Expert Systems With Applications 97 (2018) 266–275

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Measuring the emotional state among interacting agents: A game theory approach using reinforcement learning

Mireya Salgado[a,b], Julio B. Clempner[c,d,*]

[a] Centro de Alta Dirección en Ingeniería y Tecnología, Universidad Anahuac, Av. Universidad Anáhuac 46, Lomas Anáhuac, Edo. Mexico 50130, Mexico
[b] Facultad de Ingeniería, Universidad Autónoma del Estado de México, Mariano Matamoros 1042, Universidad, Toluca de Lerdo, Edo. Mexico 50130, Mexico
[c] Escuela Superior de Física y Matemáticas, Instituto Politécnico Nacional, San Pedro Zacatenco, Gustavo A. Madero, Mexico City 07738, Mexico
[d] School of Physics and Mathematics, National Polytechnic Institute, Mexico City, Mexico

ARTICLE INFO

Article history:
Received 8 September 2017
Revised 18 December 2017
Accepted 19 December 2017
Available online 20 December 2017

Keywords:
Adaptive autonomous agents
Emotional model
Kullback–Leibler distance
Game theory
Reinforcement learning

ABSTRACT

Studies on emotion perception often require stimuli that convey different emotions. These stimuli can serve as a tool to understand how agents react to different circumstances. Although different stimuli have been commonly used to change the emotions of an agent, it is not clear how to measure the emotional state of an agent.

This paper suggests a new method for measuring the emotional state among interacting agents in a given environment. We present the modeling of an adaptive emotional framework that takes into account agent emotion, interaction and learning process. For solving the problem, we employ a non-cooperative game theory approach for representing the interaction between agents and a Reinforcement Learning (RL) process for introducing the stimuli to the environment. We restrict our problem to a class of finite and homogeneous Markov games. The emotional problem is ergodic: each emotion can be represented by a state in a Markov chain which has a probability to be reached. Each emotional strategy of the Markov model is represented as a probability distribution. Then, for measuring the emotional state among agents, we employ the Kullback-Leibler distance between the resulting emotional strategies of the interacting agents. It is a distribution-wise asymmetric measure, then the feelings of one player for another are relative (different). We propose an algorithm for the RL process and for solving the game is proposed a two-step approach. We present an application example related to the selection process of a candidate for a specific position using assessment centers to show the effectiveness of the proposed method by a) measuring the emotional distance among the interacting agents and b) measuring the "emotional closeness degree" of the interacting agents to an ideal proposed candidate agent.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

1.1. Brief review

Emotions are a fundamental aspect of life and are very complicated to be model. Such complexity arises from the fact that they can be affected by many factors (Ding, Sethu, Epps, & Ambikairajah, 2012; Mill, Allik, Realo, & Valk, 2009; Vogt & Andre, 2006). Current discussion in psychology classifies emotions in two different essential categories: the basic or primary and the social emotions. We will focus on the basic set proposed by Ekman, Friesen, and Ellsworth (1982) composed by six primary emotions: i) anger, ii) disgust, iii) fear, iv) joy, v) sadness, vi) surprise. This set is characterized by the fact that are emotions which we feel instantly as a response to a stimulus.

Many real-world applications depend on knowing human's affective states (emotions). Affective states are complex psychophysiological constructs composed of three fundamental dimensions: valence, arousal, and motivational intensity. We are interesting in the motivation intensity that refers to the strength of urge to move toward or away from a particular stimulus. For instance, selecting the ideal or expected candidate (Belkaid & Sabouret, 2014; Clempner, 2010) is crucial for organizational success (employment, especially management or military command). Dealing with difficult employees (or clients) is a challenging, important part of a manager's job that can be effective handled employing information about the emotional state of that person. Selection methods that allow firms to identify the right people (from a pool of applicants)

* Corresponding author.
E-mail addresses: msalgadog@uaemex.mx (M. Salgado), julio@clempner.name, jclempnerk@ipn.mx (J.B. Clempner).

