Chapter 5

Evaluation of the global system

Contents

5.1 Motivations

In the two previous chapters, we presented several ameliorations on the way the robot elaborates and executes Shared Plans. We first endowed the robot with the ability to take into account humans mental states during Shared Plans execution. In a second time, we saw how the robot is able to compute more flexible Shared Plans where it identifies which decisions have to be taken at planning time and which one are better to be postponed. Then, the robot is able to take these decisions while smoothly adapting to the human choices.

These two ameliorations have been quantitatively and independently evaluated in simulation. In this chapter, we want to evaluate the global system including both ameliorations. Moreover, in addition to quantitative results, we want to evaluate the acceptance of the system by real users. To do so, we defined a task which allows to highlight the benefits of the system. This task has been used to evaluate the global system in simulation in order to get quantitative results. Then, the same task (with minor modifications) has been used during a user study in the real robot in order to get a subjective evaluation of the global system.

5.2 Task

The task used for the global evaluation is inspired from the "Inventory scenario" of Chapter **??**. In the task, the human and the robot have to scan several colored cubes and store them into a box of the same color. At the beginning of the interaction, both agents have a stack of colored cubes they can access (and only them can access). There are blue, green and red cubes. The stack of the human is located in another room, in a way that, to get an object, the human has to leave the sight of view of the robot (see Fig. 5.1). For the cubes to be scanned, the agents need to put them on one of the two possible areas on the table in front of the robot (see Fig. 5.1). Once a cube is on an area, the robot can scan it by orienting its head and turning on a red light in the direction of the object (see Fig.é5.3). If the robot scans an object while the human is not looking at him (e.g. he is in another room to pick a cube), the human will not be aware that the object has been scanned unless the robot tells him. Once the cube scanned, it can be stored in a box of the same color (e.g. the blue cubes in a blue box). The robot has access to a blue box, the human to a green box, and both have access to a red box. Consequently, only the robot can store the blue cubes, only the human can store the green cubes and both can store the red cubes. As well as for his stack, the boxes of the human are located in another room (see Fig. 5.1).

Both in simulation and in the user study, we compared 4 different conditions:

- using the original system, called Reference System (RS), with all decisions and instantiations performed at planning time and no estimation of the human mental state:
  - **RS-none mode:** the robot verbalizes nothing (unless it is strictly necessary)
  - **RS-all mode:** the robot informs the human when he has to perform an action, when it will act and about all actions he missed.
- using the proposed system, called New System (NS):
  - **NS-N:** the robot uses the **Negotiation** mode previously defined when a decision needs to be made concerning *X agent* action,
  - **NS-A:** the robot uses the **Adaptation** mode.

5.3 Evaluation in simulation

5.3.1 Modalities

We first evaluated our system in simulation. Different set-ups were used as initial state of the task: we randomized the composition of the stack of the human and the robot. In each case there was three cubes of each color (red, blue and green) and the robot stack was composed of 4 cubes and the human one of 5 cubes. The

