



[Keiki Takadama, Tetsuro Kawai and Yuhsuke Koyama \(2008\)](#)

Micro- and Macro-Level Validation in Agent-Based Simulation: Reproduction of Human-Like Behaviors and Thinking in a Sequential Bargaining Game

Journal of Artificial Societies and Social Simulation vol. 11, no. 2 9
<<http://jasss.soc.surrey.ac.uk/11/2/9.html>>

For information about citing this article, click [here](#)

Received: 04-Aug-2007 Accepted: 13-Mar-2008 Published: 31-Mar-2008



Abstract

This paper addresses both micro- and macro-level validation in agent-based simulation (ABS) to explore validated agents that can reproduce not only human-like behaviors externally but also human-like thinking internally. For this purpose, we employ the sequential bargaining game, which can investigate a change in humans' behaviors and thinking longer than the ultimatum game (i.e., one-time bargaining game), and compare simulation results of Q-learning agents employing any type of the three types of action selections (i.e., the ϵ -greedy, roulette, and Boltzmann distribution selections) in the game. Intensive simulations have revealed the following implications: (1) Q-learning agents with any type of three action selections can reproduce human-like behaviors but not human-like thinking, which means that they are validated from the macro-level viewpoint but not from the micro-level viewpoint; and (2) Q-learning agents employing Boltzmann distribution selection with changing the random parameter can reproduce both human-like behaviors and thinking, which means that they are validated from both micro- and macro-level viewpoints.

Keywords:

Micro- and Macro-Level Validation, Agent-Based Simulation, Agent Modeling, Sequential Bargaining Game, Reinforcement Learning

Introduction

- 1.1 The *validation of computational models and simulation results* is a critical issue in agent-based simulation (ABS) ([Axelrod 1997](#); [Moss and Davidsson 2001](#)) due to the fact that simulation results are very sensitive to how agents are modeled. To overcome this problem, several validation approaches have been proposed for social simulations, which are roughly categorized as follows ([Carley and Gasser 1999](#)): (1) *theoretical verification* that determines whether the model is an adequate conceptualization of the real world on the basis of a set of situation experts; (2) *external validation* that determines whether the results from the virtual experiments match the results from the real world; and (3) *cross-model validation* that determines whether the results from one computational model map onto, and/or extend, the results of another model. All these approaches contribute to improving the validation of computational models and simulation results. It should be noted, however, that these approaches typically validate complex social phenomena from the *macro-level* viewpoint (e.g., organizational performance caused by interactions among individual agents). This is simply because such macro-level phenomena are usually of primary interest. However, Gilbert claimed that "to validate a model completely, it is necessary to confirm that both the *macro-level* relationships are as expected and the *micro-level* behaviours are adequate representations of the actors' activity" ([Gilbert 2004](#)). From this viewpoint, few studies conducted both the micro- and macro-level validation in agent-based simulation.
- 1.2 Toward a complete validation of computational models and simulation results, this paper aims at addressing both micro- and macro-level validation in agent-based simulation. For this purpose, this paper starts by comparing several simulation results of *different agents in the same model* with subject experiment results. The point of this approach is to compare *different agents*, which differs from the general model-to-model approach ([Hales et al. 2003](#)) that compares *different models* (e.g., a comparison between *culture models* ([Axelrod 1997](#)) and *Sugarscape* ([Epstein and Axtell 1996](#)) in the work of Axtell et al. (1996)). We employ such an approach for the following reasons ([Takadama et al. 2003](#)): (1) it is difficult to fairly compare different computational models under the same evaluation criteria, since they are developed according to their own purpose; (2) common parts in different computational models are very few in number, which makes it difficult to replicate either computational model with the other; and (3) simulation results are sensitive to even a small modification in a model, which makes it difficult to find the key elements or factors that make simulation results sensitive in a model (i.e., there are several candidates that affect simulation results, which makes it hard to find the most important candidates). Since these difficulties prevent a comparison of computational models and their fair comparisons, we start by comparing the results of ABSs whose agents differ only in *one element* as the first step toward our goal. An example of such elements includes *learning mechanisms* applied to agents. In this paper, different kinds of action selection mechanisms in learning mechanisms (i.e., the ϵ -greedy, roulette, and Boltzmann distribution selections, which are all described in Section 3) are employed for agent modeling.
- 1.3 To address this issue, this paper explores agent modeling that is validated from both the micro- and macro-level viewpoints by comparing several simulation results of different agents with subject experiment results. Precisely, we conduct the simulation of agents with different action selection mechanisms and compare these simulation results with subject experiment results to conduct both micro- and macro-level validation.
- 1.4 This paper is organized as follows. Section 2 explains an example (i.e., the bargaining game) employed in this paper and an implementation of agents is described in Section 3. Section 4 presents computer simulations and Section 5 discusses the validity of agents from both the micro- and macro-level viewpoints. Finally, our conclusions are given in Section 6.

Bargaining Game

Why the bargaining game?

- 2.1 In order to address both micro- and macro-level validation in agent-based simulation as described in the previous section, we focus on *bargaining theory* ([Muthoo 1999; 2000](#)) in *game theory* ([Osborne and Rubinstein 1994](#)) and employ a *bargaining game* ([Rubinstein 1982](#)) where two or more players try to reach a mutually beneficial agreement through negotiations. This game has been proposed for investigating when and what kinds of offers from an individual player can be accepted by the other players. We selected this domain for meeting our goal because it can investigate the change in both the *payoff* of human players from the macro-level viewpoint and the *thinking* of human players from the micro-level viewpoint. In particular, the *sequential bargaining game* is employed because (1) both viewpoints can be investigated longer than in the *ultimatum game* (i.e., one-time bargaining game) and (2) *sequential negotiations* are naturally conducted in general human society instead of *one-time negotiation* (i.e., it is a rare case when a negotiation process ends after only one negotiation).

What is the bargaining game?

- 2.2 To understand the bargaining game, let us give an example from Rubinstein's work ([1982](#)), which illustrated a typical situation using the following scenario: two players, P_1 and P_2 , have to reach an agreement on the partition of a "pie." For this purpose, they alternate offers describing possible divisions of the pie, such as " P_1 receives x and P_2 receives $1-x$ at time t ," where x is any value in the interval $[0,1]$. When a player receives an offer, the player decides whether to accept it or not. If the player accepts the offer, the negotiation process ends, and each player receives the share of the pie determined by the concluded contract. If the player decides not to accept the offer, on the other hand, the player makes a counter-offer, and all of the above steps are repeated until a solution is reached or the process is aborted for some external reason (e.g., the number of negotiation processes is finite). If the negotiation process is aborted, neither player can receive any share of the pie.
- 2.3 Here, we consider the finite-horizon situation, where the maximum number of steps (MAX_STEP) in the game is fixed and all players know this information as common knowledge. In the case where MAX_STEP=1 (also known as the *ultimatum game*), player P_1 makes the only offer and P_2 can accept or refuse it. If P_2 refuses the offer, both players receive nothing. Since a rational player operates the notion that "anything is better than nothing," a rational P_1 tends to keep most of the pie to herself by offering only a minimum share to P_2 . Since there are no further steps to be played in the game, a rational P_2 inevitably accepts the tiny offer.
- 2.4 By applying a backward induction reasoning to the situation above, it is possible to perform a simulation for MAX_STEP>1. For the same reason seen in the ultimatum game, the player who can make the last offer is better positioned to receive the larger share by offering a minimum offer ([Stahl 1972](#)). This is because both players know the maximum number of steps in the game as common knowledge, and therefore the player who can make the last offer can acquire a larger share with the same behavior as in the ultimatum game at the last negotiation^[1]. From this feature of the game, the last offer is granted to the player who does not make the first offer if MAX_STEP is even, since each player is allowed to make at most MAX_STEP/2 offers. On the other hand, the last offer is granted to the same player who makes the first offer if MAX_STEP is odd.
- 2.5 After this section, we use the terms "payoff" and "agent" instead of the terms "share" and "player" for their more general meanings in the bargaining game.