https://doi.org/10.1016/j.eswa.2017.12.036
0957-4174/© 2017 Elsevier Ltd. All rights reserved.

are vital components and depend on the position (Fisher, Schoenfeldt, & Shaw, 2003). Assessment centers are most often for promotion to managerial positions because allow applicants to try on senior roles in an emotional simulated environment. An additional example is in social media where building robust emotion measuring systems is essential for human-computer interaction (Bickmore & Picard, 2005; Chatzakou, Vakali, & Kafetsios, 2017; Deng, Wang, Li, & Xu, 2015; Luo, Zeng, & Duan, 2016): the emotional state from people involved is imperative for the success of the social network. Emotional compatibility detection has attracted high interest in social media because emotional information is a main component of human communication.

Artificial Intelligence (AI) is the study of human intelligence and actions replicated artificially, such that the resultant bears to its design a reasonable level of rationality (Clempner & Poznyak, 2016b; 2017). However, rationality alone is insufficient for successful applications. Emotions play a critical role in processes such as rational decision-making and human intelligence (Picard, 1997). Until-to-date, modeling of emotions is an open research area: agents that are capable of carry out certain actions, appear to be able to think, but having emotions of their own, are still uncommon in AI. Emotions have long been considered as a fundamental complement to AI applications.

The emotional strategies analyzed in this paper represent the most frequent type of behavior founded in practice in problems of interaction considered within the AI research area. The importance of AI in the field is that it proposed a set of tools that are driving forward key parts of the futurist agenda. Many researchers and practitioners have investigated this problem using different computational tools to increase the sophistication of autonomous (interacting) agents: machine-learning algorithms (Champandard, 2003; Funge, 1999; Lim, Dias, Aylett, & Paiva, 2012), education (Core et al., 2006; Paiva et al., 2005) and social simulation (Bickmore & Picard, 2005).

In order to achieve their goal agents must adapt to a dynamic environment and need to react to an unexpected stimuli. A key challenge for any emotion framework is to explain this dynamic emotional process (Becker-Asano & Wachsmuth, 2010; Marsella & Gratch, 2009). For representing real-world application, it is necessary to compute the values of the parameters of the environment (transition probabilities) and the reward functions, which are typically, hand-tuned by experts in the field until it is acquired a satisfactory value. This results in an undesired process. Our main contribution relies on a non-cooperative game theory approach for representing the interaction between agents and a RL process for introducing the stimuli to the environment (Almahdi & Yang, 2017; Cruz & Yu, 2017; Kazemitabar, Taghizadeh, & Beigy, 2017; Radac & Precup, 2017; Radac, Precup, & Roman, 2017; Vamvoudakis, Modares, Kiumarsi, & Lewis, 2017; Zhang, Jiang, Luo, & Xiao, 2017; Zhou, Haoa, & Duval, 2016). This result in a model of an adaptive emotional framework that takes into account the interaction between agents and a learning process that can simulate human attributes.

1.2. Related works

Several computational models of emotions have been presented in the literature considering affective processing (Becker-Asano & Wachsmuth, 2010; Champandard, 2003; Chatzakou et al., 2017; Deng et al., 2015; Dias, Mascarenhas, & Paiva, 2014; El-Nasr, Yen, & Ioerger, 2000; Funge, 1999; Lim & Kim, 2017; Lim et al., 2012; Luo et al., 2016; Marsella & Gratch, 2009).