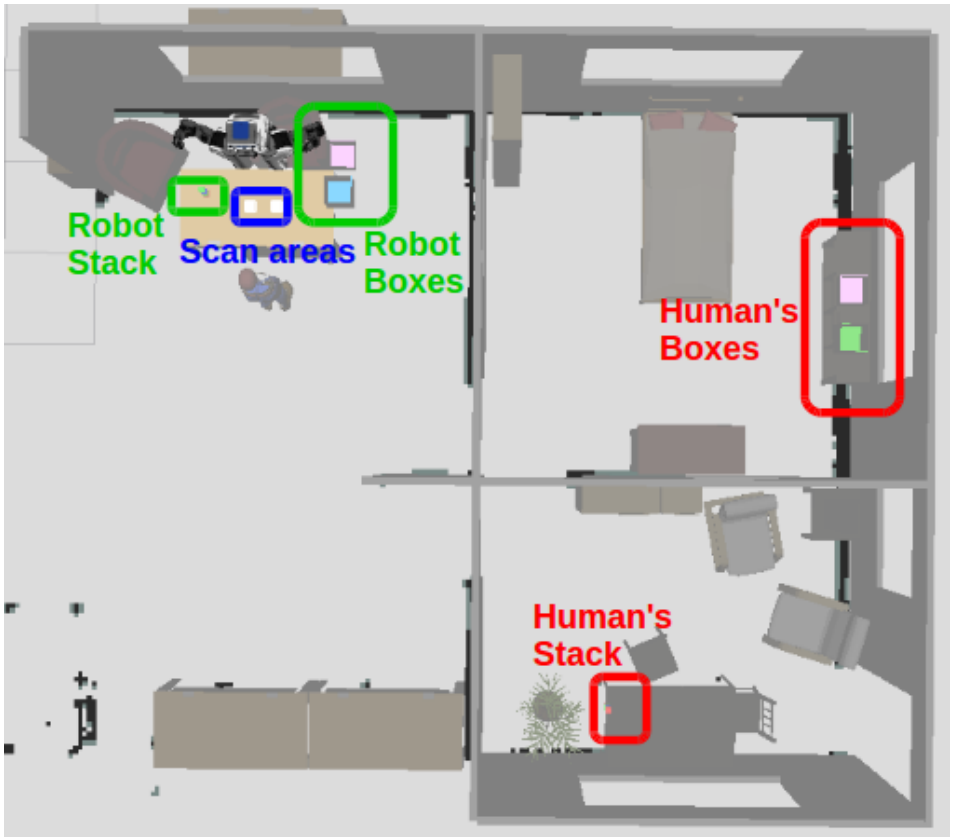


Figure 5.1: Set-up for the task used during evaluation. The human and the robot need to take the cubes from their stack and put them in the scan areas. Once a cube is in a scan area, the robot can scan it. Then, the agents can store the cubes in the boxes of the same color. The human has access to a green and a red box and the robot has access to a blue and a red box.

dispositions of the cubes in the stacks were randomized. The robot was confronted to a simulated human with different kinds of behaviors. In all cases, the human was acting as bellow:

- when the human is in front of the robot with no cube in hand and there is a green cube he knows it is scanned, it goes to the boxes to store it.
- whenever the human is idle with no cube in hand, it goes to his stack to pick a cube and then comes back to the table.
- if the human has a cube in hand to be scanned, he put it on a scan area (if free). If there is no free area, the human waits in front of the robot.
- if the human has no more cube in his stack he waits in front of the robot.

When the human is in front of the robot with no cube in hand and there is a red cube he knows it is scanned, the human chooses:

- to store it systematically (**hurry-case**)
- not to store it systematically (**lazy-case**)
- to store it with 50% chance (**50%-case**)

Then, we settled two different human behaviors:

- **the "kind" human (case=K)** who adapts his behavior to what the robot verbalizes (i.e. does an action if the robot asks him and does not execute the actions the robot says it will perform)
- **the "stubborn" human (case=S)** who does not react nor comply to robot verbalization (he will not change his decision whatever the robot says).

In all cases, the human answers to the robot questions concerning the red cubes with the answer corresponding to his decision.

We measured:

- *the number of verbal interactions* between the human and the robot (either an information given by the robot or question asked), in Tab. 5.2.
- *the number of human/robot incompatible decisions*: either both decide to perform the same action (and the robot stops its own action to avoid the conflict) or both decide not to perform the action (the robot first asks the human to perform the action after a predefined time and, if after another period the human has still not executed the action, the robot looks for a new plan where it can proceed), in Tab. 5.1.
- *the total execution time:* for the human and the robot to perform the task, in Fig. 5.2.

