the framework. On the other hand, the Action mechanism executes some functions based on the intentions formed and provided by the Coping mechanism described in this section. We group all of these functions into three categories in our framework. The first group of functions includes all of the functions capable of executing some actions with respect to the domain. The second category includes all of the functions involved in revealing the agent's utterances by writing on the screen or conveying through the agent's voice and text to speech systems. The last category includes all of the functions to express the agent's affective state. The emotions can be expressed through colors, emoticons, voice and text. For example, in the user study described in Chapter 5, we expressed the agent's emotions by using emoticons and utterances through the text on the screen as well as the agent's voice.

We use Disco as the basis of the Collaboration mechanism. Disco is the open-source successor to COLLAGEN [203, 204] which incorporates algorithms based on SharedPlans theory for discourse generation and interpretation. Disco is able to maintain a segmented interaction history, which facilitates the collaborative discourse between a human and a robot.

5.2 Evaluating Appraisal Algorithms (Crowd Sourcing)

In this section, we present a crowd-sourced user study and the results, which we conducted to validate the components of our appraisal processes.

5.2.1 Experimental Scenario

We developed an experimental scenario in which participants were asked to envision a sequence of hypothetical collaborative tasks between themselves and an imaginary friend, Mary, in order to accomplish their shared goal. To minimize the background knowledge necessary for our test subjects, we used a simple domestic example of preparing a peanut butter and jelly sandwich, and a hard boiled egg sandwich for a hiking trip. The tasks did not require the participants to do any deep problem

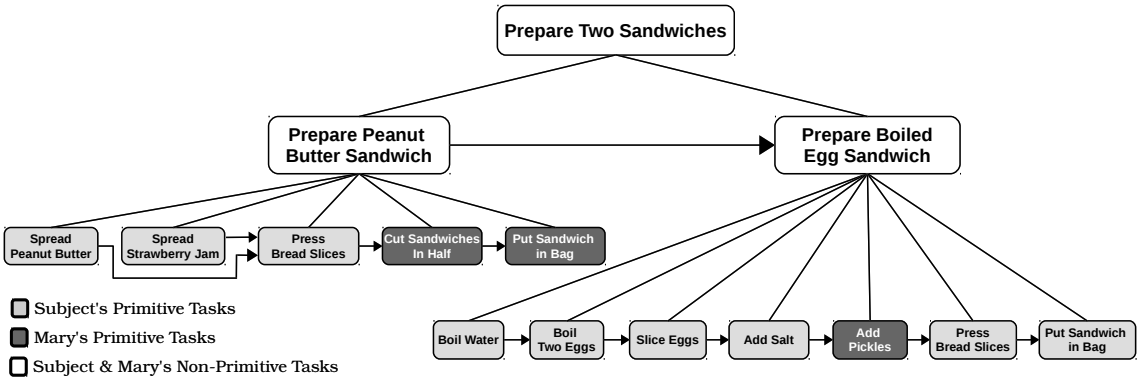

Figure 5.1: Collaboration task model for the evaluation.

solving; rather, the tasks were part of simple daily activities that should be familiar to all participants.

5.2.2 Hypothesis and Methodology

Hypothesis

We conducted this user study to test our hypothesis that humans and our algorithms will provide similar answers to questions related to different factors used to compute four appraisal variables: relevance, desirability, expectedness, and controllability.

Procedure

We conducted a between-subject user study using an online crowdsourcing website – CrowdFlower[1]. We had a questionnaire for each appraisal variable. There were 12 questions (including 2 test questions) in the controllability and expectedness questionnaires, 14 questions (including 2 test questions) in the desirability questionnaire, and 22 questions (including 3 test questions) in the relevance questionnaire.

We provided textual and graphical instructions for all questionnaires; Figure 5.1 shows the corresponding task model[2]. The instructions, provided in the Appendix A, presented a sequence of hypothetical collaborative tasks to be carried out by the

[1]http://www.crowdflower.com
[2]Figure 5.1 was not given to the participants.