For instance, Soar (Laird, Newell, & Rosenbloom, 1987; Newell, 1990) is a cognitive architecture for human cognition realized as a production system which consists of matching the contents of working memory to the precondition of rules. ACT-R (Anderson et al., 2004) is an alternative cognitive architecture, that shares similarities with Soar, based on a long-term memory model where the information is represented by networks and it is activated by a symbolic process of spreading activation. Marinier, Laird, and Lewis (2009) proposed PEACTIDM which introduces emotions into Soar by modifying the core of the architecture. Belkaid and Sabouret (2014) presented a job interview simulation in the context of human-agent interaction for improving people's social skills and supporting professional inclusion. The method focuses on the affective dimension and it is based on modal logic and inference rules about the mental states, emotions and social relations. Dias et al. (2014) suggested FAtiMA a generic and flexible architecture for emotional agents. Aylett and Louchart (2008) employed the existing cognitive appraisal mechanism presented in FAtiMA adapted to produce a second appraisal cycle, a double appraisal, in order to evaluate the emotional impact of possible actions. An alternative very successful architecture to FAtiMA is EMA (Marsella & Gratch, 2009) that proposed a single and automatic appraisal process that operates over a person's interpretation of their relationship to the environment. EMA is based on Soar (Gratch & Marsella, 2004; Laird et al., 1987; Newell, 1990) and the process theory of emotion presented by Smith and Lazarus (1990). Becker-Asano and Wachsmuth (2010) introduced the WASABI Affect Simulation Architecture, in which a virtual human's cognitive reasoning capabilities are combined with simulated embodiment to achieve the simulation of primary and secondary emotions. Marsella, Gratch, and Petta (2010) proposed a computational architecture consisting of three parts: an appraisal-derivation model, an affect-derivation model, and an affect-consequent model. Reisenzein et al. (2013) proposed a modularization approach for deconstructing emotion theories into basic Assumptions and a unification and standardization for translating different emotion theories into a common informal conceptual system and implement them in a common architecture. Bosse and Zwanenburg (2014) developed the modeling language LEADSTO based on the Temporal Trace Language (Bosse, Jonker, van der Meij, Sharpanskykh, & Treur, 2009) which focus on the simulation of prospect-based emotions where relief and disappointment depend on a combination of surprise and satisfaction. Perez-Gaspar, Caballero-Morales, and Trujillo-Romero (2016) presented a multimodal emotion recognition architecture based on hidden Markov models. García-Magariño and Plaza (2017) suggested AB-SEM which allows instructors to define different mindfulness programs and simulate their repercussions on the emotions of a group of practitioners with certain features.

1.3. Main results

This paper suggest the modeling of an adaptive emotional framework considering agent interaction and learning process.

The main contributions of this paper are as follows:

- Employs a non-cooperative game theory approach for representing the interaction between agents.
- Solves the game using an innovative two-step approach.
- Develops a RL process for introducing the stimuli to the environment.
- Presents an algorithm for the RL process.
- Employs the Kullback–Leibler distance between the resulting strategies of the interacting agents.
- Shows the effectiveness of the proposed method presenting an application example related to assessment centers.

1.4. Organization of the paper

The paper is organized as follows. In the next section we introduce all the background needed to understand the rest of the pa-

per. Section 3 suggests the introduction of the joint strategy distribution and define the emotional distance employing the Kullback-Leibler divergence. The RL process and the game theory solution is described in Section 4. Section 5 shows the effectiveness of the proposed method presenting an application example related to assessment centers. Section 6 concludes the paper.

2. Markov models

2.1. Markov chains

Let the set of all nonnegative integers be denoted by $\mathbb{N}$. A Borel subset X of a complete and separable metric space is called a Borel space, and its Borel σ-algebra is denoted by $\mathbb{B}(X)$. Let X and Y be Borel spaces. A *stochastic kernel* on X given Y is a function $P(\cdot\,|\,\cdot)$ such that $P(\cdot\,|y)$ is a probability measure on X for every fixed $y \in Y$, and $P(B|\,\cdot)$ is a measurable function on Y for every fixed $B \in \mathbb{B}(X)$. If $X = Y$, then P is called a (Markov) *transition kernel*.