5.3.2 Results

	RS-none	**RS-all**	**Neg**	**Adapt**
50%-K	2.4 (0.84)	20.7 (1.34)	3.4 (1.51)	2 (1.33)
hurry-K	1.8 (0.79)	21.1 (2.08)	1.9 (1.10)	2.2 (1.13)
lazy-K	3.0 (1.33)	21 (1.56)	3.3 (1.42)	1.6 (1.17)
50%-S	2.5 (1.43)	23.9 (1.59)	3.3 (1.49)	1.7 (0.95)
hurry-S	1.5 (0.97)	20.9 (1.29)	2.4 (1.89)	1.9 (0.99)
lazy-S	3.2 (0.92)	25.2 (1.55)	2.8 (1.68)	1.8 (1.14)

Table 5.1: Results for the reference system (RS) and the proposed system (NS-N for the negotiation mode and NS-A for the adaptation mode). Number of verbal interactions (i.e. question asked by the robot in the negotiation mode or an information given with the reference system). The numbers correspond to means in 10 runs and their associated standard deviations.

RS-none performance: Even if the robot is supposed not to speak in this mode, we can see in Tab. 5.1 that there are still verbalizations, especially in the cases where the human is lazy. These verbal interactions are due to two reasons. First, when the robot decides that the human should store a red cube and the human decides he will not do it, the robot unlocks the situation by asking the human to store the object (and in the stubborn case, as the human will still not do it, it then changes its plan). Secondly, when the human does not see that a cube has been scanned he will wait before storing it. As previously, the robot unlocks the situation by asking the human to store the object (as it detects that the human is not executing his action).

This mode is also the mode where there are the most incompatible decisions as the robot verbalize nothing. These incompatible decisions are mainly conflicts concerning the red cubes to store and the scan areas (as there is no notions of *similar* objects in this mode, the robot stops its actions if the human puts a cube in the same area it was aiming for even if the other is free).

Concerning the execution time, this mode is the one with the highest ones. The execution times are especially high in the stubborn and lazy cases, as, when the human decides not to store a cube, the robot wastes time to ask the human to do it and only then looks for a new plan where it stores the cube.

	RS-none	**RS-all**	**NS-N**	**NS-A**
50%-K	2.9 (0.99)	0.9 (0.57)	0.6 (0.7)	0.3 (0.48)
hurry-K	2.5 (0.97)	1.0 (0.94)	0.6 (0.52)	0.4 (0.52)
lazy-K	3.5 (1.08)	0.8 (0.63)	0.5 (0.7)	0.5 (0.53)
50%-S	2.9 (1.45)	1.9 (0.99)	0.6 (0.52)	0.5 (0.97)
hurry-S	2.3 (1.34)	1.0 (0.82)	0.5 (0.53)	0.4 (0.52)
lazy-S	3.5 (0.97)	2.6 (1.84)	0.3 (0.67)	0.4 (0.52)

Table 5.2: Results for the reference system (RS) and the proposed system (NS-N for the negotiation mode and NS-A for the adaptation mode). Number of incompatible decisions between the human and the robot (i.e. either both agents decide to perform the same action or both decide not to perform a given action). The numbers correspond to means in 10 runs and their associated standard deviations.

RS-all performance: In this mode, as expected, there is a lot of verbal interactions. Indeed, the robot informs not only about who should perform the actions but also about all actions that the human missed. However, even with the "kind" human, it is not enough to get ride of all conflicts. Indeed, there are still conflicts concerning the scan areas as the robot has to stop its action if the human puts his cube in the area it was aiming for. There are even more conflicts with the "stubborn" human as, even if the robot gives information, the human does not change his choices.

Concerning the execution times, they are low for the "kind" human as the human follows what the robot asks. However, with the "stubborn" human (not with the

"hurry" one as the human takes the initiative to execute all possible actions), the execution times become higher. Indeed, when the robot has decided that the human has to perform an action, the robot wastes time to wait for the human to perform it before looking for another plan. These execution times are still lower than with RS-NONE because, as the robot informs for all missed actions, there is no time where the human waits to know that a cube has been scanned.

