Modeling Human Emotional Intelligence in Virtual Agents

Alexei V. Samsonovich
Krasnow Institute for Advanced Study, George Mason University
4400 University Drive MS 2A1, Fairfax, Virginia 22030-4444
asamsono@gmu.edu

Abstract
A candidate framework for integration of theoretical, modeling and experimental approaches to understanding emotional intelligence is described. The framework includes three elements of a new kind that enable representation of emotional cognition: an emotional state, an appraisal, and a moral schema. These elements are integrated with the weak semantic cognitive map representing the values of emotional appraisals. The framework is tested on interpretation of results obtained in two new experimental paradigms that reveal general features of human emotional cognition, such as the emergence of subjectively perceived persistent roles of individual virtual actors. Implications concern heterogeneous human-robot teams.

Introduction
Integration of disparate theoretical and modeling approaches (Gray, 2007; Laird, 2012; Parisi & Petrosino, 2010; Phelps, 2006; Russell, 1980) to understanding the nature of human emotional intelligence becomes a key challenge in artificial intelligence (Samsonovich, 2012). A grand-unifying theory of emotional cognition seems to be around the corner (Sellers, 2013; Treur, 2013). In development of a generalizing theory based on phenomenology, first steps usually consist in analysis of fundamental logical possibilities and the introduction of conceptual building blocks that together constitute a framework for further analysis. A candidate for this framework is described below. The architecture builds essentially on two prototypes: GMU BICA (Samsonovich and De Jong 2005) and its variation called Constructor (Samsonovich, 2009).

The architecture has six components, or memory systems: working, semantic, episodic, procedural, the value system and the interface virtual environment (Figure 1). Building blocks for these components include schemas, mental states, semantic maps, and various primitives performing specialized functions, implemented as algorithmic automata, neural networks, etc. These building blocks and components are explained in more detail below.

Figure 1. The cognitive architecture.

Figure 1 shows a bird’s eye view of the cognitive architecture. Semantic memory is populated by schemas that are universal units of symbolic representations. Bound instances of schemas fill mental states (I-Now, I-Next, etc.) that populate working and episodic memories. The value system includes semantic maps (also called semantic spaces, or semantic cognitive maps) that give values to appraisals, and drives that are linked to moral schemas. Procedural memory includes specialized primitives. Interface virtual environment can be the medium used for interactions with other agents and the external world – or it can be the virtual environment where the agent is embedded.

Here is a brief overview of the components (Figure 1; see also Samsonovich et al. 2009). Working and episodic memories are populated by mental states, that in turn are populated by bound instances of schemas. Semantic memory is the collection of all available schemas. The value system includes drives that represent primary values and a semantic map that assigns values to all appraisals of schemas and mental states. Procedural memory includes specialized primitives.
The special notion of a schema in the present framework was described earlier (Samsonovich & De Jong, 2005; Samsonovich et al. 2006), and so was the notion of a mental state (Samsonovich et al., 2009). In addition, Samsonovich and Ascoli (2010) described weak semantic cognitive map that is also at the core of the value system. The focus in the present work is on the integration of the above components in computational description of emotional intelligence.

**Emotional Elements**

Emotional cognitive elements are parsimoniously added to the base theoretical framework as three new categories of elements (Figure 2; Samsonovich 2012a, 2012b):

1. an emotional state, understood as an attribute (an appraisal) of a mental state, based on the current situation perceived by the agent in this state;
2. an appraisal (understood as an attribute of an instance of a schema), and
3. a moral schema, that represents a higher-order appraisal (i.e., an appraisal of appraisals) associated with a certain pattern of appraisal values specified as “normal” for this schema.

![Figure 2. UML class diagram of the architecture.](image)

Figure 2 shows essential core UML class diagram of the architecture. Emotional elements (shown in red) include two attributes (Emotional State is an attribute of a Mental State, Appraisal is an attribute of a Schema) and Moral Schemas representing higher-order appraisals.

Thus, in the first case (1), the emotional characteristic is assigned to a mental state as a whole. In the second case (2), the emotional characteristic is assigned to an instance of a schema. This schema may represent another agent or its mental state: then it is an appraisal of the mental state of one agent by another agent. Finally, a moral schema (3) is a special kind of a schema that binds to patterns of appraisals, and produces effects (e.g. actions of the agent) that can change the values of those appraisals.

**Experimental Paradigms and Methods**

Two paradigms were used in this study to assess the framework. They involved human behavioral experiments as well as virtual agent simulations. Paradigm 1 was cooperative shape construction, and Paradigm 2 was random social interactions, all described below. Other proposed paradigms are entertained in Discussion.

