CHAPTER 4

Design of Experiments

This chapter describes trial and error for designing experiments, then describes the purpose, procedure, and conditions of the five web-based experiments performed. Hypotheses are described in Chapter 5.

Lessons Learned from the Test Experiment

Before conducting the web-based experiments, a test experiment was conducted within the Media Lab to refine the design of the experiment. There were several important lessons about how to design an experiment to get meaningful and significant results.

The first lesson concerned controlling the conditions. The test experiment was designed to test all conditions (the effect of having a face, the difference between gender, realism, humanity, and expressiveness) with only one experiment. The poker game had four poker

playing agents which were randomly chosen out of 18 combinations.[1] The results were hard to analyze, since there were too many factors which might affect the users' impressions of the faces. Thus, the web-based experiments were designed to have only one factor per experiment. For example, in order to study the difference between gender, two faces which differ in terms of gender only (Caricature Male and Caricature Female, or Realistic Male and Realistic Female) were used.

Secondly, I found that there was a need for some initial training to let subjects familiarize themselves to the facial expressions of the poker playing agents. During the test experiment, subjects reported that they could not understand the meaning of each facial expression. Thus, the web-based experiments started with familiarization.

The third lesson concerned designing questionnaires. Yes/no questions should not be used in a questionnaire survey in order to have quantitative measurements. Instead of asking yes/no questions, fixed scale questions should be used to maximize the quantitative data. The scale should have odd numbers to permit the neutral opinion. As for the manner of asking questions, questions that are too direct should be avoided. For example, instead of asking, "How intelligent do you think player A is?", "Do you think the player is good at playing poker?" should be used.

The test experiment was useful to make sure whether the poker game was appropriate for the experiment. Since the main experiment was conducted on the World Wide Web, there was a

1. There were six facial representations (Realistic Male/ Realistic Female/Caricature male/Caricature Female/Smiley/ Dog) and three modes of expressiveness (Honest/Deceiving/Stoic). This means that there are 18 possible combinations of representation for the poker playing agents.

need to evaluate whether people consider the poker game as a serious experiment and to make sure subjects respond to the questionnaire. Some subjects were video taped and observed during the test experiment. They took the game as a serious experiment though they enjoyed playing the game. The response rate for the questionnaire was more than 30%. The test subjects made many comments and suggestions on the interface design, which was used to modify the interface before starting the web-based experiments (see Chapter 3 Designing the Interface). The test experiment was also useful to determine the appropriate number of rounds for playing poker and speed of playing in turn.

Design of Web-based Experiments

Purpose

The first purpose of the experiments is to understand the effect of using faces and facial expressions in the interface of the poker game environment. The second purpose is to analyze whether people's impression of an agent is determined by its representation or by its performance or both. All agents use exactly the same algorithm for playing poker, which the subjects were not aware of.

The questions tested in the experiments are as follows:

- Do people pay attention to a face or facial expressions of an agent?
- Are people distracted by the face or facial expressions?
- Does having a face increase people's engagement to a task?

- Do people use information from the face to interpret the agent's behavior?
- What kind of facial features make the agent look intelligent, likable, and comfortable to interact with?
- Does people's impression of an agent is determined based on its representation or on its performance or both?
- Does people's opinion about personification affect their impression of the faces?
- Is there any difference between the subject's gender in their impression of faces?

The purpose of conducting experiments on the World Wide Web is to have as many subjects with various backgrounds as possible in a condition where software agents are mostly used.

Subjects

Intended subjects are people who accessed the poker game site on the World Wide Web. Anyone who knows the URL (http://poker.www.media.mit.edu:81/epoker) can play the poker game. The poker game site was advertised to two poker related news groups, added to the Gamelan Java directory, Yahoo, the Media Lab web page, and the MIT fun stuff web page.

Procedure

A typical scenario of an experiment is as follows:

- A subject accesses the poker web site.