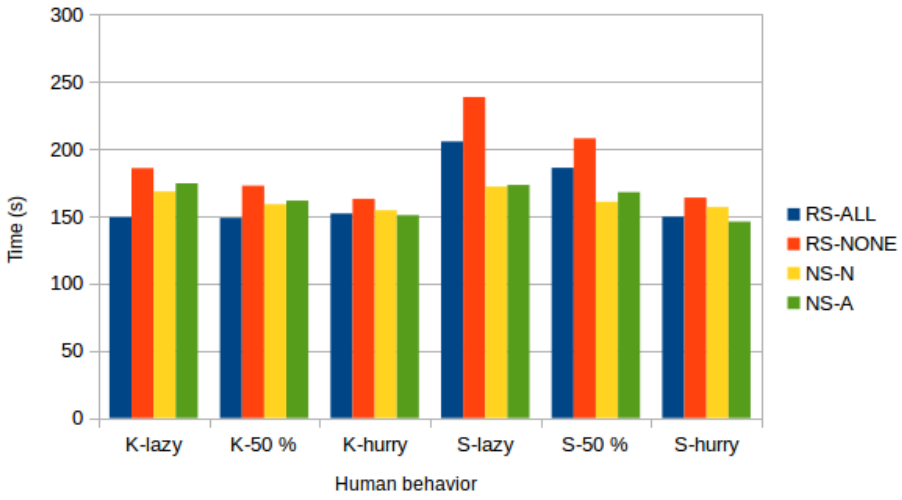


Figure 5.2: Time in seconds each system spent performing the task for each kind of human behavior (mean in 10 runs).

New system performance: We can see that the performance of the new system is globally better than in the two other modes. Concerning the incompatible decisions, it only remains the conflicts when the human puts a cube on the last available scan area and the robot was trying to put an object on it too. The execution times are lower than the reference system when the human is stubborn. Indeed, the robot does not wait for the human to perform actions he does not want to execute (it either asks or adapts). Moreover, as the robot informs the human about the cube which has been scanned during his absence (and which the human can store), the human does not wait to store cubes.

Concerning the verbal interactions, they are higher for the negotiation mode as the robot asks to the human if he wants to store the red cubes (but only when both agents are available). For the adaptation mode, these verbal interactions correspond to the information concerning the missing scan actions of the green or red objects.

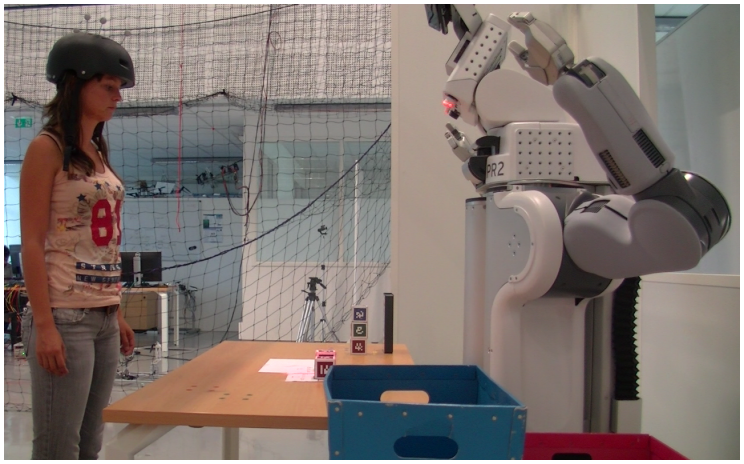


Figure 5.3: The PR2 robot interacting with a subject to achieve the task. The robot is scanning the cube before storing it.

5.4 User study

5.4.1 Adaptations of the task for the study

Before realizing the real study, we made some pre-tests by running the task in the robot with few subjects. During these pre-tests we noticed several possible problems that we fixed by proceeding to small adaptations of the task.