**Virtual Agents**

The virtual agents used in both paradigms were implemented in Matlab R2011a following the principles outlined above. While implementations were simplistic and tailored specifically for selected tasks, elements of all components can be identified in them. Schemas of moves (elements of semantic memory) and their activation in working memory were implemented algorithmically. A set of primitives (procedural memory) were used to implement constraints on move selection. Goals represented by drives (the value system) determined probabilities of move selection. Implicitly implemented mental states in working memory represented the agent and the partner(s) at the present moment and the agent at the next moment (these mental states should be labeled “I-Now”, “Partner-Now”, and “I-Next”, respectively). Their essential attribute used in this study was the appraisal of the partner A. All action schemas were also assigned certain appraisal values a that were kept fixed. Episodic memory was used in the shape construction paradigm and represented memory of past actions and their authorship. The interface between the virtual agent and the physical world was mediated by a human operator, who was either the experimenter or the participant.

Emotional elements, in addition to the variable appraisals A of partner mental states and fixed appraisals a of action schemas, included a moral schema that was defined by a set of “normal” values of appraisals, as described below. Appraisals in general are understood here as vectors that take values on the semantic map (a part of the value system). They affect the probabilities of action selection. The main two dimensions of the map represented valence (“like-dislike”) and dominance-arousal (dominant vs. subordinate). Here dominance was merged with arousal. A rationale for merging these dimensions in a simplified version of semantic map comes from, e.g., the widely-used affective database ANEW, in which
dominance and arousal are strongly correlated. Each appraisal was represented by a vector in the semantic map. Appraisals of actors A determined probabilities of action selection based on the match between A and the action appraisal a. And vice versa, appraisals of performed actions determined dynamics of the appraisals of actors.

Participants

Four George Mason University college students, 3 females and 1 male, participated in this study. The ethnic breakdown was as follows: 1 White, 1 Hispanic, and 2 Asian. All students were from Northern Virginia. All students reported English as their native language. In terms of their student status, they were 1 freshmen, 1 sophomore, and 2 juniors. All of the students were full time students. The age range was between 20 and 25. Students were majoring in Psychology, Computer Game Design, Neuroscience and Bioengineering, and one reported undecided major.

Paradigm 1: Cooperative Shape Constructions

In this paradigm, two partners cooperatively construct an abstract shape from wooden blocks, following a given template and a set of rules. Legal moves are specified precisely and therefore can be entered into, and generated by a computer. This allows one of the partners to be a virtual agent. By making their moves in turn, partners interact with each other. Moves are performed in response to partner’s moves and in turn have consequences for the partner, resulting in emotional appraisals by a human participant.

Here the idea is to study dynamics and behavioral effects of the emergent appraisals, that are described with the model. A research task is to be able to replicate the laws of dynamics inferred from human behavior within the agent. Indeed, the virtual agent is inspired by the human mind and models internally emotional elements of cognition and their behavioral effects. This allows for the following research questions to be addressed.

• What are the effects of appraisals on success in collaboration?
• Do intrinsic emotional features of the virtual agent affect human experience and human behavior?
• Can a human and a virtual agent establish social relationships with each other in the process of collaboration?

It is interesting to note that making moves is the only communication channel in this case: the virtual agent has no sensors and no visible appearance that could be used to convey or detect emotional states. This circumstance makes most of the studies of expression and detection of emotions in artifacts only weakly related to the present study. Therefore, measures used here to address research questions included performance measures (e.g., the number of moves used to assemble a complete shape and characteristics of the shape), values of intrinsically generated appraisals in the virtual agent, and subjects’ reports in response to the questionnaire.

Settings and Procedures

Blocks used in the experiment were two sets of Melissa & Doug, LLC, 100 Wood Blocks #481. The set of pieces includes nine kinds of blocks that for convenience were labeled “long”, “short”, “thick”, “arc”, “cylinder”, “cube”, “half”, “small”, and “large” (Figure 3B). Shapes constructed from blocks were constrained by the template shown in Figure 3A (the shape stands vertically). Piece labels together with position numbers from the template were used to enter moves in the computer. Positions 1 through 13 could be filled with any blocks that fit into them, given additional constraints described below. The color of blocks was ignored. Positions 4, 5, 9 and 10 in the template were optional. Additional constraints were the following: (i) nothing could be put on top of the closing pieces: “half”, “small” and “large”, that had to be placed with the convex part or the corner pointing upwards; (ii) any intermediate or final shape must hold physically and be stable; (iii) only one piece can be manipulated during each move.