Modeling Agents

Why reinforcement learning agents?

- 3.1 For the bargaining game, we employ *reinforcement learning agents* ([Sutton and Barto 1998](#)) because a lot of research has shown that reinforcement learning agents have a high reproduction capability of human-like behaviors ([Roth and Erev 1995; Erev and Roth 1998; Iwasaki et al. 2005; Ogawa et al. 2005](#)). For example, Roth and Erev compared simulation results of simple reinforcement learning agents with results of subject experiments in several examples ([Roth and Erev 1995; Erev and Roth 1998](#)) revealing that (1) computer simulation using simple reinforcement learning agents can better explain the result of subject experiments than economic theory; and (2) the former approach has greater potential of predicting results than the latter approach. In related work, Ogawa and their colleagues compared simulation results with subject experiment results in *monopolistic intermediary games* ([Spulber 1999](#)), which more real-world complexity than examples addressed in Roth and Erev's works ([Roth and Erev 1995; Erev and Roth 1998](#)), and revealed that simple reinforcement learning agents can reproduce the subject experiment results more precisely than the best response agents and random agents ([Iwasaki et al. 2005; Ogawa et al. 2005](#)).
- 3.2 Since these results are validated from the macro-level viewpoint which means that they are not sufficient for micro-level validation, this paper investigates the reproduction capability of reinforcement learning agents in terms of both micro- and macro-level validation in order to explore validated agents through comparisons of them. The reinforcement learning agents, specifically, Q-learning agents ([Watkins and Dayan 1992](#)) in computer science literature are employed among other famous agents like Roth's learning agents ([Roth and Erev 1995; Erev and Roth 1998](#)) in social science literature. This is because our previous research revealed that Q-learning agents can learn consistent behaviors and acquire sequential negotiation in the sequential bargaining game, while Roth's agents cannot (Roth's agents work well in one-time negotiation) ([Takadama et al. 2006](#)).
- 3.3 Furthermore, an employment of reinforcement learning agents including Q-learning agents is useful for model-to-model comparisons in terms of *transferability* of agents to other domains. This is because the agent's model is very simple which makes it easy to replicate for further analysis, in comparison with conventional models which are generally very complex, ad hoc, and created for their own purpose.

An implementation of agents

- 3.4 This section explains an implementation of reinforcement learning agents in the framework of the sequential bargaining game as follows.

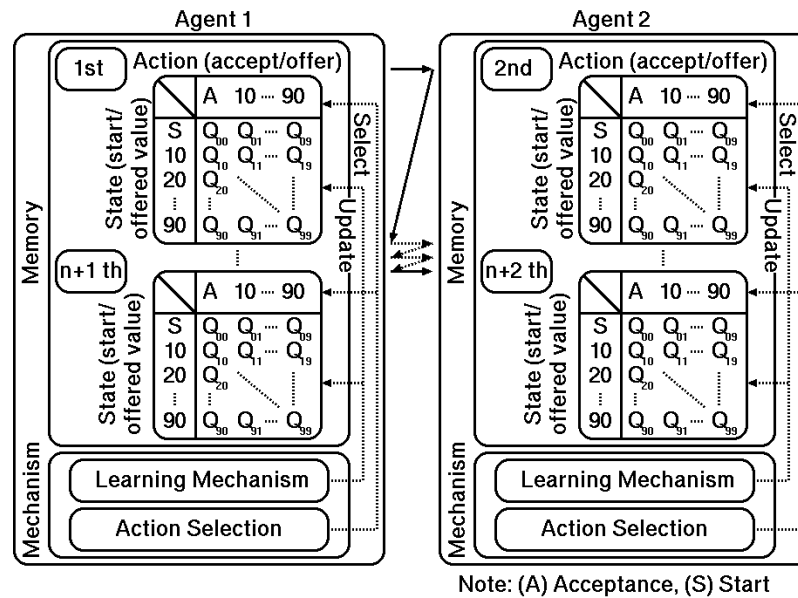


Figure 1. Reinforcement learning agents

• **Memory**

As shown in Figure 1, memory stores a fixed number of matrices of *state* (which represents the start or the offered value from the opponent agent) and *action* (which represents the acceptance of the offered value or the counter-offer value). In particular, the $\text{MAX_STEP}/2 + 1$ number of matrices are prepared in each agent and used in turn at each negotiation to decide to accept an offer or make a counter-offer (see an example presented later in this section). In Figure 1, both agents have $n+1$ number of matrices. In this model, agents independently learn and acquire different worths^[2] of the state and action pair, called *Q-values*, in order to acquire a large payoff. Q-value, represented by $Q(s,a)$, indicates an expected reward (*i.e.*, the payoff in the bargaining game) that an agent will acquire when performing the action *a* in the situation *s*. Note that (1) both state and action in this model are represented by discrete values in units of 10 (*i.e.*, 10, 20, ..., 90); and (2) in addition to these 10-90 values, the matrix has a column labeled (S) and a row labeled (A), which are used to indicate the start to determine the value of the first offer and accept an offer, respectively.

This modeling of memory indicates that the agents decide their actions according the latest offer received. This does not fully represent human behaviors because humans decide their actions according to the offers received not only in the latest but also in the past. However, the agents have multiple numbers of matrices for each negotiation as shown in Figure 1 which implicitly enables agents to consider offers received in the past by employing the mechanism that the current matrix is affected by the previous matrix which is done by equation (1) described in the next.

• **Mechanism**

Q-learning employed in our simulation updates the worth of pairs of state and action by the following equation (1). Variables in this equation are summarized in Table 1.

$$Q(s,a) = Q(s,a) + \alpha [r + \gamma \max_{a' \in A(s')} Q(s',a') - Q(s,a)] \quad \dots (1)$$

$Q(s,a)$	worth of selecting the action (<i>a</i>) at the state (<i>s</i>)
$Q(s',a')$	worth of selecting the next action (<i>a'</i>) at the next state (<i>s'</i>)
<i>r</i>	reward corresponding to the acquired payoff
$A(s')$	a set of possible actions at the next state (<i>s'</i>)
$\alpha (0 < \alpha \leq 1)$	learning rate
$\gamma (0 \leq \gamma \leq 1)$	discount rate

Table 1. Variables in Q-learning

3.5 For the above mechanism, Q-learning mechanism estimates the expected rewards by using the next Q-values, which strengthens the sequential state and action pairs that contribute to acquiring the reward. This is done by updating $Q(s,a)$ to be close to $r + \max_{a' \in A(s')} Q(s',a')$. Precisely, $Q(s,a)$ is close to $\max_{a' \in A(s')} Q(s',a')$ until the final negotiation because *r* is set to 0 due to the fact that the reward is not obtained until the bargaining game is completed, while $Q(s,a)$ is close to *r* at the final negotiation because *r* is set by the acquired reward calculated from the payoff and $\max_{a' \in A(s')} Q(s',a')$ is set to 0 which indicates that there is no further negotiation.