Introduction of a red tape: In certain cases, the configuration coupled to the decision of the subjects led to not having any decision in the task concerning the red cubes. Indeed, there were cases where, each time there was a red cube to store, one of the two agents were busy (either the human were in another room to pick or store an object or the robot was performing another action).

To ensure that, at each interaction, there is at least one decision to take between the human and the robot, we added to the objects to scan and store a red tape. The human and the robot both have a red tape in the same emplacement as their stacks of cubes. At the end of the task, when all the cubes are scanned and stored (and so both agent are available), **only one** of the two tapes (the one of the human or the one of the robot) needs to be put on a scan area. Then, as well as for the cubes, the robot scans the tape. Finally, as the tape is red, it needs to be stored in a red box either by the human or the robot.

Distraction task: We noticed that some subjects tried not to miss any action of the robot (they stayed in front of the robot each time there was a cube to scan and they hurried in the places where they cannot see the robot). Consequently, there was no missing knowledge during the task for these subjects. To ensure that all subjects miss some actions of the robot, at one predefined point of the task, the experimenter asked the subject to leave the task for a while to perform another

task. In this task, the subject has to build a construction shown in a picture with Lego bricks. Once the construction achieved, the subject is free to go back to the main task.

5.4.2 Protocol

Each subject of the study had to interact with the robot to achieve the task previously described, and in the four conditions described in Sec. 5.2. The order in which they were confronted to the different conditions was randomized. There were four different compositions for the stacks of the human and the robot. The attribution of each composition to a condition was also randomized for each participant.

At their arrival, the participants were introduced to the robot and the environment of the study by the experimenter. Then, participants were asked to read instructions explaining the task and its constraints. The experimenter checked the good understanding of the instructions and showed the emplacements of the different objects of the task. The participants were then asked to perform a quick familiarization task. In this task, the human and the robot had only one cube in their stacks (a blue for the human and a green for the robot). They had to put them in the scan areas, scan them and then store them in the appropriate boxes. There was no tape in the familiarization task.

After each interaction with the robot (for each condition), the participants were asked to fill a questionnaire in order to evaluate their feeling concerning the robot and the interaction. To do so, several questionnaires have already been developed to evaluate human-robot interaction. [Hoffman 2013] allows to evaluate several aspects of the interaction as the "trust" in the robot or the "fluency" of the interaction. It has been used in several studies as [Gombolay 2015] or [Dragan 2015]. However, this questionnaire lacks some aspects needed in our evaluation as "acceptability" or "usability". The Godspeed questionnaire [Bartneck 2009] allows to measure the perception of the robot by the human with questions relative to "anthropomorphism" or "perceived intelligence". However, this questionnaire is focused on the evaluation of the perception of the robot and lacks parts on the evaluation of the interaction and on the usability of the system. The SUS (System Usability Scale) questionnaire [Brooke 1986] allows to measure the interaction of a user with an electronic system with 10 affirmations which subjects need to evaluate with a Lickert scale from "totally agree" to "totally disagree". On the contrary of the Godspeed, the SUS questionnaire measures the usability of the system but lacks of measure concerning the perception of the robot or the interaction. Finally, [Heerink 2009] presents a toolkit to measure acceptance for assistive social robots. This toolkit is based on the UTAUT (Unified Theory of Acceptance and Use of Technology) questionnaire [Venkatesh 2003]. It has been well conceived in order to evaluate the perception and usability of the robot and more particularly for social robots. However, the questionnaire is more oriented toward the perception of the robot than the interaction and the collaboration.