The goal for the team was to assemble a complete shape. The shape was considered complete when nothing else could be added to it within the template by a legal move, or the position 12 is filled with a closing piece while positions 11 and 13 are empty. Given this goal, the shape should be maximally tall and as symmetrical and nice as possible. While the last two notions were interpreted intuitively by human participants, they had to be formalized for the virtual agent. For this purpose, it was assumed that a shape is symmetrical if it has a vertical axis of approximate symmetry, and it is “nice” if it has some sparseness and/or a sharp top. To include these criteria into the biases that affected decision making in the virtual agent, measures called “symmetry” and “sparseness” reflecting the above intuitive descriptions were defined algorithmically and implemented as primitives in the procedural memory of the agent.

The experimental procedure was the following. The team starts with an empty table and a virtually unlimited number of pieces of each kind. Then the team makes moves, each partner in turn. Legal moves include: add a piece to the shape at a legal position; remove one or two pieces while leaving the rest in place; or skip the move. The process ends when no pieces can be added to the shape given constraints, or the third level is symmetrical and includes closing piece(s), as e.g. in Figure 3 CC1, or the time limit expires. Each sequence of moves was recorded. No misinformation was used in this study.
Virtual Agent
The virtual agent implemented as described above had no physical embodiment, and therefore needed a human operator to read and enter commands and to move pieces following its commands. In some sessions the experimenter played the role of this operator, while the participant was informed that moves were generated by the virtual agent. In other sessions, however, participants themselves performed functions of the operator, entering their moves into the computer and implementing moves generated by the virtual agent.

Appraisal of the partner remained in effect fixed during a session. Only one component of the partner appraisal A was varied across sessions and was addressed by this study. Behaviorally, A affected the probability of only one sort of action: the action of undoing a previous partner’s move. This action can be interpreted as an attempt to dominate by not allowing the partner to implement her intent, or alternatively as an aversive action expressing a negative valence (“dislike” of the partner), or both. In the first interpretation, the corresponding action appraisal a can be understood as a positive dominance, and in the second interpretation it would be a negative valence. More generally, the value of this action appraisal can be viewed as belonging to the upper-left quadrant in the semantic map (Figure 2 in Samsonovich, 2012). While these interpretations may be speculative, model dynamics related to A and a are specified by the implementation: A is directly proportional to the probability of aversive action selection.

Paradigm 2: Random Social Interactions
In this paradigm, participants are first presented with a short demo showing three animated shapes – Circle, Square and Triangle – moving in a square environment. Then participants are asked to keep this scene in mind and, imagining being in control of Circle, engage in interactions with the two other actors in this environment. Interactions are performed by typing commands and reading simple standard messages, e.g., “Triangle greets Circle”. There are four possible actions that actors can perform: hit, yield, greet, and ignore. Actors are performing actions in turn, one after another, choosing themselves what action to perform, while the target of each action is sampled randomly and independently of the actor. This corresponds to random encounters during spontaneous 2-D motion.

The general paradigm involves a group of N agents interacting by means of the 4 possible actions. Any environmental (including spatial) aspects of the interaction are ignored. The interacting actors and their actions are appraised by the virtual agents identically, because they all receive identical information and use identical internal cognitive models. Therefore, one and the same generic working memory model is presumed to correctly describe working memory of any of the participating agents. There are no false beliefs or subjective biases. This also assumes that all agents always “correctly interpret” each other’s motivations. The obvious indexical differences between the N agent perspectives are easily taken into account in derivation of the dynamic equations.

The first task is to observe from simulations what stable patterns of emotional relationships among agents (understood as their mutual appraisals) can emerge in this environment, and how the outcome may depend on the architecture parameters.

Implementation Details
The virtual agent used in this case is implemented as outlined above and is capable of interacting with N-1 other agents by performing the four available actions. Probabilities of action selection are determined by the action appraisals, which take their values on the two-dimensional semantic map (Figure 2), with the two dimensions labeled “valence” and “dominance-arousal”. The map coordinates of the four actions are given by positions of their names on the map (Table 1). Normalized values of the principal components PC1 (valence) and PC2 (dominance, arousal) are taken from the map available as part of the materials of Samsonovich and Ascoli (2010).