- The subject reads the consent form and instructions. The instructions contain the poker playing rules, how to play, and the procedure (see Appendix 2 for more details).
- The subject agrees to participate in the experiment and submits an electronic agreement form.
- The subject sees the poker playing agents faces to familiarize himself/herself with the agents' faces and facial expressions.
- The subject's name, email address, and login time are sent to the server and stored in a disk.
- The poker java applet is downloaded to the subject's computer.
- The subject plays a poker game with one out of five conditions for 15 rounds, which takes from 10 to 15 minutes to complete.
- The subject's and agents' hands, actions, and the ending moneys are logged into the server.
- A questionnaire appears after playing 15 rounds[1].
- The subject answers and submits the questionnaire.
- The subject's answers are stored in the server.
- The subject can continue another poker game with one out of five conditions or quit the game.

1. Prior to the web-based experiment, a test experiment was conducted to observe how long it takes for subjects to master how to play the poker game. According to my observation and subjects' comments, it takes about 5 rounds to become familiar with the poker game. I assume playing for 15 rounds is long enough to observe the opponents' strategies, but not long enough to get bored of playing the game.

The poker game was designed so that subjects can continue playing the poker game with another condition only after submitting their answers. This is to encourage subjects to answer the questionnaire[1]. Conditions are randomly assigned to each subject. Appendix 2 shows the instructions presented to the subjects.

Conditions for the Web-based Experiments

The poker game contains five experiments. Table 2 shows the facial representations used in each experiment. There are two or three poker playing agents in each experiment. All facial images used in the experiments have the same size, same resolution, and the same background color. Except for experiment 5, all poker playing agents are in Honest mode.

Experiment 1 (FACE experiment) is designed to study the effect of having a face or not. There are two poker playing agents in this experiment. Their representations are the Caricature Female face which has a white square background, and a white square with the same size which is called “Invisible Man”. Figure 13 shows the poker game environment of experiment 1. Experiment 2 (GENDER experiment) studies the difference in subjects’ impressions of characters’ gender. Once again there are two poker playing agents. The representations used in this experiment are the Realistic Male and the Realistic Female face. Figure 14 shows the poker game environment for experiment 2. Experiment 3 (HUMANITY experiment) is designed to compare the difference in subjects’ impressions between a human face and a non-human face. The representations used are the Caricature Male face

1. Some subjects experienced the same condition and answered the same questionnaire more than once. Only the first answer of those was analyzed in order to design within-subjects (or repeated-measures) experiments. In a within-subject design, a given subject participates in all experimental conditions only once.

and the Dog's face. Figure 15 shows the poker game environment for experiment 3. Experiment 4 (REALISM experiment) uses three faces to compare the effects of three different levels of realism. The representations used are the Realistic Male face, the Caricature Male face, and the Smiley face. Figure 16 shows the poker game environment for experiment 4. Experiment 5 (EXPRESSIVENESS experiment) has three players with the same facial representation. This experiment was designed to evaluate the difference in subjects' impression between three modes of expressiveness. The Caricature Female face with different color of clothes is used to distinguish each face. One agent has Honest expressions, the other has Deceiving expressions, and the last one has Stoic expressions. Figure 17 shows the poker game environment for experiment 5.

TABLE 2. The facial representations used in each experiment

The expressive mode is Honest unless noted.

	Player 1	**Player 2**	**Player 3**
FACE	Caricature Female	White square (Invisible man)	-
GENDER	Realistic Male	Realistic Female	-
HUMANITY	Caricature Male	Caricature Dog	-
REALISM	Realistic Male	Caricature Male	Smiley
EXPRESSIVENESS	Caricature Female with Honest expressions	Caricature Female with Deceiving expressions	Caricature Female with Stoic expressions