Let $M = (S, A, \{A(s)\}_{s\in S}, \mathbb{K}, P)$ be a Markov chain (Clempner & Poznyak, 2014; Poznyak, Najim, & Gómez-Ramírez, 2000b), where S is a finite set of *states*, $S \subset \mathbb{N}$ and A is a finite *set of actions*. For each $s \in S$, $A(s) \subset A$ is the non-empty set of admissible actions at state $s \in S$. Without loss of generality we may take $A = \cup_{s\in S} A(s)$. Whereas, $\mathbb{K} = \{(s, a) | s \in S, a \in A(s)\}$ is the set of *admissible state-action pairs*, which is a measurable subset of $S \times A$. The variable $p_{j|ik}$ is a stationary controlled *transition matrix* which defines a stochastic kernel on S given $\mathbb{K}$, where $p_{j|ik} := P(X_{t+1} = s_j | X_t = s_i, A_t = a_k)$ $\forall t \in \mathbb{N}$ represents the probability associated with the transition from state s_i to state s_j, $i = \overline{1, N}$ $(i = 1, \ldots, N)$ and $j = \overline{1, N}$ $(j = 1, \ldots, N)$, under an action $a_k \in A(s_i)$, $k = \overline{1, K}$ $(k = 1, \ldots, K)$. A *Markov Decision Process* is a pair $MDP = (M, U)$ where M is a controllable Markov chain and $U : \mathbb{K} \to \mathbb{R}$ is a *utility function*, associating to each state-action pair a real value.

This *MDP* is interpreted as follows: At each time $t \in \mathbb{N}$ the decision maker observes the state of a dynamical system, say $X_t = s_i \in S$, and selects the action (control) $A_t = a_k \in A(s_i)$. Then, a utility $U_{(s,\,a)}$ is incurred and, regardless of the previous states and actions, the state of the system at time $t + 1$ will be $X_{t+1} = s_j \in S$ with probability $p_{j|ik}$; this is the Markov property of the decision process. Let $\mathbb{F}$ denote the set of measurable functions f: $S \to A$ such that $f(s)$ is in $A(s)$ for all $s \in S$, and Φ stands for the set of stochastic kernels φ on A given S for which $\varphi(A(s)|s) = 1$ for all $s \in S$.

The space $\mathbb{H}_t$ of possible histories up to time $t \in \mathbb{N}$ is defined by $\mathbb{H}_0 := S$ and $\mathbb{H}_t := \mathbb{K}_t \times S$, $t \geq 1$, $h_t = (s_0, a_0, \ldots, s_i, a_i, \ldots, s_t)$ stands for a generic element of $\mathbb{H}_t$, where $a_k \in A(s_i)$. A *control policy* $\pi = \{\pi(t)\}$ is a special sequence of stochastic kernels: for each $t \in \mathbb{N}$ and $h_t \in \mathbb{H}_t$, $\pi_{(\cdot|h_t)}(t)$ is a probability measure on A concentrated on $A(s_t)$ then the controller chooses actions according to the control A_t applied at time t belongs to A with probability $\pi_{(A|h_t)}(t) = 1$, where h_t is the observed history of the process up to time t. Then, the policy $\pi_{k|i}(t) := P(A_t = a_k | X_t = s_i)$ represents the probability measure associated with the occurrence of an action $A_t = a_k$ from state $X_t = s_i$ at time $t \in \mathbb{N}$.

Let Π be the class of all policies. Given the π policy being used for choosing actions and the initial state $X_0 = s \in S$, the distribution of the state-action process denoted by $\{(S_t, A_t) | t \in \mathbb{N}\}$ is uniquely determined and such a distribution and the corresponding expectation operator are P and E respectively.