For this study, like in [Heerink 2010, Fischer 2016], we use a questionnaire that

we conceived for the experiment. This questionnaire is composed of several dimensions:

- *Verbal:* this dimension, composed of 3 questions allows to evaluate how the human perceived the verbal interaction with the robot.
- *Acting:* this dimension, composed of 3 questions allows to evaluate how the human perceived the decisions of the robot concerning its actions.
- *Collaboration:* this part, based on [Weistroffer 2014] and composed of 5 questions, allows to evaluate how the human perceived the collaboration between the human and the robot.
- *Interaction:* this dimension, based on the AttrakDiff questionnaire [Lallemand 2015] and composed of 5 questions, allows to evaluate how the human perceived the interaction between the human and the robot.
- *Robot perception:* this dimension, based on the Godspeed questionnaire [Bartneck 2009] and composed of 8 questions, allows to evaluate how the human perceived the robot in general.

For all the questions, the subject was asked to place himself in a scale of 100 between two antonym adjectives. The English translation of the questionnaire can be found in Appendix **??**.

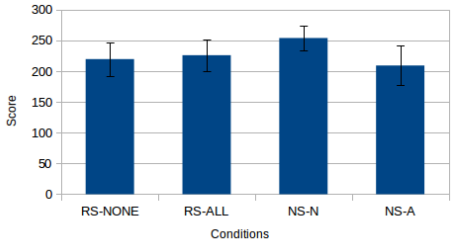
In addition to these two questionnaires, after each interaction with the robot (including the familiarization task), we asked participants to answer a small yes/no questionnaire. This questionnaire contains general questions about what happened during the interaction (e.g. "Do you think all the cubes have been scanned?"). The aim of this questionnaire was to remind the key points of the interaction to the subjects (because we noticed during the pre-tests that subjects were kind of "lost" and did not know on what to focus their attention). This questionnaire can also be found in Appendix **??**.

5.4.3 Results

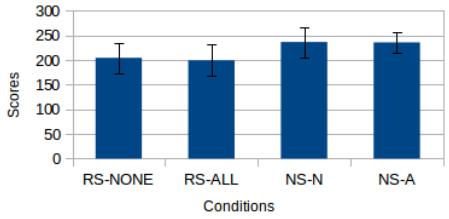
21 subjects took part in the study (8 women and 13 men). They were all fluent in french and had no significant experience in robotics. The results for the questionnaire evaluating the subjects feeling concerning the robot and the interaction can be found in Fig. 5.5. We will discuss the results here bellow.

Verbal dimension: The scores of the different modes for the verbal dimension of the questionnaire can be found in Fig. 5.4(a). The negotiation mode of the new system (NS-N) has been found significantly better than the RS-none condition and the adaptation mode (NS-A) ($p < 0.05$[1]). Even if the negotiation mode had

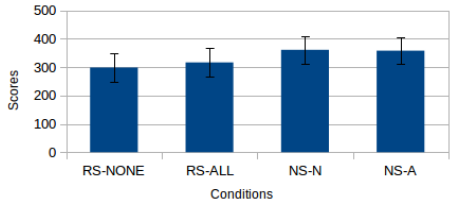
[1]The p values have been calculated with the Student's t-test when the data were normally distributed and with the Wilcoxon test otherwise



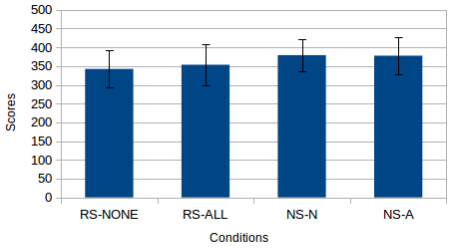
(a) Scores for the "Verbal" dimension of the questionnaire



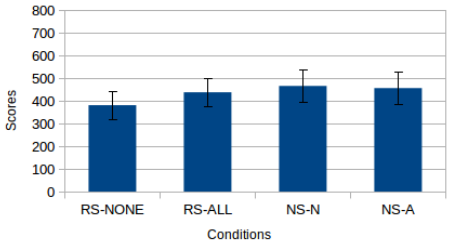
(b) Scores for the "Acting" dimension of the questionnaire