Table 5.1: Number of participants

appraisal variables	# of participants
Relevance	29
Desirability	35
Expectedness	33
Controllability	33

test subject and an imaginary friend, Mary, in order to accomplish their goal of preparing two sandwiches. We also provided a simple definition and an example of each appraisal variable. The collaboration structure and the instructions were the same for all questionnaires. The questions introduced specific situations related to the shared plan, which included blocked tasks and failure or achievement of a shared goal. Each question provided three answers which were counterbalanced in the questionnaire. We provided an option like C in all questions (see Figure 5.2), because we did not want to force participants to choose between two options when they did not have a good reason. We derived two questions for different factors involved in each algorithm (see Section 4.3). For instance, we prepared two questions about the influence of the strength of a belief as a key factor involved in relevance algorithm. The questions were randomly placed in the questionnaire. Figure 5.2 shows an example question from the relevance questionnaire which was designed to test whether participants perceive saliency as a factor in relevance. The input for our algorithms was the task model depicted in Figure 5.1.

Participants

Each participant group originally had 40 participants. We limited the participant pools to those with the highest confidence level on the crowdsourcing website in the United States, Britain, and Australia. Test questions were included to check the sanity of the answers. We eliminated participants providing wrong answers to our sanity questions, and participants with answering times less than 2 minutes. The final number of accepted participants in each group is provided in Table 5.1.

5.2.3 Results

Each question in our questionnaires was designed based on different factors that we use in our algorithms (see Section 4.3). For each of the four questionnaires we provide an example question, and describe how each question relates to a specific factor within the corresponding algorithm. The input for our algorithms was the task model depicted in Figure 5.1. The complete list of questions is provided in the Appendix A. Additionally, we provide the p-value for each question, using a binomial distribution, with a probability of success of 0.33, which is the probability of selecting the right answer if the participant is simply guessing.

Expectedness

Figure 5.2 shows an example question from the expectedness questionnaire. In this example, with respect to Algorithm 3 (line 6), option A is more expected because the task related to this option provides the next available task in the focus stack (see the task model in Figure 5.1). Although the task in option B is part of the existing task model, it is considered as UNEXPECTED by our algorithm, since it is not live in the plan. We provided option C to determine whether the participants will differentiate between these two options. This question was presented to the participants to determine whether their decision for the expectedness of this event is similar to the output of the expectedness algorithm. For this question, the human decision was 97% similar to the algorithm's output.

Results for the expectedness questionnaire are presented in Table 5.2. As shown in this table, the results are statistically significant; in fact, for questions 1-6 and 9-10, human participants showed between 67 and 100 % agreement with our algorithms, with p-values of $\ll 0.001$. Questions 7 and 8 were the only two questions that did not show a statistically significant p-value. It should be noted that these questions are comparing equally expected or equally unexpected situations, none of which our algorithms would consider most-expected or most-unexpected.

Table 5.2: Expectedness results (the Equally Expected column indicates for which questions our algorithm provides option C as the response)

Question	Factor	Equally Expected	Percentage of Matching Answers	*p*-Value
1	Live goal vs. Necessary focus shift	No	94%	« 0.001
2	Live goal vs. Not part of shared plan	No	97%	« 0.001
3	Live goal vs. Not part of current branch	No	82%	« 0.001
4	Necessary focus shift vs. Not part of shared plan	No	100%	« 0.001
5	Necessary focus shift vs. Not part of current branch	No	97%	« 0.001
6	Not part of shared plan vs. Not part of current branch	No	73%	« 0.001
7	Live goal	Yes	42%	0.093
8	Not part of current branch	Yes	42%	0.093
9	Necessary focus shift	Yes	67%	« 0.001
10	Not part of shared plan	Yes	88%	« 0.001

Controllability

Figure 5.3 shows an example question from the controllability questionnaire. The algorithm's output is option B, and is determined by Algorithm 4 (line 3), similarly to the expectedness example above. In this example, option B is more controllable than option A, because the self over total ratio of the responsibility of the predecessors of the given task (see *Autonomy* in Section 4.3.4) is higher than the ratio in

Imagine you have pressed the two slices of bread (one covered with strawberry jam and one covered with peanut butter) together and passed it to Mary. Which of the following two actions is **more expected**?