3.6 For the action selection mechanisms that determine the acceptance of an offer or counter-offer, the following methods are employed.

• **ε-greedy selection**

This method selects an action of the maximum worth (Q-value) at the 1-ε probability, while selecting an action randomly at the ε probability (0 ≤ ε ≤ 1) probability.

• **Roulette selection**

This method probabilistically selects an action based on the ratio of Q-values over all actions, which is calculated by the following equation (2).

$$P(a|s) = Q(s,a) / \sum_{a_i \in A(s)} Q(s,a_i) \quad \dots (2)$$

• **Boltzmann distribution selection**

This method probabilistically selects an action based on the ratio of Q-values over all actions, which is calculated by the following equation (3). In this equation, *T* is the temperature parameter that adjusts randomness of action selection. Agents select their actions at random when *T* is high, while they select their greedy actions when *T* is low.

$$P(a|s) = e^{Q(s,a)/T} / \sum_{a_i \in A(s)} e^{Q(s,a_i)/T} \dots (3)$$

- 3.7 As a concrete negotiation process, agents proceed as follows. Defining $\{\text{offer}, \text{offered}\}_i^{A\{1,2\}}$ as the i th offer value (action) or offered value (state) of agent A_1 or A_2 , A_1 starts by selecting one Q-value from the line $S(\text{Start})$ (i.e., one Q-value from $\{Q_{01}, \dots, Q_{09}\}^{[3]}$ in the line S), and makes the first offer, offer_1^{A1} , according to the selected Q-value (for example, A_1 makes an offer of 10% if it selects Q_{01}). Here, we count one *step* when either agent makes an offer. Then, A_2 selects one Q-value from the line $\text{offered}_1^{A2} (= \text{offer}_1^{A1})$ (i.e., one Q-value from $\{Q_{10}, \dots, Q_{19}\}$, where $V = \text{offered}_1^{A2} (= \text{offer}_1^{A1})$). A_2 accepts the offer if Q_{V0} (i.e., the acceptance (A)) is selected; otherwise, it makes a counter-offer, offer_2^{A2} , according to the selected Q-value in the same way as A_1 . This cycle is continued until either agent accepts the offer of the other agent or the negotiation is over (i.e., the maximum number of steps (MAX_STEP) is exceeded by deciding to make a counter-offer instead of acceptance at the last negotiation step).
- 3.8 To understand this situation, let us consider the simple example where MAX_STEP=6 as shown in Figure 2. Following this example, A_1 starts to make an offer of 10% (= offer_1^{A1}) to A_2 by selecting Q_{01} from the line $S(\text{Start})$. However, A_2 does not accept the first offer because it determines to make a 10% (= offer_2^{A2}) counter-offer by selecting Q_{11} from the line 10% (= offered_1^{A2} , corresponding to A_1 's offer). Then, in this example, A_1 makes a 90% (= offer_3^{A1}) counter-offer by selecting Q_{19} from the line 10% (= offered_2^{A1}), A_2 makes a 90% (= offer_4^{A2}) counter-offer by selecting Q_{09} from the line 90% (= offered_3^{A2}), A_1 makes a 10% (= offer_5^{A1}) counter-offer by selecting Q_{09} from the line 90% (= offered_4^{A1}), and A_2 makes a 10% (= offer_6^{A2}) counter-offer by selecting Q_{11} from the line 10% (= offered_5^{A2}). Finally, A_1 accepts the 6th offer from A_2 by selecting Q_{10} from the line 10%, which results in A(acceptance). But, if A_1 makes a counter-offer instead of accepting the 6th offer from A_2 at the last negotiation step (which means to exceed the maximum number of steps), both agents can no longer receive any payoff, i.e., they receive 0 payoff.

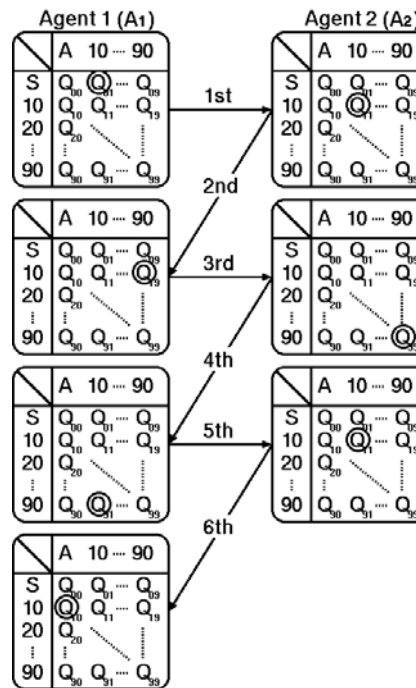


Figure 2. Example of a negotiation process

- 3.9 Here, we count one *iteration* when the negotiation process ends or fails. In each iteration, Q-learning agents update the worth pairs of state and action in order to acquire a large payoff.

Simulation

Simulation cases

- 4.1 The following simulations were conducted in the sequential bargaining game as comparative simulations shown in Table 2.
- **Case 1: Q-learning agents with the constant random parameter**
The aim of case 1 is to explore agent modeling that can reproduce human-like behaviors from the macro-level viewpoint. To address this issue, we compare the results of Q-learning agents, applying one of the three action selection mechanisms (i.e., the ϵ -greedy, roulette, and Boltzmann distribution selection mechanisms) and investigate which agent modeling can reproduce simulation results that are close to subject experiment results. Note that the random parameters, ϵ and T , in the ϵ -greedy and Boltzmann distribution selection mechanisms are set as the constant value (see Section 4.2).
 - **Case 2: Q-learning agents with changing the random parameter**
The aim of case 2 is to explore agent modeling that can reproduce human-like behaviors and thinking from both the micro- and macro-level viewpoints. To address this issue, we compare the results of Q-learning agents, applying either of the ϵ -greedy or Boltzmann distribution selection mechanism with changing the random parameter, ϵ and T , as the following equations (4) and (5) and investigate which agent modeling can reproduce simulation results that are close to subject experiment results. In the following equations, $ChangeRate$ ($0 < changeRate < 1$) indicates the randomness decreasing parameter which is set in Section 4.2.

$$\epsilon = \epsilon \times (1 - \text{ChangeRate}) \text{ in each interaction, } \dots \quad (4)$$

$$T = T \times (1 - \text{ChangeRate}) \text{ in each interaction. } \dots \quad (5)$$

Note that (1) the above equations implicitly represent the thinking of human players (*i.e.*, the micro-level behaviors) according to the subject experiment, which is discussed in Sections 5.1 and 5.6; and (2) we do not conduct the simulation of Q-learning agents with the roulette selection mechanism because there is no random parameter in this mechanism.

