC, SciTePress (2017). https://doi.org/10.5220/0006191601820189
3. Castro, V.F., Clodic, A., Alami, R., Pacherie, E.: Commitments in human-robot interaction (2019)
4. Clark, H.H., French, J.W.: Telephone goodbyes. Language in Society **10**(1), 1–19 (1981), http://www.jstor.org/stable/4167185
5. Dautenhahn, K., Werry, I.: A quantitative technique for analysing robot-human interactions. In: IEEE/RSJ International Conference on Intelligent Robots and Systems. vol. 2, pp. 1132–1138 vol.2 (Sep 2002). https://doi.org/10.1109/IRDS.2002.1043883
6. Devin, S., Alami, R.: An implemented theory of mind to improve human-robot shared plans execution. In: The Eleventh ACM/IEEE International Conference on Human Robot Interaction. pp. 319–326. HRI '16, IEEE Press, Piscataway, NJ, USA (2016), http://dl.acm.org/citation.cfm?id=2906831.2906886
7. Feil-Seifer, D., Matarić, M.J.: Human RobotHuman–robot interaction (HRI)InteractionInteractionhuman robot, pp. 4643–4659. Springer New York, New York, NY (2009). https://doi.org/10.1007/978-0-387-30440-3_274, https://doi.org/10.1007/978-0-387-30440-3_274
8. Foster, M.E., Craenen, B., Deshmukh, A., Lemon, O., Bastianelli, E., Dondrup, C., Papaioannou, I., Vanzo, A., Odobez, J.M., Canévet, O., Cao, Y., He, W., Martínez-González, A., Motlicek, P., Siegfried, R., Alami, R., Belhassein, K., Buisan, G., Clodic, A., Mayima, A., Sallami, Y., Sarthou, G., Singamaneni, P.T., Waldhart, J., Mazel, A., Caniot, M., Niemelä, M., Heikkilä, P., Lammi, H., Tammela, A.: Mummer: Socially intelligent human-robot interaction in public spaces (2019)
9. Genise, P.: Usability evaluation: Methods and techniques: Version 2.0 (August 2002), https://web.archive.org/web/20050303232908/http://www.cs.utexas.edu/users/almstrum/cs370/elvisino/usaEval.html, accessed: 2019/09/23

10. Hoffman, G.: Evaluating fluency in human–robot collaboration. IEEE Transactions on Human-Machine Systems **49**(3), 209–218 (June 2019). https://doi.org/10.1109/THMS.2019.2904558
11. Itoh, K., Miwa, H., Nukariya, Y., Zecca, M., Takanobu, H., Roccella, S., Carrozza, M.C., Dario, P., Takanishi, A.: Development of a bioinstrumentation system in the interaction between a human and a robot. In: 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 2620–2625 (Oct 2006). https://doi.org/10.1109/IROS.2006.281941
12. Movellan, J.R., Tanaka, F., Fasel, I.R., Taylor, C., Ruvolo, P., Eckhardt, M.: The rubi project: a progress report. In: 2007 2nd ACM/IEEE International Conference on Human-Robot Interaction (HRI). pp. 333–339. IEEE (2007)
13. Olsen, D.R., Goodrich, M.A.: Metrics for evaluating human-robot interaction. In: Proc. NIST Performance Metrics for Intelligent Systems Workshop (2003)
14. Robinson, J.D.: Overall structural organization. In: Sidnell, J., Stivers, T. (eds.) The Handbook of Conversation Analysis, pp. 257–280. John Wiley & Sons, Ltd (2012). https://doi.org/10.1002/9781118325001.ch13, `http://doi.wiley.com/10.1002/9781118325001.ch13`
15. Schegloff, E., Sacks, H.: Opening up closings. Semiotica **8**, 289–327 (1973). https://doi.org/10.1515/semi.1973.8.4.289
16. Shuyin Li, Kleinehagenbrock, M., Fritsch, J., Wrede, B., Sagerer, G.: "biron, let me show you something": evaluating the interaction with a robot companion. In: 2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583). vol. 3, pp. 2827–2834 vol.3 (Oct 2004). https://doi.org/10.1109/ICSMC.2004.1400761
17. Sidner, C.L., Lee, C.: Engagement rules for human-robot collaborative interactions. In: SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme - System Security and Assurance (Cat. No.03CH37483). vol. 4, pp. 3957–3962 vol.4 (Oct 2003). https://doi.org/10.1109/ICSMC.2003.1244506
18. Steinfeld, A., Fong, T., Kaber, D., Lewis, M., Scholtz, J., Schultz, A., Goodrich, M.: Common metrics for human-robot interaction. In: Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-robot Interaction. pp. 33–40. HRI '06, ACM, New York, NY, USA (2006). https://doi.org/10.1145/1121241.1121249, `http://doi.acm.org/10.1145/1121241.1121249`
19. Tanevska, A., Rea, F., Sandini, G., Sciutti, A.: Towards an Affective Cognitive Architecture for Human-Robot Interaction for the iCub Robot. In: 1st Workshop on "Behavior, Emotion and Representation: Building Blocks of Interaction". Bielefeld, Germany (Oct 2017), `https://hal.inria.fr/hal-01615491`
20. Weiss, A., Bernhaupt, R., Lankes, M., Tscheligi, M.: The usus evaluation framework for human-robot interaction. In: Proceedings of the Symposium on New Frontiers in Human-Robot Interaction (2009). https://doi.org/10.1075/ais.2.07wei