There are N mental states in working memory of every agent, each corresponding to the state of awareness of one agent taken at the present moment. Again, the appraisal of a given mental state is the same for all appraisers in this

Figure 3. Cooperative shape construction. (A) The template for the shape. Legal positions for pieces are labeled 1 through 13. (B) The labeling of pieces. (CC1) Two examples of shapes constructed by two virtual agents working cooperatively. (KJ1, KJ2) Examples of shapes constructed by two human participants working cooperatively. (C1, C2, C3m KJcomp, Hm) examples of shapes constructed by a human and a virtual agent working cooperatively.
model. Emotional states and higher-order appraisals are not simulated.

Table 1. Weak semantic cognitive map coordinates for the appraisals of actions (values taken from materials of Samsonovich and Ascoli, 2010).

<table>
<thead>
<tr>
<th>Action</th>
<th>PC1</th>
<th>PC2</th>
</tr>
</thead>
<tbody>
<tr>
<td>hit</td>
<td>0.26</td>
<td>1.07</td>
</tr>
<tr>
<td>yield</td>
<td>0.43</td>
<td>-1.03</td>
</tr>
<tr>
<td>greet</td>
<td>0.93</td>
<td>0.15</td>
</tr>
<tr>
<td>ignore</td>
<td>-1.83</td>
<td>-0.37</td>
</tr>
</tbody>
</table>

Therefore, there are only N dynamic emotional characteristics in this model: appraisals of the N agents. They are initiated to very small random values at the beginning of the simulation epoch, in order to break the symmetry. Each of the four possible actions has a fixed appraisal given in Table 1. Appraisal values $A$ are 2-D vectors that are treated here for convenience of implementation as complex numbers:

$$A = (\text{valence}, \text{dominance}).$$

In this case, $valence = \text{Re}(A)$, and $dominance = \text{Im}(A)$. The simulation epoch consists of a sequence of iterations performed at each moment of discrete time $t$. One iteration includes the following essential steps: (i) compute action probabilities, (ii) select and perform action, (iii) update appraisals of the actor and the target of action. Dynamical equations used here to update the appraisals are:

$$A_{target}^{t+1} = (1 - r)A_{target}^t + rA_{action}$$
$$A_{actor}^{t+1} = (1 - r)A_{actor}^t + rA_{action}^*$$

Here $t$ is the moment of discrete time, and $r$ is a small positive number (a model parameter that was set to 0.01). The likelihood $L$ of a specific action is proportional to

$$L_{action} \sim \left[ \text{Re} \left( A_{action} (A_{actor}^* + A_{target}) \right) \right]_+.$$  

Here $[x]_+$ is equal to the positive values of $x$ and is zero otherwise, $A^*$ is the complex conjugate of $A$. Intuitively, this formula means that the action is more likely to be selected, when its appraisal matches the appraisal of the actor and also matches the appraisal of the target, in which the dominance component is inverted.

Results and Analysis

Paradigm 1: Cooperative Shape Constructions

At the beginning, simulations of two virtual agents interacting with each other were performed. These simulations demonstrated the ability of the architecture to construct appealing shapes satisfying the constraints (Figure 3 CC1). Examples shown in the figure are typical and were among the first valid shapes rather than specifically selected ones.

Sessions with human participants (human – human and human – virtual agent) resulted in similar shapes (Figure 3). In this sense, virtual agents did not demonstrate the lack of domain expertise, compared to human participants. However, on average it took fewer moves for two human partners to assemble a complete shape, compared to a team including one or two virtual agents.

Typical individual sessions of shape construction by homogeneous and heterogeneous teams are represented by data in Figs. 7, 8. Not surprisingly, aversive appraisal of the partner by the virtual agent increased the time of task completion (the two dashed lines in Figure 4). At the same time, results (Figs. 5, 7) appear to be comparable across all kinds of teams.

Figure 4. Typical examples of recorded sessions of shape construction. Solid lines: homogeneous teams; dashed lines: heterogeneous teams; (1) human-human; (2) human-computer, $A = 0$; (3) computer-computer, $A = 0.1$ and 0; (4) human-computer, $A = 0.45$.

One interesting observation can be made from Figure 4: the intensity of aversive moves of a human participant appears to vary systematically during sessions with $A > 0$. Here moves that undo previous partner’s moves are arbitrarily called “aversive”. To remind, the appraisal $A$ controls (is proportional to) the probability of aversive moves performed by the virtual agent. Interestingly, it seems that the probability of aversive moves of the human participant also increases in response to aversive behavior of the virtual partner, when $A > 0$. 