	ϵ -greedy selection	Roulette selection	Boltzmann distribution selection
Q-learning with the constant random parameter	Case 1-a	Case 1-b	Case 1-c
Q-learning with changing the random parameter	Case 2-a	—	Case 2-c

Table 2. Simulation cases

Evaluation criteria and parameter setting

- 4.2** In each simulation, (a) the payoff for two agents and (b) the negotiation process size are investigated. Here, the negotiation process size is the number of steps until an offer is accepted or MAX_STEP if no offer is accepted. All simulations are conducted for up to 10,000,000 iterations, which is enough for the agents to learn appropriate behaviors, and the results show the moving average of 10,000 iterations, which are all averaged over 10 runs. As for the parameter setting, the variables are set as follows.
- **Common parameters of the game:** MAX_STEP (the maximum number of steps in one iteration) is 6; reward r (the maximum payoff) is 10; ϵ (the ϵ -greedy selection) is 0.25 in case 1 and 0.9 in case 2; T (the Boltzmann distribution selection) is 0.5 in case 1 and 1000 in case 2; and *ChangeRate* is 0.000001.
 - **Q-learning parameters:** α (learning rate) is 0.1; γ (discount rate) is 1.0; and initial Q-value is 0.1.
- 4.3** Note that (1) preliminary examinations found that the tendency of the results does not drastically change according to the above parameter setting. We have confirmed, in particular, that the results do not drastically change when varying the sensitive parameter ϵ and T around 0.25 and 0.5, respectively; (2) we have confirmed in case 2 that $\epsilon (=0.9)$ in the ϵ -greedy selection and $T (=1000)$ in the Boltzmann distribution selection show mostly the same high randomness in the action selection; and (3) *ChangeRate* in case 2 is set as 0.000001 to reduce the randomness of agents' behaviors around the end of simulations (*i.e.*, 10,000,000 iterations).
- 4.4** Finally, all simulations were implemented by Java with standard libraries and conducted in Windows XP OS with Pentium 4 (2.60GHz) Processor^[4].

Simulation results

- 4.5** Figures 3 and 4 show the simulation results of Q-learning agents with the constant random parameter and those with changing the random parameter, respectively. The upper, middle, and lower figures in Figure 3 correspond to cases 1-a, 1-b, and 1-c, respectively, while the upper and lower figures in Figure 4 correspond to cases 2-a and 2-c, respectively. The left and right figures in all cases indicate the payoff and negotiation process size, respectively. The vertical axis in these figures indicates these two criteria, while the horizontal axis indicates the iterations. In the payoff figure, in particular, the red and skyblue lines indicate the payoff of agents 1 and 2, respectively. Finally, all results are averaged from 10 runs at 10,000,000 iterations. The variances across the 10 runs of all simulation results are less than 0.3 in both payoffs and the negotiation process size, which is enough to be small, *i.e.*, the simulation results across the 10 runs are consistent.

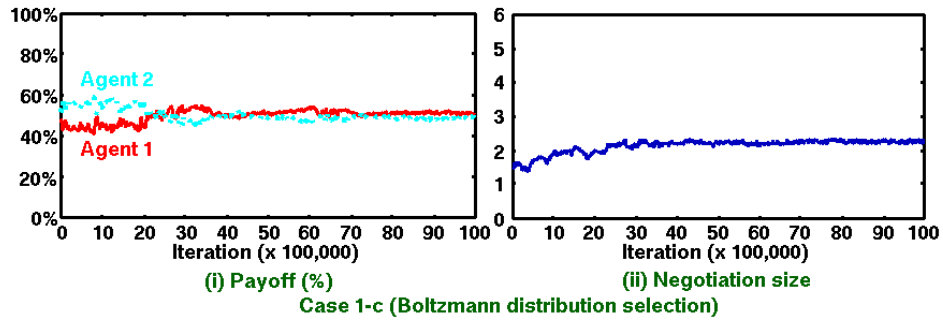
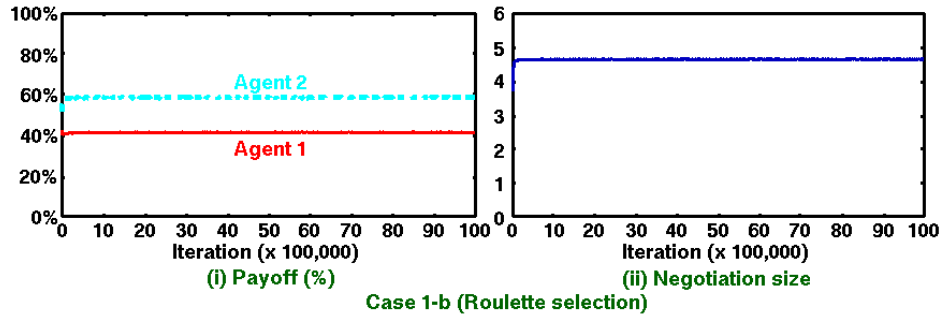
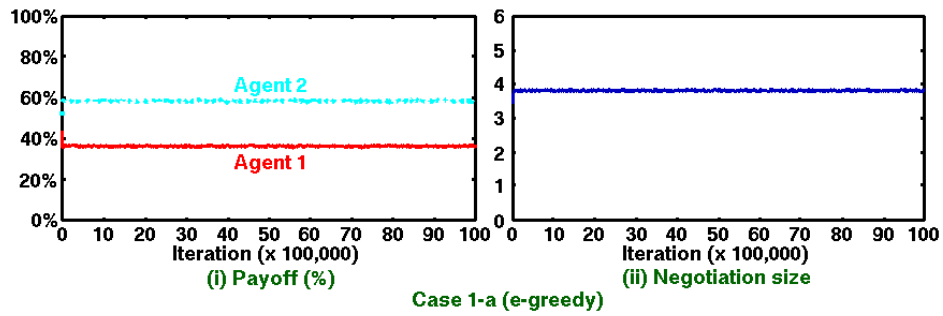


Figure 3. Simulation results of Q-learning agents (Case 1): Average values over 10 runs through 10,000,000 iterations

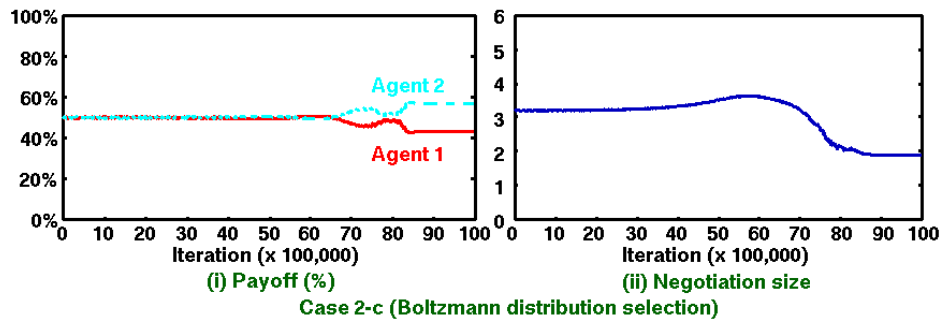
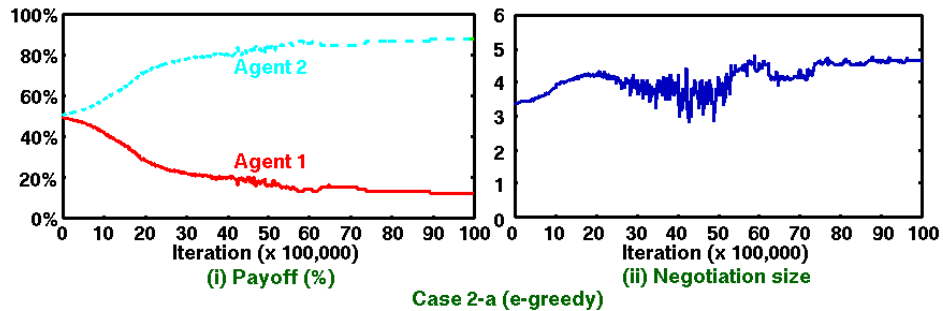