A. Mary puts the given sandwich into a zip lock bag after cutting it in half.

B. Mary puts some pickles on another slice of bread.

C. Equally expected.

Figure 5.2: Example expectedness question.

Table 5.3: Controllability results (the Equally Controllable column indicates for which questions our algorithm provides option C as the response)

Question	Factor	Equally Controllable	Percentage of Matching Answers	*p*-Value
1	Agency	No	85%	« 0.001
2	Autonomy (contributors)	No	52%	0.009
3	Autonomy (predecessors)	No	91%	« 0.001
4	Succeeded predecessors ratio	No	58%	0.001
5	Available inputs	No	91%	« 0.001
6	Agency	Yes	91%	« 0.001
7	Autonomy (contributors)	Yes	73%	« 0.001
8	Autonomy (predecessors)	Yes	55%	0.003
9	Succeeded predecessors ratio	Yes	70%	« 0.001
10	Available inputs	Yes	76%	« 0.001

option A, i.e., self is responsible to spread peanut butter on one slice of bread and strawberry jam on another slice of bread. In this question, the humans decision was 90% in agreement with the algorithm's output.

Results for the controllability questionnaire are presented in Table 5.3. As shown in the table, the p-value is <0.01 for each of the ten questions. The two questions with the lowest human agreement with the algorithms both relate to autonomy (Questions #2 and #8) of the participants with 52% and 55%.

Imagine you want to make a peanut butter sandwich. Which of the following two actions is **more controllable**?

A. You can spread the peanut butter on one slice of bread and you need Mary to spread strawberry jam on the second slice of bread.

B. You can spread the peanut butter on one slice of bread and strawberry jam on the second slice of bread.

C. Equally controllable.

Figure 5.3: Example controllability question.

Table 5.4: Desirability results (the Equally Desirable column indicates for which questions our algorithm provides option C as the response)

Question	Factor	Equally Desirable	Percentage of Matching Answers	*p*-Value
1	Top level goal is failed	No	100%	« 0.001
2	Top level goal is achieved	No	83%	« 0.001
3	Predecessors or preconditions of the top level goal	No	100%	« 0.001
4	Focus is achieved	No	98%	« 0.001
5	Focus is failed	No	100%	« 0.001
6	Predecessors or preconditions of the focus	No	100%	« 0.001
7	Pending or in-progress focus	Yes	46%	0.040
8	Top level goal is failed	Yes	66%	« 0.001
9	Predecessors or preconditions of the top level goal	Yes	54%	0.003
10	Focus is achieved	Yes	57%	0.001
11	Focus is failed	Yes	60%	« 0.001
12	Predecessors or preconditions of the focus	Yes	77%	« 0.001

Desirability

Figure 5.4 shows an example question from the desirability questionnaire. The output based on the Algorithm 2 (line 14) is option C, since in both option A and option B, the focus goal has been achieved successfully. Therefore, in this example, both options A and B are desirable. The humans' decision was 77% in agreement with the algorithm's output in this question.

The results of the desirability questionnaire are presented in Table 5.4. As shown in the results table, the p-value is less than 0.05 for all of the desirability questions. However, an interesting trend is that human participants had a level of agreement of 83%-100% when the algorithm's output selected one alternate as more desirable than another alternate. When the algorithm's output chose option C (i.e. rating

Which of the following two actions is **more desirable**?

A. Imagine you pressed two slices of bread together with peanut butter and strawberry jam on them, and passed them to Mary. Mary cuts the peanut butter sandwich in half and puts them in the zip lock bag.

B. Imagine you want to make the egg sandwich. You have sliced the eggs, put them on one slice of bread, salted them, and waiting for Mary to put some pickles on your eggs. Mary puts some pickles on your eggs.