75
In order to validate this intuition, the following analysis was performed. Each recorded session was divided into two equal halves, and the change in frequency of aversive moves from the first half to the second was assessed with the t-test. Results are summarized in Table 2. Statistics were computed for 6 sessions, in 3 of which $A=0$, and in the other 3 the value of $A$ varied from 0.45 to 0.71. All sessions ended with a complete shape.

Table 2.

<table>
<thead>
<tr>
<th>Appraisal of the partner</th>
<th>$A = 0$</th>
<th>$A &gt; 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actor</td>
<td>human</td>
<td>human</td>
</tr>
<tr>
<td>Aversive move frequency</td>
<td></td>
<td></td>
</tr>
<tr>
<td>increase from 1st to 2nd half, confidence interval</td>
<td>(-0.14, 0.22)</td>
<td>(0.21, 0.04)</td>
</tr>
<tr>
<td>$P$-value</td>
<td>0.65</td>
<td>0.0056</td>
</tr>
<tr>
<td>Number of sessions</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Average number of moves per session</td>
<td>13</td>
<td>63</td>
</tr>
</tbody>
</table>

In the virtual agent, the probability of aversive move selection did not change over time. The fact that their number significantly decreases from the first to the second half of the session is due to the decrease in the number of opportunities to undo a partner’s move, when the partner starts making aversive moves. The fact that the decrease in the virtual agent aversive move frequency is comparable to the increase in the human aversive move frequency is due to the high frequency of virtual agent aversive moves in the case $A > 0$.

In contrast, when two virtual agents with fixed mutual appraisals $A$ worked cooperatively as a team, no significant increase in aversive move frequency from the first to the second half of the session was observed. This is because the probability of aversive moves was determined by fixed appraisals. Therefore, the simplistic virtual agent with fixed appraisal of the partner behaves differently from a human.

In order to model human behavior, at the second step in this study the dynamics of the partner appraisal $A$ was enabled in one of the two virtual agents that formed a team, according to the following empirically motivated rule:

$$A_{t+1} = A_t + r(1 - A_t)a,$$

where $t$ is the discrete time (the move number), $r = 0.01$, $A$ is the appraisal of the partner and $a$ is the appraisal of the last partner’s action (1 for aversive actions and 0 otherwise). The initial value of $A$ was zero, in which case the fixed value of $A$ in the other virtual partner was equal to 0.1. In this case, the results were consistent with the human data, in the sense that the number of aversive moves produced by the partner with variable $A$ increased systematically from the first half of the session to the second ($P < 0.0004$ in 26 trials). The amount of increase was not significantly different from the human data. In the alternative condition, the starting value of the dynamical $A$ was 0.1, and the fixed value in the partner was 0. In this case, results were obviously not different from the previous simulations with both $A$’s fixed.

In terms of the proposed framework, the rule (3) has a natural interpretation: appraisal of the agent changes in response to the agent’s actions. Aversive actions should result in aversive appraisal of the actor. In fact, based on this logic, the rule (3) was implemented in the virtual agent a priori. The reason why it was not enabled in the first step is that it was necessary to understand the simplest case first. The findings thus confirm the rule (3) and support the underlying interpretation.

There is, however, a possible alternative interpretation (mentioned above) of the partner’s move reversal: as an expression or intention to achieve dominance over the partner in collaboration. In this case, it seems like the sign of the last term in (3) should be opposite: perception of partner’s actions that force me to submit to the partner should, it seems, decrease the likelihood of my similar actions with respect to the partner, which is given by $A$. The same logic underlies the model description of the second paradigm used here. Why, then, do we see the wrong sign in the experimentally found rule (3)?
A parsimonious answer is that (3) may, in fact, represent the working of a moral schema that is pre-existent in the human mind. Suppose that it is normal for a human participant to expect cooperation from a partner, even if the partner is a virtual agent. Then, the normal value of \( A \), according to the schema, is zero (this is consistent with the observation that two human participants working as a team do not undo each other’s moves). If during interactions with the partner the value of \( A \) becomes negative (the partner appears to dominate me), then the schema should initiate behavior that would restore the normal value of \( A \) (I will show to the partner that I can dominate too). These speculations sound consistent with the subjects’ reports. For example, referring to a condition of an increased value of \( A \) in the virtual agent, a subject wrote:

As the dislike setting grew, my frustration grew, and my need to control the layout of the structure also grew. I kind of felt as though the computer operated much like a child at times.

It sounds like in order to put the “child” its place, the subject had to use aversive moves. Other possible interpretations may involve metacognitive mental states, but at the same time are less parsimonious and at this stage may be unwarranted.