Figure 4. Simulation results of Q-learning agents (Case 2): Average values over 10 runs through 10,000,000 iterations

4.6 These results suggest that (1) in case 1, the payoff of Q-learning agents with the constant random parameter converges within 40% and 60% in any type of the three action selections (*i.e.*, the ϵ -greedy, roulette, and Boltzmann distribution selections), while the negotiation of those agents is more than two-time negotiations in any type of the three action selections, which means that the agents acquire the sequential negotiation; and (2) in case 2, the payoff of the Q-learning agents with changing the random parameter mostly converges at 10% and 90% in the

ε-greedy selection and converges within 40% and 60% in Boltzmann distribution selection, while the negotiation process size of those agents tends to increase as a whole in the ε-greedy selection and shows the increasing and decreasing trend in Boltzmann distribution selection.

Discussion

Subject experiment result

5.1 Before discussing the simulation results of Q-learning agents, this section briefly describes the subject experiment result found in Kawai et al. (2005). Figure 5 shows this result indicating the payoff of two human players in the left figures and the negotiation process size in the right figures. The vertical and horizontal axes have the same meaning as in Figures 3 and 4. In the payoff figure, in particular, the red and skyblue lines indicate the payoff of human players 1 and 2, respectively. Note that (1) all values in this figure are averaged from 10 cases through 20 iterations, where 10 cases were done by 10 pairs created from 20 human players. The variances across the 10 cases of all experiment results are less than 0.9 in both payoffs and the negotiation process size, which is a little larger than that in simulation results but the experiment results across the 10 cases are mostly consistent; (2) all human players are not well aware of the bargaining game and can make an offer or counter-offer either of 10%, 20%, ..., 90% or accept an offer (which is the same as simulations); (3) human players negotiated not face-to-face but by *Windows Messenger* in order not to know each other and to avoid the influence of facial emotion. For this purpose, human players participated in the bargaining game in separated rooms; and (4) human players received one actual payoff from 1 iteration (game) selected from among 20 iterations (games). By informing human players of this reward decision rule before starting the bargaining game, they were motivated to concentrate on every game because they did not know which game would be selected for determining the payment.

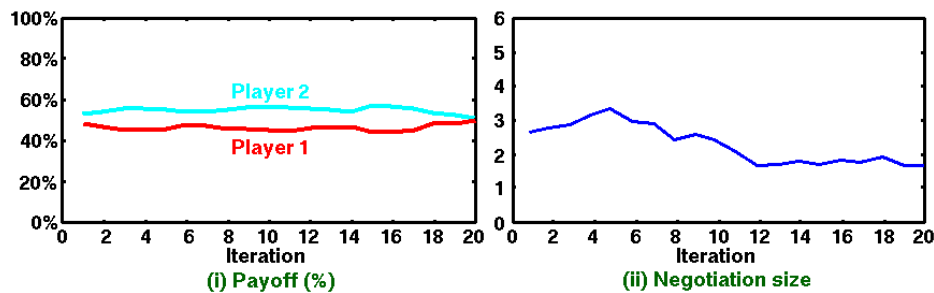


Figure 5. Subject experiment results in (Kawai 2005):
Average values over 10 experiments through 20 iterations

- 5.2 The result shows that (1) the payoff of two human players mostly converges within 40% and 60%, which indicates that human players never accept the tiny offer, unlike the rational players analyzed in Section 2; and (2) the negotiation process size increases a little bit around the first several iterations, decreases and converges around two around the last several iterations, which indicates that (2-i) human players acquire *sequential negotiations*; and (2-ii) the *increasing and decreasing trend* occurs in the subject experiment. This result also differs from the theoretical analysis done by (Rubinstein 1982), which indicates that the rational players (calculate and) offer the optimal payoff (*i.e.*, the minimum payoff) and accept the offer right away without any further negotiation.
- 5.3 To analyze the reason why we obtained the above results, we conducted a questionnaire survey of the human players. The results are summarized as follows: (1) human players that can make the last offer come to be aware of their advantage of having a chance to acquire a larger payoff; (2) human players search for a mutually agreeable payoff not by one-time negotiation but by sequential negotiations; and (3) the increasing and decreasing trend emerges because of the following reasons: (3-i) the negotiation process size increases around the first several iterations because both players do not know their strategies each other which promotes them to explore possibilities of obtaining a larger payoff by competing with each other which requires further negotiations (*i.e.*, a larger negotiation process size is required to explore a larger payoff) and (3-ii) the negotiation process size decreases around the last several iterations because both players find a mutually agreeable payoff by knowing their strategies each other which decreases the motivation of human players to negotiate again (*i.e.*, a few negotiation process size is enough to determine their payoffs). These results suggest that the trend change of the negotiation process size (*i.e.*, the increasing trend to the decreasing trend) represents the change in thinking of human players (*i.e.*, thinking for exploring a larger payoff to thinking for a mutually agreeable payoff).

Case 1: Q-learning agents with the constant random parameter

- 5.4 Regarding Q-learning agents with the constant random parameter, Figure 3 shows that (1) the payoff of two agents converges within 40% and 60% (one of them is rather close to 50% and 50%) in any type of the three action selections (*i.e.*, the ε-greedy, roulette, and Boltzmann distribution selections); and (2) the negotiation process size of those agents is more than two in any type of the three action selections, which means that the agents acquire the sequential negotiation. We obtain these results for the following reasons:
- **Payoff viewpoint:** Q-learning agents updates their Q-values by estimating the expected reward, but they sometimes select actions probabilistically or randomly. This results in acquiring around 40% to 60% payoffs instead of acquiring the maximum and minimum payoffs^[5]. This corresponds to human behaviors, *i.e.*, a human does not always take the optimal actions.
 - **Negotiation process size viewpoint:** Since Q-values that determine players' actions (*i.e.*, the offer, counter-offer, or acceptance of an offer) are not so much different, the negotiation may continue, *i.e.*, some games are completed by one-time negotiation, while others are completed by the maximum time negotiation. This causes more than two-time negotiations and corresponds to human behaviors, *i.e.*, a human does not always complete the same number of negotiations.

5.5 The above analysis suggests that Q-learning agents can acquire the similar results of the subject experiment described in Section 5.1 from both the payoff viewpoint (*i.e.*, an acquisition of the payoff within 40% and 60%) and the negotiation process size viewpoint (*i.e.*, an acquisition of sequential negotiations). This derives the implication that Q-learning agents with any type of action selections are validated from the macro-level viewpoint.