C. Equally desirable.

Figure 5.4: Example desirability question.

two situations as equally desirable), the human participants only showed 46%-77% agreement. This may indicate that a higher level of granularity is required in the algorithm when evaluating options with similar levels of desirability.

Relevance

In the example shown in Figure 5.5, with respect to Algorithm 1, option A is relevant because of Mary's perceived negative affective state (see Equation 4.1). Although option B is relevant (since it achieves the next goal in the shared plan), 83% of participants consider it as less relevant than option A; we believe this is due to the effect of Mary's perceived negative affective state which also generates a higher utility value in our relevance algorithm. Another question also tested belief saliency. However, the options provided only related to the shared plan (i.e., no human emotions in the options). In this case 87% of participants chose the option that accomplished the next goal in the shared plan. Interestingly, when confronted with a negative affective state from their collaborator, human participants deviated from the shared plan and found their collaborator's affective state more relevant than the original plan. It is noteworthy that in both the absence and the presence of emotions the participants chose the more salient option with respect to our definition of saliency, which was not referenced or provided in the questionnaire.

The complete summary of results for the relevance questionnaire is provided in

Table 5.5: Relevance results (the Equally Relevant column indicates for which questions our algorithm provides option C as the response)

Question	Factor	Equally Relevant	Percentage of Matching Answers	*p*-Value
1	Belief Saliency	No	86%	« 0.001
2	Belief Strength	No	45%	0.063
3	Belief Recency	No	97%	« 0.001
4	Motive Insistence	No	86%	« 0.001
5	Motive Urgency	No	66%	« 0.001
6	Motive Intensity	No	72%	« 0.001
7	Goal Proximity	No	69%	« 0.001
8	Goal Specificity	No	79%	« 0.001
9	Belief Saliency	Yes	90%	« 0.001
10	Belief Strength	Yes	76%	« 0.001
11	Belief Recency	Yes	72%	« 0.001
12	Motive Insistence	No	90%	« 0.001
13	Motive Urgency	Yes	100%	« 0.001
14	Motive Intensity	Yes	100%	« 0.001
15	Goal Proximity	Yes	83%	« 0.001
16	Goal Specificity	Yes	90%	« 0.001
17	Belief Saliency	No	59%	« 0.001
18	Motive Insistence	No	10%	0.995
19	Goal Proximity	No	14%	0.982

.

Table 5.5. As shown in the table, all questions show 59%-100% agreement with our algorithms and statistically significant p-values except for questions 2, 18 and 19. Question 2 addresses belief strength. This question presents a situation in which participants must choose whether a self related goal, or collaborator's goal is more relevant. Questions 18 and 19 address motive insistence and goal proximity, respectively; both of these questions present situations in which participants must choose whether an intense emotional circumstance, or adherence to the collaboration plan is more relevant (the questionnaire is provided in the Appendix A). Our algorithms

Imagine you have made the peanut butter sandwich and passed it to Mary to cut it in half. Which of the following two actions is **more relevant**?

A. Mary starts crying since she cut her finger with a knife.

B. You begin to boil the water to boil the eggs for your second sandwich.

C. Equally relevant.

Figure 5.5: Example relevance question.

choose that the strong emotional circumstance will be more relevant; however, human participants generally selected adherence to the collaboration plan to be more relevant.

5.2.4 Discussion

As shown in the preceding results tables, the human participants agreed 100% on some questions, while on some other questions there was a much lower level of agreement. Our results indicate that people largely performed as our hypothesis predicted. The very small p-values indicate that the data set is not random; in fact, the high percentage of similarity confirms our hypothesis and shows that the algorithms can help us to model appraisal in a collaboration. The very low level of agreement on a handful of questions may indicate algorithm components that require further refinement before implementation; therefore, we made limited changes to our algorithms in light of this study.

5.3 End-to-End System Evaluation

As mentioned earlier, collaborative robots need to take into account humans' internal states while making decisions during collaboration. Humans express affect to reveal their internal states in social contexts including collaboration [35]. Due to the existence of such expressions, robots' affect-awareness can improve the quality of collaboration in terms of humans' perception of performance and preferences. Hence, col-