(c) Scores for the "Collaboration" dimension of the questionnaire



(d) Scores for the "Interaction" dimension of the questionnaire

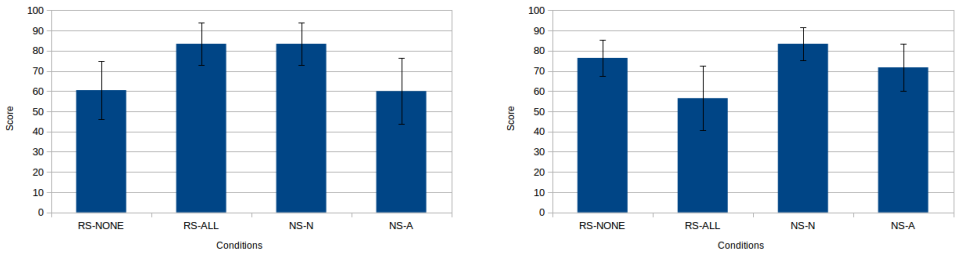


(e) Scores for the "Robot perception" dimension of the questionnaire

Figure 5.4: Results on the questionnaire evaluating the subjects feeling concerning the robot and the interaction given to the participants during the user study. The different modes are for the reference system RS-all when the robot verbalizes everything and RS-none when the robot verbalize nothing and for the proposed system NS-N for the negotiation mode and NS-A for the adaptation mode.

a higher score than the RS-all condition, the difference was not found significant. Indeed, when discussing with subjects after the experiment, we found that, for some of them, the fact that the robot was speaking a lot was reassuring. However, they also point out the fact that, even if they found it reassuring the first time, if they had to interact with the robot several time in this mode, they would quickly find it "annoying". Indeed, if we look at some details of the questions asked into the verbal part of the questionnaire, the verbal interaction of the robot have been found more superfluous in the RS-all mode than in the other modes (see Fig. 5.5(b)). Moreover,

the verbal interactions in the RS-none mode and the adaptation mode have been found less sufficient than in the other modes (see Fig. 5.5(a)). Indeed, the fact that the robot does not inform about its choices (and more particularly concerning the red objects) was found disturbing by the participants.



(a) Score for the question concerning the verbal interactions where subjects were asked to choose between "insufficient" (0) and "sufficient" (100)

(b) Score for the question concerning the verbal interactions where subjects were asked to choose between "superfluous" (0) and "pertinent" (100)

Figure 5.5: Details of the results on the verbal dimension of the questionnaire. The different modes are for the reference system RS-all when the robot verbalizes everything and RS-none when the robot verbalize nothing and for the proposed system NS-N for the negotiation mode and NS-A for the adaptation mode.

Other dimension: Concerning the rest of the questionnaire, in all other dimension, the new system (combination of the negotiation and adaptation modes) had scored significantly higher ($p < 0.05$) than the reference system (combination of RS-all and RS-none). The difference was particularly visible for the *Acting* part of the questionnaire ($p \simeq 0.003$). It shows that the algorithms developed for the robot to be able to take the appropriate decisions at the right time during Shared Plan achievement have been appreciated by the subjects.

Questionnaire validation: In order to validate the coherence and uniformity of the questionnaire used during the study, we calculated Cronbach's alpha for each dimension of the questionnaire. We calculated these values for the RS-none condition which is the closest of based condition. These values can be found in Tab. 5.3. To consider that the coherence of a dimension is validated, alpha should be of 0.7 or higher. We can see that all dimensions of questionnaire (the french version) are validated here.

Dimensions	Cronbach's alpha
Verbal	0.73
Acting	0.85
Collaboration	0.76
Interaction	0.9
Robot perception	0.84

Table 5.3: Cronbach's alpha for the different dimensions of the questionnaire. An alpha of 0.7 and higher means the dimension is validated.