Case 2: Q-learning agents with changing the random parameter

- 5.6 It should be noted here, however, that the validation described in the previous section is not accurate because Q-learning agents cannot reproduce the increasing and decreasing trend found in the subject experiment from the precise negotiation process size viewpoint. This suggests that the macro-level validation is not enough to explore validated agents. More importantly, the only macro-level validation may derive incorrect implications. To overcome this problem, we should conduct the micro-level validation in addition to the macro-level validation as Gilbert

claimed (Gilbert 2004).

5.7 For this purpose, we focus on the *thinking* of human players from the micro-level validation and consider why the increasing and decreasing trend is occurred in the negotiation process size. Concretely, the change in thinking of human players is investigated from the viewpoint of the negotiation process size because the trend change of the negotiation process size represents the change in thinking of human players which is revealed from the questionnaire survey to human players conducted in the subject experiment in Section 5.1. Repeating the analysis of the subject experiment, such a trend emerges by players competing with each other to obtain a larger payoff around the first several iterations which promotes further negotiations (*i.e.*, the negotiation process size increases), and by finding a mutually agreeable payoff around the last several iterations which decreases the motivation of human players to negotiate again (*i.e.*, the negotiation process size decreases). Considering these characteristics of human players, we introduce the *randomness decreasing parameter* in equations (4) and (5) described in Section 4.1. This parameter decreases the randomness of the action selection of human players as the iterations increase, which have the following functions: (1) the high randomness of the action selection in the first several iterations corresponds to the stage where players try to explore a larger payoff by competing with each other; while (2) the low randomness of the action selection in the last several iterations corresponds to the stage where players make a mutually agreeable payoff with a small number of negotiations.

5.8 By using Q-learning agents with the above randomness decreasing parameter, we conducted the simulation and acquired Figure 4 showing that (1) the payoff of agents employing the ϵ -greedy selection converges mostly at the maximum and minimum payoffs, while that of agents employing Boltzmann distribution selection converges within 40% and 60%; and (2) the negotiation process size of agents employing the ϵ -greedy selection increases as a whole tendency although it sometimes vibrates, while that of agents employing the Boltzmann distribution selection shows the increasing and decreasing trend. We obtain these results for the following reasons.

- **Payoff viewpoint:** When the random parameter is high, both Q-learning agents employing the ϵ -greedy and Boltzmann distribution selections explore their offer or counter-offer values randomly with the high value of ϵ and T . When the random parameter becomes low, however, the agents employing the ϵ -greedy selection become to select the best action, while the agents employing Boltzmann distribution selection become to select actions considering the past experience (*i.e.*, the better actions are selected in a high probability, while the worse actions are selected in a low probability). This difference derives the implication that agents employing the ϵ -greedy selection can estimate the expected reward that contributes to acquiring mostly the maximum and minimum payoffs, while those employing Boltzmann distribution selection cannot estimate the expected reward that results in acquiring around 40% to 60% payoffs like human players. The latter result (*i.e.*, the result of agents employing Boltzmann distribution selection) corresponds to human behaviors, *i.e.*, humans do not always select their best actions but select them considering the past experience.
- **Negotiation process size viewpoint:** The above difference between the best action selection in the ϵ -greedy selection and the learned actions selection in Boltzmann distribution selection also causes the different Q-tables after 10,000,000 iterations as shown in Table 3(a) and (b). In both tables, the column and line indicate the *action* (*e.g.*, "acceptance" represented by A or "offer/counter-offer value") and *state* (*e.g.*, "start" represented by S or "offered value"), respectively. For example, the Q-value of counter-offering 10% is 8.1 when an opponent agent offers 10% in Table 3(a). The tables indicate that (1) agents employing the ϵ -greedy selection continue to make a 10% or 20% counter-offer in a high probability because the highest Q-values (*i.e.*, 8 (counter-offering a 10% payoff) or 8.1 (counter-offering a 20% payoff) in Table 3(a)) is usually selected, which contributes to increasing the negotiation process size as shown in Figure 4(a); and (2) agents employing Boltzmann distribution selection, on the other hand, make a 50% offer and accept it in a high probability because the highest Q-values (*i.e.*, 5 (counter-offering a 50% payoff) in Table 3(b)) is usually selected (precisely, agents may make a 50% counter-offer with the same probability of the acceptance of the offer because both Q-values are 5). This contributes to decreasing the negotiation process size as shown in Figure 4(b). This directly corresponds to human behaviors, *i.e.*, humans behave under the consideration of fairness (or equity) discussed in Section 5.10.

(a) Q-learning agents employing the ϵ -greedy selection

Action State	Acceptance(A) or offer/counter-offer value									
	A	10	20	30	40	50	60	70	80	90
S	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
10	1	8.1	8	7	6	5	4	3	2	1
20	2	8	8	7	6	5	4	3	2	1
Start(S)	30	3	8	8	7	6	5	4	3	2
or	40	4	8.1	8	7	6	5	4	3	2
Offered	50	5	8	8	7	6	5	4	3	2
values	60	6	8	8	7	6	5	4	3	2
	70	7	8	8	7	6	5	4	3.2	2
	80	8	8	8	7	6	5	4	3	2
	90	9	8.2	8	7.1	6.2	5	4	3.1	2.1

(b) Q-learning agents employing Boltzmann distribution selection

Action State	Acceptance(A) or offer/counter-offer value									
	A	10	20	30	40	50	60	70	80	90
S	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
10	1	2.1	2.7	2.6	2.3	5	4	3	2	1
20	2	2.1	2	2.3	3.1	5	4	3	2	1
Start(S)	30	3	2.6	2.7	2.5	2.1	5	4	3	2
or	40	4	2.3	2.8	2.5	2.6	5	4	3	2
Offered	50	5	2.6	3.1	2.6	2.8	5	4	3	2
values	60	6	3.5	3.2	2.9	3.1	3.9	4.1	3	2
	70	7	3.6	3.6	3.1	4.3	4.7	3.8	3	1.9
	80	8	1.4	0.2	1	1.4	1.3	1.6	0.6	1.2
	90	9	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.2

Table 3. Q-table in Q-learning agents

5.9 The above analysis suggests that Q-learning agents employing Boltzmann distribution selection with changing the random parameter can acquire the results similar to the subject experiment described in Section 5.1 from both the payoff viewpoint (*i.e.*, an acquisition of the payoff within 40% and 60%) and the negotiation process size viewpoint (*i.e.*, an acquisition of both sequential negotiations and the increasing and

decreasing trend). This is because the agents employing Boltzmann distribution selection with changing the random parameter become to select their actions considering the past experience, while the agents employing the ϵ -greedy selection with changing the random parameter become to select the best action (which is not usual human behaviors). It goes without saying that the agents employing the roulette section cannot reproduce the increasing and decreasing trend in the negotiation process size due to a lack of the random change parameter, while the agents employing both ϵ -greedy and Boltzmann distribution selections with changing the random parameter can reproduce such trend change. This derives the implication that Q-learning agents employing Boltzmann distribution selection with changing the random parameter are validated in the sequential bargaining game from both the micro- and macro-level viewpoints. This result stresses the importance of both the micro- and macro-level validation in an exploration of validated agents.