2.2. Markov games

A game consists of a set $\mathcal{M} = \{1, \ldots, m\}$ of players (indexed by $l = \overline{1, m}$) (Clempner & Poznyak, 2011; 2016a; Solis, Clempner, & Poznyak, 2016). The dynamics of a Markov game is described as follows. Each of the players l is allowed to randomize actions, with probabilities $\pi^l_{k|i}(t)$, over the pure action choices $a^l_k \in A^l(s^l_i)$, $i = \overline{1, N}$ and $k = \overline{1, K}$. Herein, we will consider only stationary strategies $\pi^l_{k|i}(t) = \pi^l_{k|i}$. In the ergodic case when all Markov chains are ergodic for any stationary strategy $\pi^l_{k|i}$ the probabilities $P^l\left(X^l_{t+1} = s^l_j\right)$ exponentially quickly converge to their limit probabilities, given by the solution to the linear system of equations $P^l(s^l_j) = \sum_{i=1}^{N}\left(P^l(s_i)\sum_{k=1}^{K} p^l_{j|ik}\pi^l_{k|i}\right)$.

The expected reward function R^l of each player l depends on the states and actions of all the other players and it is given by the values of V^l_{ik} as follows

$$R^l\left(\pi^1, ., \pi^m\right) := \mathrm{E}\left\{\sum_i^N \sum_k^K V^l_{ik} \prod_{l=1}^{m} \pi^l_{k|i} P^l(s^l_i)\right\} \tag{1}$$

given that

$$V^l_{ik} = \sum_i^N \sum_k^K \left(\sum_j^N U^l_{ik} p^l_{j|ik}\right) \tag{2}$$

The limiting average reward is as follows

$$\begin{aligned} R_t\left(\pi^1, ., \pi^m\right) &= \lim_{t\to\infty} t^{-1}\mathrm{E}\left\{\sum_{n=0}^{t}\sum_i^N\sum_k^K V^l_{ik}(n)\prod_{l=1}^{m}\pi^l_{k|i}(n)P^l(s^l_i)\right\} \\ &= \lim_{t\to\infty} t^{-1}\mathrm{E}\left\{\sum_{n=0}^{t}\sum_i^N\sum_k^K\left(\sum_j^N U_{ijk}{}^l(n)\,p^l_{j|ik}(n)\right)\prod_{l=1}^{m}\pi^l_{k|i}(n)P^l(s^l_i)\right\} \\ &= \max_{\pi^1,.,\pi^m} R\left(\pi^1, ., \pi^m\right) \end{aligned} \tag{3}$$

The *individual aim* of each player is to maximize the reward, i.e. $R^l(\pi^l) \to \max_{\pi^l}$.

3. Emotional distance

We will consider $z^l := \left[z^l_{ik}\right]_{i=\overline{1,N};k=\overline{1,K}}$ $(l = \overline{1, m})$ as follows

$$z^l_{ik} = \pi^l_{k|i} P^l(s^l_i) \tag{4}$$

Note that by (4) it follows that

$$P^l(s^l_i) = \sum_k^K z^l_{ik} \qquad \pi^l_{k|i} = \frac{z^l_{ik}}{\sum_k^K z^l_{ik}} \tag{5}$$

where $\sum_{ik}^{NK} z^l_{ik} = 1$, for all $l = \overline{1, m}$. The strategy z^l_{ik} is called "joint strategy" and it is a probability distribution. In the ergodic case $\sum_k^K z^l_{ik} > 0$ for all $l = \overline{1, m}$. z^l_{ik}. The admissible set Z^l_{adm} is defined as follows:

$$z^l \in Z^l_{adm} = \begin{cases} z^l : \sum_i^N \sum_k^K z^l_{ik} = 1,\ z^l_{ik} \geq 0 \\ h^l_j(z) = \sum_i^N \sum_k^K p^l_{j|ik} z^l_{ik} - \sum_k^K z^l_{jk} = 0 \end{cases} \tag{6}$$

Then, in terms of z^l_{ik}, the expected reward function R^l of each player l depends on the states and actions of all the other players and it is given by the values of $V^l_{i,k}$ as follows

$$R^l\left(z^1, ., z^m\right) := \mathrm{E}\left\{\sum_i^N \sum_k^K V^l_{ik} \prod_{l=1}^{m} z^l_{ik}\right\} \tag{7}$$

given that

$$V^l_{ik} = \sum_i^N \sum_k^K \left(\sum_j^N U^l_{ik} p^l_{j|ik}\right) \tag{8}$$