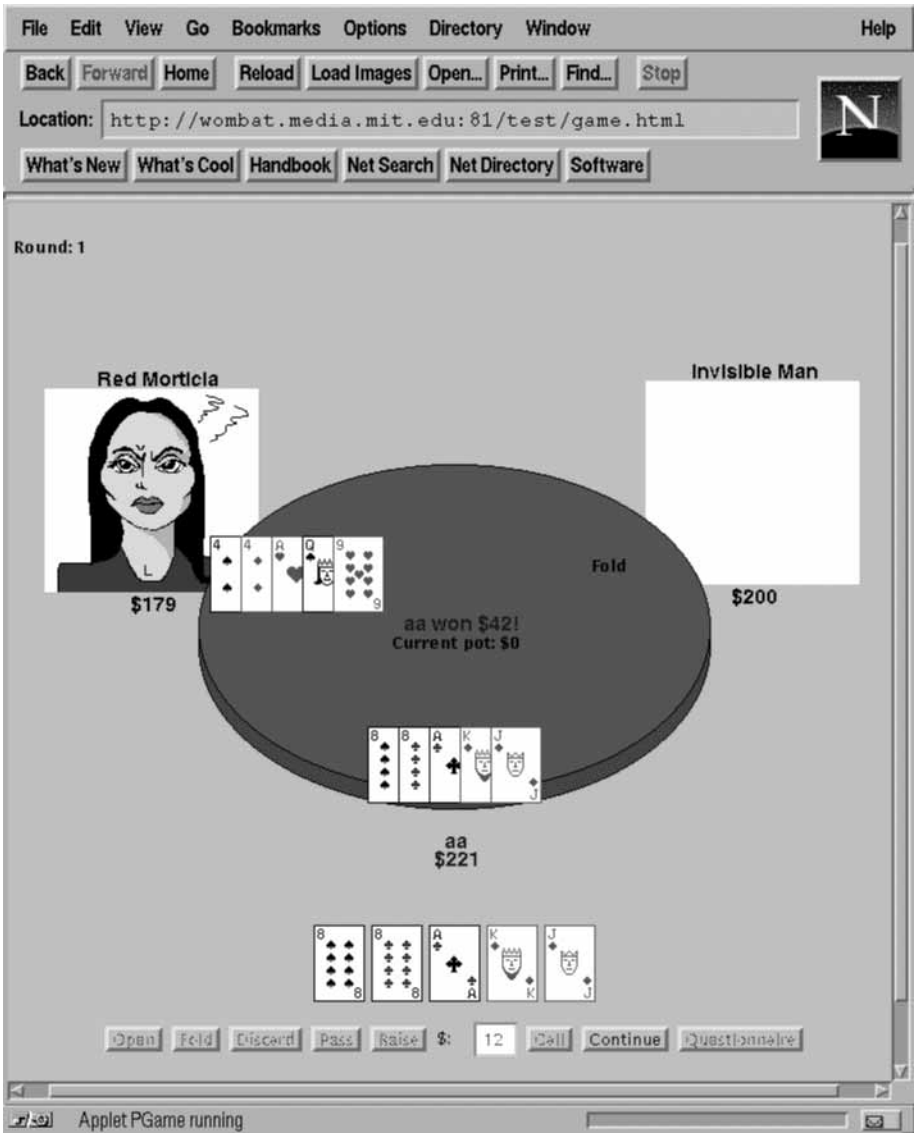

FIGURE 13. **The poker game environment for the FACE experiment.**

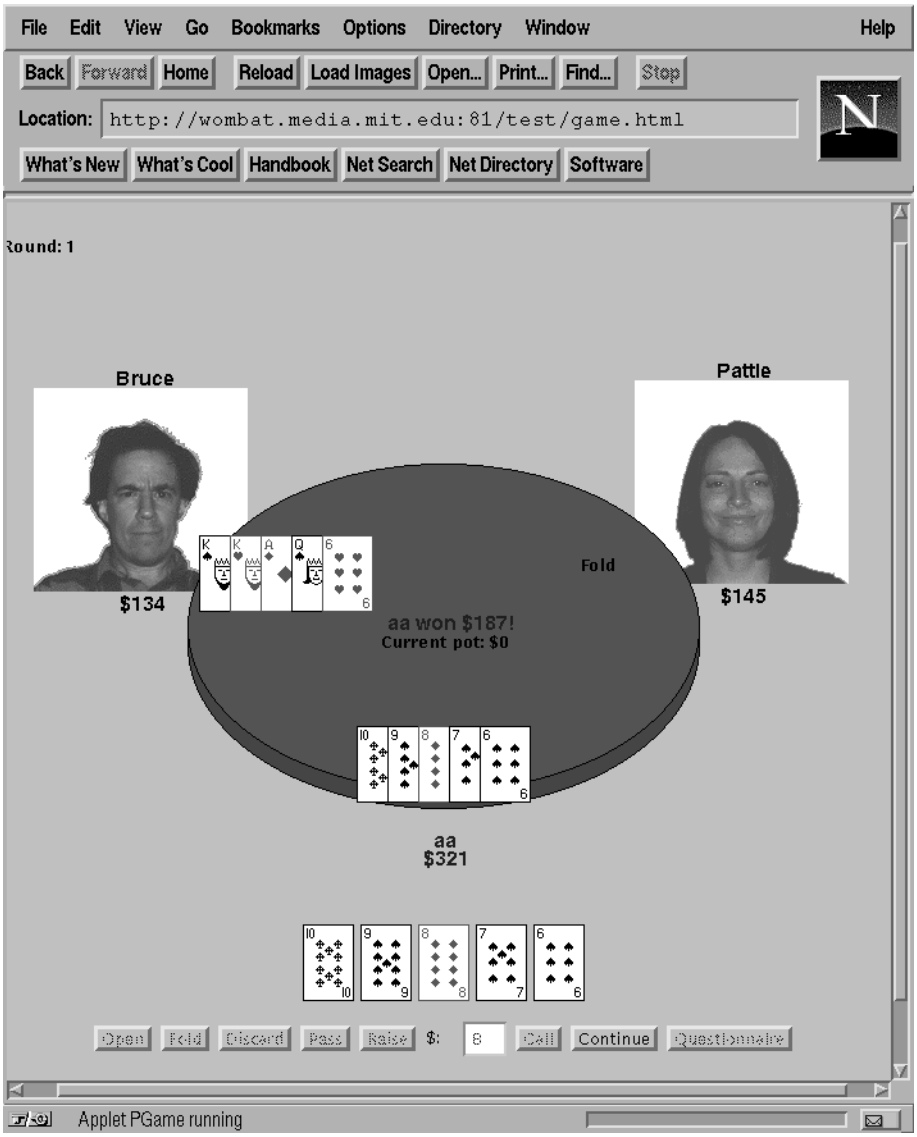

FIGURE 14. **The poker game environment for the GENDER experiment**

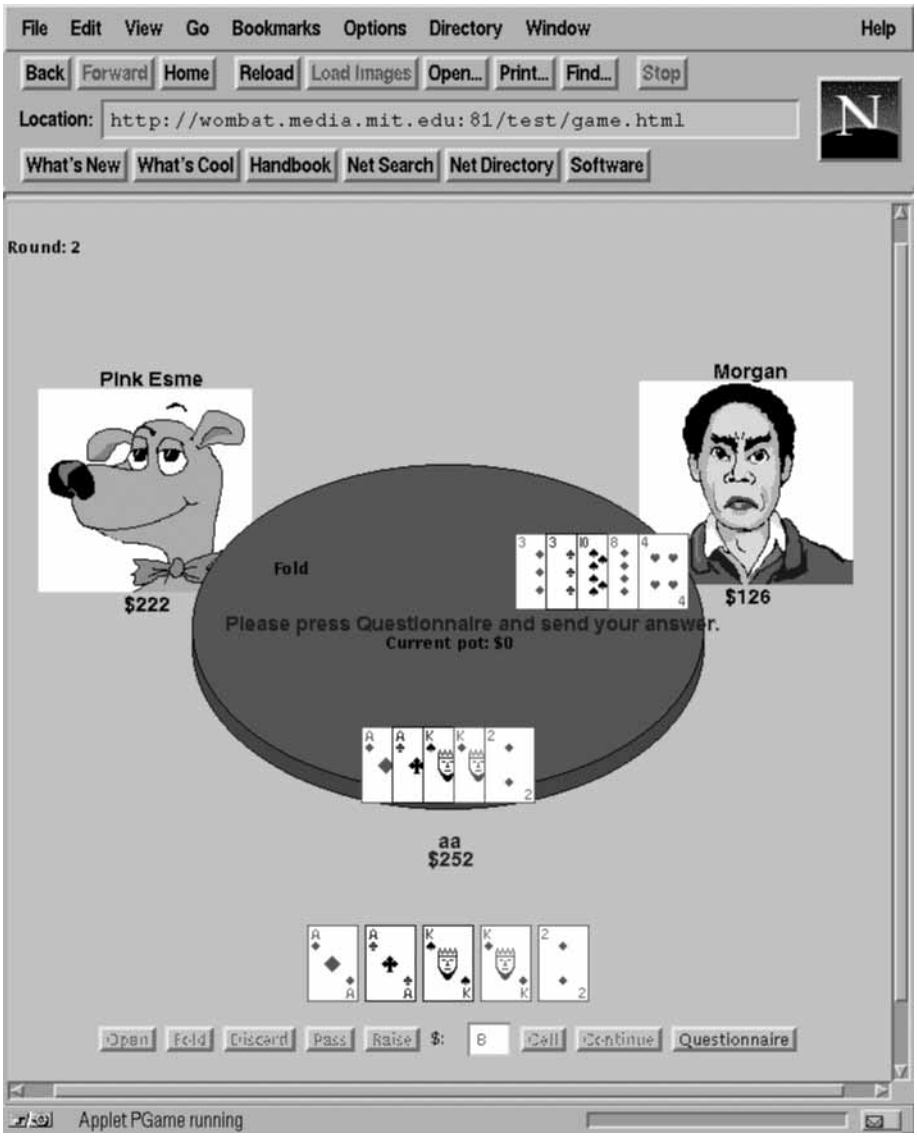

FIGURE 15. The poker game environment for the HUMANITY experiment

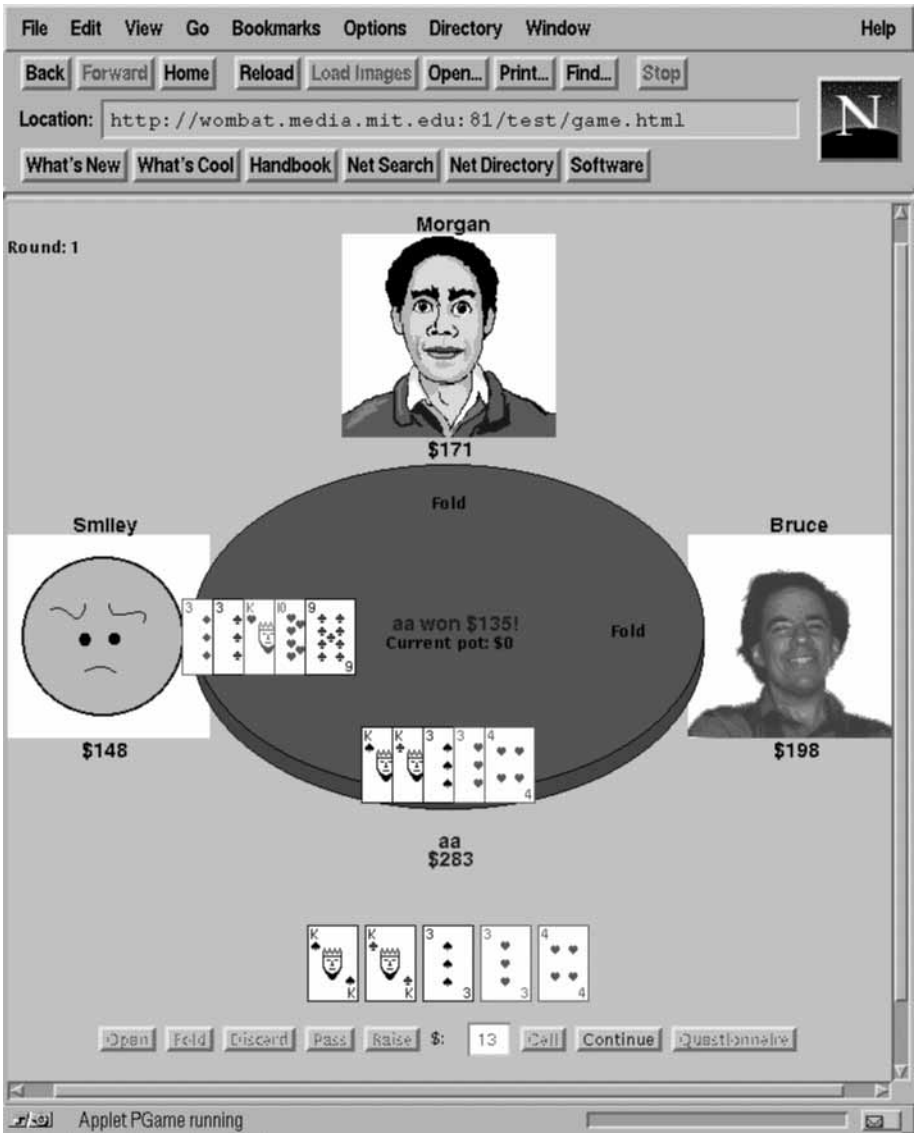

FIGURE 16. The poker game environment for the REALISM experiment

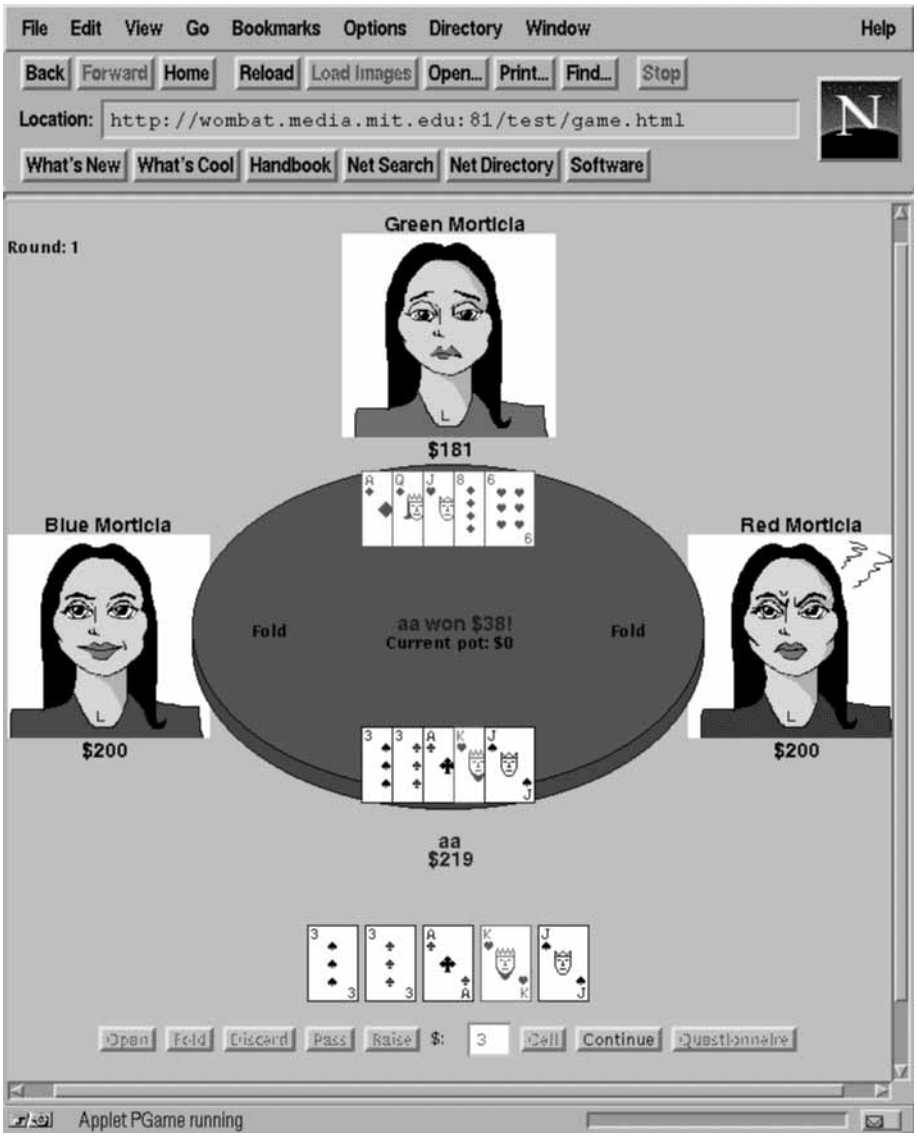

FIGURE 17. The poker game environment for the EXPRESSIVENESS experiment.

Questionnaire

A questionnaire contains 8 to 12 questions. Questions vary according to the experiment. There are 4 background questions (subjects's gender, age, computer expertise, and