Validity of Q-learning agents

5.10 The above analysis derives the following implications: (1) the macro-level validation is not sufficient to explore validated agents, which suggests that the micro-level validation should be conducted in addition to the macro-level validation; and (2) Q-learning agents employing Boltzmann distribution selection with changing the random parameter are validated in the sequential bargaining game from both the micro- and macro-level viewpoints. In order to further strengthen the validity of the above Q-learning agents, this section discusses it from the following aspects.

- **Other micro- and macro-level viewpoints**

Other micro- and macro-level viewpoints can be analyzed instead of the payoffs and the negotiation process size evaluated in our simulations. This indicates that Q-learning agents employing Boltzmann distribution selection with changing the random parameter is only validated from the viewpoints of the payoffs and the negotiation process size. Therefore, further investigations should be done from other micro- and macro-level viewpoints to generalize our results.

- **Interaction**

Comparing the iterations between the subject experiment and computer simulation, humans require only 20 iterations to learn consistent behaviors and acquire sequential negotiation, while Q-learning agents require 10,000,000 iterations. It seems that Q-learning agents cannot completely reproduce the human-like behaviors from the iteration viewpoint. This is true if agents should be validated in terms of iteration aspect, but the tendency and consistency of the simulation results are important aspects in such comparisons for the following reasons: (1) it is difficult to fairly compare both humans' and agents' results in terms of iteration aspect due to humans by nature having much higher capabilities than Q-learning agents (*e.g.*, Q-learning agents do not have the capability of modeling opponent players). This requires a lot of learning time for agents in comparison with human players; and (2) when we validate agents in terms of the iteration aspect, we should also consider the time of one iteration in the sequential bargaining game. This is because one iteration in a short consideration time is not the same as one in a long consideration time. For example, human players can consider opponents' actions in future steps in a long consideration time. From this viewpoint, human players have the a lot of time to consider in comparison with agents due to the fact that average time of completing 10,000,000 iterations for agents (less than 1 minute) is smaller than that of 20 iterations in human players (10 minutes ([Kawai et al. 2005](#))). It seems that 10,000,000 iterations in agents is not so large for comparing the results in terms of the time aspects. But, as you can easily imagine, it is also not a fair comparison due to the different capabilities of human players and agents.

From the above difficulty of validating agents in terms of iterations, a comparison of humans' and agents' results in terms of the tendency and consistency is important for the first stage of validation. However, an exploration of agents modeling that produces human-like behaviors in short iterations (like 20 iterations) is the challenging issue to overcome the above validation problem.

- **Fairness (Equity)**

Focusing on the fairness (or equity) of the payoff, Q-learning agents employing Boltzmann distribution selection derive the roughly equal division of the payoff, which is most similar to the subject experiment result. It should be noted here, however, that (1) the Q-learning mechanism *itself* does not consider fairness (or equity) of the payoff because it is an optimization method but (2) the *integration* of the Q-learning mechanism with action selections enables agents to acquire the fairness of behaviors. Especially in the case of introducing the *randomness decreasing parameter* that reflects human behaviors (*i.e.*, (1) the high randomness of the action selection in the first several iterations corresponds to the stage where players try to explore a larger payoff by competing with each other; while (2) the low randomness of the action selection in the last several iterations corresponds to the stage where players make a mutually agreeable payoff with a small number of negotiations), agents acquires 50% offer for any offers from the opponent agents as shown in Table 3(b). Such results cannot be obtained in the case of other action selection mechanisms. In this sense, Q-learning employing the Boltzmann distribution selections has great potential for providing the fairness of behaviors.

This implication can be supported by other research of the bargaining game in the context of *experimental economics* ([Friedman and Sunder 1994](#); [Kagel and Roth 1995](#)). For example, Nydegger and Owen showed that there is a focal point ([Schelling 1960](#)) around the 50% split in the payoff of two players ([Nydegger and Owen 1974](#)); Binmore (1988: 209) suggested that fairness norms evolved to serve as an equilibrium selection criterion when members of a group are faced with a new source of surplus and have to divide it among its members without creating an internal conflict; and the results obtained by Roth et al. showed the fairness even though the subjects playing the ultimatum game had distinct characteristic behaviors depending on their countries of origin (precisely, four different countries: Israel, Japan, USA, and Slovenia) ([Roth et al. 1991](#)).

Model comparison vs. agent comparison

5.11 In general, the model-to-model approach ([Hales et al. 2003](#)) compares *different models* to investigate the validity of "computational models" and "simulation results." This also contributes to promoting transfer of knowledge on "models" by clarifying the limits of the applicability of their "models". In comparison with this approach, our approach compares *different agents* in the same model to validate "agents" and "simulation results" by comparison with subject experiment results. This also contributes to promoting transfer of knowledge on "agents" by clarifying the limits of applicability of their "agents." From this analysis, we find that (1) both approaches pursue the same goal and (2) the only difference is to focus on model-level design (*i.e.*, the framework of the model) or agent-level design (*i.e.*, the framework of the agent). This viewpoint suggests that our approach can be regarded as one of the model-to-model approaches.

Conclusions

6.1 This paper addressed both *micro- and macro-level validation* in agent-based simulation (ABS) to explore validated agents that can reproduce not only human-like behaviors *externally* but also human-like thinking *internally*. For this purpose, we employed the *sequential bargaining game* for the long investigation of a change in humans' behaviors and thinking and compared simulation results of Q-learning agents employing any type of the three types of action selections (*i.e.*, the ϵ -greed, roulette, and Boltzmann distribution selections) in the game. Intensive

simulations have revealed the following implications: (1) Q-learning agents with any type of the three action selections can acquire sequential negotiation, but they cannot show the *increasing and decreasing trend* found in subject experiments. This indicates that Q-learning agents can reproduce human-like behaviors but not human-like thinking, which means that they are validated from the *macro-level* viewpoint but not from the *micro-level* viewpoint; and (2) Q-learning agents employing Boltzmann distribution selection with changing the random parameter cannot only acquire sequential negotiation but also show the *increasing and decreasing trend* in the game. This indicates that the Q-learning agents can reproduce both human-like behaviors and thinking, which means that they are validated from both *micro- and macro-level* viewpoints.

- 6.2 What should be noted here is that these results have only been obtained from one example, *i.e.*, the sequential bargaining game. Therefore, further careful qualifications and justifications, such as analyses of results using other learning mechanisms and action selections or in other domains, are needed to generalize our results. Such important directions must be pursued in the near future in addition to the following future research: (1) an exploration of other ChangeRage settings; (2) modeling agents that produce human-like behaviors in the short iterations (such as 20 iterations as subject experimental results); (3) simulation with more than two agents; (4) an analysis of the case where humans play the game with agents like in (Bosse and Jonker 2005); and (5) investigation of the influence of the discount factor (Rubinstein 1982) in the bargaining game.

Acknowledgements

The research reported here was supported in part by a Grant-in-Aid for Scientific Research (Young Scientists (B), 19700133) of the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan. The authors wish to thank Juliette Rouchier for introducing Gilbert's interesting paper and also thank anonymous reviewers for useful, significant, and constructive comments and suggestions.