**Paradigm 2: Random Social Interactions**

First, simulations of a homogeneous group consisting of virtual agents only were performed. It was found computationally that stable patterns of mutual appraisals (that correspond to certain configurations of emotional relationships among agents) develop in this model in ~100 iterations. E.g., with the choice of parameters specified above, and given randomly sampled initial appraisal values with all positive valence components, a pattern always develops in which all \( N \) appraisals have positive valence. (This choice of initial conditions for appraisals is consistent with human data: jumping ahead, it should be mentioned here that in all but one sessions with human participants, the first action selected by a subject was “greet”.) If, however, the initially sampled values of appraisals had negative valence components, then a stable configuration with negative valence in all appraisals developed. This kind of an outcome, however, is not considered further.

With a small number of actors in the group \((N<4)\), each actor controlled by a virtual agent reaches its stationary position on the semantic map in approximately 100 iterations, or cycles (each cycle consists in every actor acting on a randomly selected target). In all simulated sessions, the resultant dominance hierarchy that was spontaneously selected as one out of \( N! \) possibilities never changed after 100 cycles with \( N<4 \). At a large \( N \), however, the final configuration remained stationary only macroscopically (as a “cloud”), while there was no microscopically stationary configuration. The qualitative outcome for \( N=2 \) is nearly obvious based on the above analysis. It is a stationary configuration in which the two agent vectors tend to be complex conjugates of each other with the positive real part. At \( N=3 \) the stable configuration is qualitatively the same for two out of three vectors, while the third vector takes a position in the middle between the two, at approximately zero dominance. This result is found both analytically, in a simplified version of the model, and numerically. At \( N \geq 4 \) the configuration is microscopically undetermined: actors do not have permanent stationary positions in the cloud. They tend to spread uniformly in a vertical line at a positive valence, and keep drifting up and down, continuously switching their positions in the dominance hierarchy.

Similar sessions were performed with one of the \( N=3 \) virtual agents replaced by a human participant. Measures used in this case were again appraisals defined by (1) for all actors, including the one controlled by a human. In most cases the outcome for homogeneous and for heterogeneous groups looked similar; however, in some sessions the human behavior seemed less consistent compared to virtual agents. These intuitively noticeable differences in individual sessions, however, were difficult to quantify.

A striking difference between human and virtual actors can be seen at the level of all recorded sessions (13 sessions in total, with 3 participating students in this part of the study, one subject participating in one heterogeneous team). All but one of these sessions ended with placing the human-controlled actor consistently in one and the same position in the dominance hierarchy, which was determined individually for each participating subject. For most subjects this was the middle position, while for one subject it was the top position.

This outcome is nontrivial, because it was not predetermined by initial conditions: the human position in the hierarchy typically changed many times during a session. Even after 20 moves, a crossover occurred on average 3.7 times during a session. Therefore, this observation indicates that individual subjects each had a (pre)determined role and place in the hierarchy in a group of 3 actors \((P < 0.0026)\).

Interestingly, for 2 out of 3 subjects, stable positions of the two virtual agents in the hierarchy were also persistent across different sessions. E.g., in all sessions with Subject 1, Triangle ended up at the top of the hierarchy, and Square at the bottom. In all sessions with Subject 3, Square ended up always in the middle, and Triangle at the bottom.

All virtual agents in these sessions, however, were controlled by one and the same script and had no memory of a previous session; their initial conditions were sampled from one and the same uniform probability distribution.
Moreover, in the first sessions of Subjects 2 and 3, initial conditions for virtual agents were identical across subjects (this is because the same seed for the random number generator was used in both initial sessions; however, this was not the case with other sessions and other subjects, where initial conditions were sampled independently in each session).

Given these settings, the persistence of any significant differences across subjects, e.g., in the relative positions of actors in the hierarchy, can only be a result of interactions with the human possibly influenced by images of the virtual teammates that spontaneously developed in the human mind. Indeed, for the human participant, virtual agents had persistent names (Square and Triangle) and therefore could acquire persistent individual roles attributed to them in the human mind. The subject was not aware of the details of implementation: e.g., that both virtual partners were controlled by the same script and had no memory of previous sessions. The emergence of subjectively perceived persistent roles of individual actors in the human mind seems intuitively obvious in the presented data.

The study needs to be continued to see whether moral schemas can also stabilize social relations among agents in a large group, and also account for other complex/social emotions, as illustrated by examples discussed below.

References