their opinion about personification), and 4 to 8 questions about the subjects' impressions of the faces against which they played the poker game. The questions about their impressions are seven point scale (1: strongly disagree to 7: strongly agree). Appendix 3 shows all questions presented to subjects.

The common questions across the experiments are perceived intelligence, likability, engagingness, and level of comfort for each face. Intelligence is measured by how good at playing poker each player is perceived to be. Likability is measured by how much the user enjoyed player poker against each player. Engagingness is measured by how engaging playing poker against each player was. Comfortableness is measured by how much the user wants to play poker with the same player.

Specific questions are attention required and level of distraction (for FACE and EXPRESSIVENESS experiments), level of correspondence to the actual intelligence of the player (for HUMANITY and REALISM experiments), believability of the facial expressions, and the usefulness of the facial expressions (for EXPRESSIVENSS experiment). To measure level of distraction, subjects are asked to rate how much they were distracted by the existence of the face or facial expressions with 7 point scale. Level of correspondence is measured by how suitable each face was in representing the actual poker playing skills of each agent. Believability is measured by how much the user believed each player was honest in its facial expressions about its poker hands. Usefulness is measured by how helpful the

facial expressions of each player were to understand the agent's strategy. Table 3 shows the experiment name and the questions asked in each experiment.

TABLE 3. The questions asked in each experiment

experiment name	**FACE**	**GENDER**	**HUMANITY**	**REALISM**	**EXPRESSIVENESS**
conditions **dependent measurements**	**Face/ NoFace**	**Male/ Female**	**Human/Dog**	**Smiley/ Caricature/ Realistic**	**Honest/Deceiving/ Stoic**
Intelligence	Y	Y	Y	Y	Y
Likability	Y	Y	Y	Y	Y
Engagingness	Y	Y	Y	Y	Y
Comfortableness	Y	Y	Y	Y	Y
Required attention	Y				Y
Distraction	Y				Y
Correspondence		Y	Y	Y	
Believability					Y
Usefulness					Y

Pilot Experiment

A pilot experiment was conducted before starting a web-based experiment. The purpose of the pilot experiment was to collect subjects' impressions of each face based solely upon its visual appearance. Seventeen MIT undergraduate and graduate students participated in the pilot experiment. Subjects were shown one of the sets of faces used in the GENDER, HUMANITY, and REALISM experiments with a neutral expression. The name of the experiments are GENDER_LOOK, HUMANITY_LOOK, and REALISM_LOOK, respec-

tively. The subjects answered questions about each face's intelligence, likability, and engagingness as a opponent poker player, just by looking at the facial images. The questions asked in the pilot experiments are shown in Appendix 3.