Notes

¹ In detail, Rubinstein concluded that the solution for the bargaining game is unique, *i.e.*, it reaches the *perfect equilibrium partition (P.E.P)* under the assumptions that (1) the discount factors are common knowledge to the players and (2) the number of stages (or steps) to be played is infinite (Rubinstein 1982). In other words, in an exchange between rational players, the first offerer should (calculate and) offer the P.E.P; the responder (the opponent players) should then accept the offer right away, making an instantaneous deal with no need of further interaction. Concretely, assuming that players P_1 and P_2 are penalized with discount factors δ_1 and δ_2 , respectively, and P_1 is granted the first offer, the composition of the P.E.P contract is that player P_1 receives a share of the pie that returns her a utility of $U_1 = (1 - \delta_2) / (1 - \delta_1 \delta_2)$, whereas player P_2 gets a share that returns him a utility of $U_2 = \delta_2 (1 - \delta_1) / (1 - \delta_1 \delta_2)$. For values of δ close to 0, the finite-horizon alternating-offers bargaining games give a great advantage to the player making the last offer. In this research, we employ the finite-horizon alternating-offers bargaining game in the case where $\delta_1 = \delta_2 = 0$.

² In the context of reinforcement learning, worth is called "value." We select the term "worth" instead of "value" because the term "value" is used as a numerical number represented in the state and action.

³ At the first negotiation, one Q-value is selected from $\{Q_{01}, \dots, Q_{09}\}$, not from $\{Q_{00}, Q_{01}, \dots, Q_{09}\}$. This is because the role of the first agent is to make the first offer and not to accept any offer (by selecting Q_{00}) due to the fact that a negotiation has not started yet.

⁴ Source code can be downloaded from <http://www.cas.hc.uec.ac.jp/bargaining-game/index.html>.

⁵ In other words, the Q-learning agents get into the local minimum solution.

References

- AXELROD, R. M. (1997), *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration*, Princeton University Press.
- AXTELL, R., Axelrod, R., Epstein J., and Cohen, M. D. (1996), "Aligning Simulation Models: A Case Study and Results," *Computational and Mathematical Organization Theory (CMOT)*, Vol. 1, No. 1, pp. 123-141.
- BINMORE, K. G. (1998), *Game Theory and the Social Contract: Just Playing*, Volume 2, The MIT Press.
- Bosse, T. and Jonker, C. M. (2005), "Human vs. Computer Behaviour in Multi-Issue Negotiation," *First International Workshop on Rational, Robust, and Secure Negotiations in Multi-Agent Systems (RRS'05)*, IEEE Computer Society Press, pp. 11-24.
- CARLEY, K. M. and Gasser, L. (1999), "Computational and Organization Theory," in Weiss, G. (Ed.), *Multiagent Systems - Modern Approach to Distributed Artificial Intelligence -*, The MIT Press, pp. 299-330.
- EREV, I. and Roth, A. E. (1998), "Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria," *The American Economic Review*, Vol. 88, No. 4, pp. 848-881.
- EPSTEIN J. M. and Axtell R. (1996), *Growing Artificial Societies*, MIT Press.
- FRIEDMAN, D. and Sunder, S. (1994), *Experimental Methods: A Primer for Economists*, Cambridge University Press.
- GILBERT, N. (2004), "Open problems in using agent-based models in industrial and labor dynamics," In R. Leombruni and M. Richiardi (Eds.), *Industry and Labor Dynamics: the agent-based computational approach*, World Scientific, pp. 401-405.
- HALES, D., Rouchier J., and Edmonds, B. (2003), "Model-to-Model Analysis," *Journal of Artificial Societies and Social Simulation (JASSS)*, Vol. 6, No. 4, 5 <http://jasss.soc.surrey.ac.uk/6/4/5.html>.
- IWASAKI, A., Ogawa, K., Yokoo, M., and Oda, S. (2005), "Reinforcement Learning on Monopolistic Intermediary Games: Subject Experiments and Simulation," *The Fourth International Workshop on Agent-based Approaches in Economic and Social Complex Systems (AESCS'05)*, pp. 117-128.

- KAGEL, J. H. and Roth, A. E. (1995), *Handbook of Experimental Economics* Princeton University Press.
- KAWAI et al. (2005) "Modeling Sequential Bargaining Game Agents Towards Human-like Behaviors: Comparing Experimental and Simulation Results," *The First World Congress of the International Federation for Systems Research (IFSR'05)*, pp. 164-166.
- MOSS, S. and Davidsson, P. (2001), *Multi-Agent-Based Simulation*, Lecture Notes in Artificial Intelligence, Vol. 1979, Springer-Verlag.
- MUTHOO, A. (1999), *Bargaining Theory with Applications*, Cambridge University Press. MUTHOO, A. (2000), "A Non-Technical Introduction to Bargaining Theory," *World Economics*, pp. 145-166.
- NYDEGGER, R. V. and Owen, G. (1974), "Two-Person Bargaining: An Experimental Test of the Nash Axioms," *International Journal of Game Theory*, Vol. 3, No. 4, pp. 239-249.
- OGAWA, K., Iwasaki, A., Oda, S., and Yokoo, M. (2005) "Analysis on the Price-Formation-Process of Monopolistic Broker: Replication of Subject-Experiment by Computer-Experiment," *The 2005 JAFEE (Japan Association for Evolutionary Economics) Annual Meeting* (in Japanese).
- OSBORNE, M. J. and Rubinstein, A. (1994), *A Course in Game Theory*, MIT Press.
- ROTH, A. E., Prasnikar, V., Okuno-Fujiwara, M., and Zamir, S. (1991), "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study," *American Economic Review*, Vol. 81, No. 5, pp. 1068-1094.
- ROTH, A. E. and Erev, I. (1995) "Learning in Extensive-Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term," *Games and Economic Behavior*, Vol. 8, No. 1, pp. 164-212.
- RUBINSTEIN, A. (1982), "Perfect Equilibrium in a Bargaining Model," *Econometrica*, Vol. 50, No. 1, pp. 97-109.
- SHELLING, T. C. (1960), *The Strategy of Conflict*, Harvard University Press.
- STÅHL, I. (1972), *Bargaining Theory*, Economics Research Institute at the Stockholm School of Economics.
- SPULBER, D. F. (1999), *Market Microstructure - Intermediaries and the theory of the firm-*, Cambridge University Press.
- SUTTON, R. S. and Barto, A. G. (1998), *Reinforcement Learning: An Introduction*, The MIT Press.
- TAKADAMA et al. (2003), "Cross-Element Validation in Multiagent-based Simulation: Switching Learning Mechanisms in Agents," *Journal of Artificial Societies and Social Simulation (JASSS)*, Vol. 6, No. 4, 6. <http://jasss.soc.surrey.ac.uk/6/4/6.html>
- TAKADAMA et al. (2006), "Can Agents Acquire Human-like Behaviors in a Sequential Bargaining Game? - Comparison of Roth's and Q-learning agents -," *The Seventh International Workshop on Multi-Agent-Based Simulation (MABS'06)*, pp. 153-166.
- WATKINS, C. J. C. H. and Dayan, P. (1992), "Technical Note: Q-Learning," *Machine Learning*, Vol. 8, pp. 55-68.

[Return to Contents of this issue](#)

© Copyright Journal of Artificial Societies and Social Simulation, [2008]