5.5 Conclusion

The aim of this chapter was to evaluate the algorithms presented in the last two chapters in order to improve the Shared Plan elaboration and execution by the robot. The new system, with its two possible modes (negotiation and adaptation) has been compared to a reference system corresponding to the state of the art before the ameliorations (with two possible options for verbalization). The evaluation has been done both in simulation and with a real study in the real robot.

Both evaluations have shown that the new system performs better than the old one. In simulation, the adaptation mode performed a little better than the negotiation mode (a little less verbalizations). However, the naive users during the user study preferred the negotiation mode mainly because it was reassuring to have the robot asking when there was a choice. In conclusion, maybe the negotiation mode should be preferred for first or punctual interactions with the robot, and, when the user becomes more used to the robot, the adaptation mode should be preferred.

Moreover, during the user study, we constructed a questionnaire in order to evaluate users feeling concerning the collaboration with the robot which has been validated (in therm of intern coherence) thanks to the study data. This tool is generic enough to be used for other studies where a robot collaborates with a human.

Bibliography

[Bartneck 2009] Christoph Bartneck, Dana Kulić, Elizabeth Croft and Susana Zoghbi. *Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots.* International journal of social robotics, vol. 1, no. 1, pages 71–81, 2009. (Cited in pages 8 and 9.)

[Brooke 1986] John Brooke. *System usability scale (SUS): a quick-and-dirty method of system evaluation user information.* Reading, UK: Digital Equipment Co Ltd, 1986. (Cited in page 8.)

[Dragan 2015] Anca D Dragan, Shira Bauman, Jodi Forlizzi and Siddhartha S Srinivasa. *Effects of robot motion on human-robot collaboration.* In Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, pages 51–58. ACM, 2015. (Cited in page 8.)

[Fischer 2016] Kerstin Fischer, Lars C Jensen, Stefan-Daniel Suvei and Leon Bodenhagen. *Between legibility and contact: The role of gaze in robot approach.* In Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on, pages 646–651. IEEE, 2016. (Cited in page 8.)

[Gombolay 2015] Matthew C Gombolay, Reymundo A Gutierrez, Shanelle G Clarke, Giancarlo F Sturla and Julie A Shah. *Decision-making authority, team efficiency and human worker satisfaction in mixed human–robot teams.* Autonomous Robots, 2015. (Cited in page 8.)

[Heerink 2009] Marcel Heerink, Ben Krose, Vanessa Evers and Bob Wielinga. *Measuring acceptance of an assistive social robot: a suggested toolkit.* In Robot and Human Interactive Communication, 2009. RO-MAN 2009. The 18th IEEE International Symposium on, pages 528–533. IEEE, 2009. (Cited in page 8.)

[Heerink 2010] Marcel Heerink, Ben Kröse, Vanessa Evers and Bob Wielinga. *Relating conversational expressiveness to social presence and acceptance of an assistive social robot.* Virtual reality, vol. 14, no. 1, pages 77–84, 2010. (Cited in page 8.)

[Hoffman 2013] Guy Hoffman. *Evaluating fluency in human-robot collaboration.* In International conference on human-robot interaction (HRI), workshop on human robot collaboration, volume 381, pages 1–8, 2013. (Cited in page 8.)

[Lallemand 2015] Carine Lallemand, Vincent Koenig, Guillaume Gronier and Romain Martin. *Création et validation d'une version française du questionnaire AttrakDiff pour l'évaluation de l'expérience utilisateur des systèmes*

interactifs. Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology, vol. 65, no. 5, pages 239–252, 2015. (Cited in page 9.)

[Venkatesh 2003] Viswanath Venkatesh, Michael G Morris, Gordon B Davis and Fred D Davis. *User acceptance of information technology: Toward a unified view*. MIS quarterly, pages 425–478, 2003. (Cited in page 8.)

[Weistroffer 2014] Vincent Weistroffer. *Étude des conditions d'acceptabilité de la collaboration homme-robot en utilisant la réalité virtuelle*. PhD thesis, Paris, ENMP, 2014. (Cited in page 9.)