The limiting average reward is as follows

$$R_t^l(z^1,.,z^m)=\lim_{t\to\infty}t^{-1}\mathrm{E}\left\{\sum_{n=0}^{t}\sum_{i}^{N}\sum_{k}^{K}V_{ik}^l(n)\prod_{l=1}^{m}z_{ik}^l(n)\right\}$$
$$=\lim_{t\to\infty}t^{-1}\mathrm{E}\left\{\sum_{n=0}^{t}\sum_{i}^{N}\sum_{k}^{K}\left(\sum_{j}^{N}U_{ijk}{}^l(n)p_{j|ik}^l(n)\right)\prod_{l=1}^{m}z_{ik}^l(n)\right\} \quad (9)$$
$$=\max_{z^1,..,z^m\in Z_{adm}}R^l(z^1,.,z^m)$$

The*individual aim* of each player is to maximize the reward, i.e. $R^l(z)\to\max_{z\in Z_{adm}}$.

We will determine a measuring mechanism for establishing the emotional distance between two different players $l\neq l'$. The Kullback–Leibler divergence is a measure of how one probability distribution z_{ik}^l diverges from a second expected probability distribution $z_{ik}^{l'}$. In our case the Kullback–Leibler divergence is defined to be the expectation of the logarithmic difference between the probabilities distributions z_{ik}^l and $z_{ik}^{l'}$ where the expectation is taken using the probabilities distribution z_{ik}^l as follows

$$D_{KL}(z_{ik}^l|z_{ik}^{l'})=\textstyle\sum_{ik}z_{ik}^l\log\frac{z_{ik}^l}{z_{ik}^{l'}},\ \text{for all } l\neq l',\ l=\overline{1,m} \quad (10)$$

As a result of applying the a Kullback–Leibler divergence we have that: $D_{KL}(z_{ik}^l|z_{ik}^{l'})=0$ means that we can expect similar, if not the same, emotion of two different distributions z_{ik}^l and $z_{ik}^{l'}$, while a $D_{KL}(z_{ik}^l|z_{ik}^{l'})=1$ means that the two distributions z_{ik}^l and $z_{ik}^{l'}$ have completely different emotions.

4. Reinforcement learning

4.1. Learning rules

In a RL process (Poznyak, Najim, & Gomez-Ramirez, 2000a; Sánchez, Clempner, & Poznyak, 2015; Sutton & Barto, 1998; Trejo, Clempner, & Poznyak, 2016) the objective of an agent is to maximize the reward in each period of time n. The maximization goal is restarted after each end of each period of time. In our case, we consider to optimize the average reward over the whole process. In such reinforcement learning process the agent and its environment are represented considering to be in a state $s\in S$ and perform actions $a\in A$, each of which are members of a finite and discrete sets. A state s contains all relevant information about the current situation to predict future states, e.g. an example would be the current emotion of the agent. An action a is a control and it is employed to change the state of the model. For instance, an action can change from joy to sadness or fear to anger. At each step, the agent obtains a reward R, which is a scalar value and presumed to be a function of the states and action. Rewards can be represented by utilities for reaching emotional targets. The main goal of a RL process is to find a policy π which selects an action a in a given state s maximizing the expected reward. The RL agent needs to find the relations between states, actions, and rewards.

We propose a model from experiences that is computed by counting the number η of observed experiences defining the following variables recursively (Najim & Poznyak, 1994; Poznyak et al., 2000b). Let us define $\eta_{i|k}$ as the total number of times that the process evolves from state i applying action k recursively as follows

$$\eta_{ik}(t)=\sum_{n=1}^{t}\chi\left(X_t=s_i,A_t=a_k\right) \quad (11)$$

as well, let us define η_{ijk} as the total number of times that the process evolves from state i to state j applying action k recursively as follows

$$\eta_{ijk}=\sum_{n=1}^{t}\chi\left(X_{t+1}=s_j|X_t=s_i,A_t\right) \quad (12)$$

such that

$$\chi(\mathcal{E}_n)=\begin{cases}1 & \text{if the event } \mathcal{E} \text{ occurs at interaction } n\\ 0 & \text{otherwise}\end{cases}$$

where $\eta_{ik}(t)$ the number of visits in the observed state s_i for time n applying action k in the RL process ($n\in\mathbb{N}$) and, $\eta_{ijk}(t)$ denote the total number of times that the process evolves from the observed state s_i to s_j applying action k. We have that $\eta_{ik}(t)=\sum_{j=1}^{N}\eta_{ijk}(t)$. The frequency is defined by $f_{ijk}(t)=\frac{\eta_{ijk}(t)}{t}$.

When the MDP is observable the learning rule for π is defined as follows

$$\hat{p}_{(\hat{\jmath}|\hat{\imath}\hat{k})}(t)=\frac{\eta_{ijk}(t)}{\eta_{ik}(t)} \quad (13)$$

and the frequency $f_{ijk}(t)=\frac{\eta_{ijk}(t)}{t}$ as we expected. The learning process considers the maximum likelihood model $\left(\frac{0}{0}:=0\right)$.

The estimation rule for the utility $\hat{U}_{\hat{\imath}\hat{\jmath}\hat{k}}(t)$ at the entry (ijk) is given by

$$\hat{U}_{\hat{\imath}\hat{\jmath}\hat{k}}(t)=\frac{\sum_{n=0}^{t}\xi_U(n)\,\chi\left(X_{t+1}=s_{\hat{\jmath}}\mid X_t=s_{\hat{\imath}},A_t=a_{\hat{k}}\right)}{\sum_{n=0}^{t}\chi\left(X_{t+1}=s_{\hat{\jmath}}\mid X_t=s_{\hat{\imath}},A_t=a_{\hat{k}}\right)}$$
$$=\frac{K_{\hat{\imath}\hat{\jmath}\hat{k}}(t_0)+\alpha_{\hat{\imath}\hat{\jmath}\hat{k}}(n)}{n_{\hat{\imath}\hat{\jmath}\hat{k}}(t_0)+\beta_{\hat{\imath}\hat{\jmath}\hat{k}}(n)} \quad (14)$$

where

$$\alpha_{\hat{\imath}\hat{\jmath}\hat{k}}(n)=\sum_{n=t_0+1}^{t}\xi_U(n)\,\chi\left(X_{t+1}{=}s_{\hat{\jmath}}\mid X_t{=}s_{\hat{\imath}},A_t{=}a_{\hat{k}}\right) \quad (15)$$

$$\beta_{\hat{\imath}\hat{\jmath}\hat{k}}(n)=\sum_{n=t_0+1}^{t}\chi\left(X_{t+1}{=}s_{\hat{\jmath}}\mid X_t{=}s_{\hat{\imath}},A_t{=}a_{\hat{k}}\right) \quad (16)$$

$$K_{\hat{\imath}\hat{\jmath}\hat{k}}(t)=\sum_{n=1}^{t}\xi_U(n)\,\chi\left(X_{t+1}=s_{\hat{\jmath}}\mid X_t=s_{\hat{\imath}},A_t=a_{\hat{k}}\right) \quad (17)$$

and

$$\xi_U:=U_{ijk}+\mu_U\mathrm{rand}\big[-1,1\big],\quad \mu_U\leq U_{\hat{\imath}\hat{\jmath}\hat{k}}$$

4.2. RL algorithm

The goal of the emotional RL Algorithm 1 is to find for the Markov game a policy $\hat{\pi}_{\hat{k}|\hat{\imath}}^l(t)$ which maximizes the reward $R_t^l\left(\hat{\pi}_{\hat{k}|\hat{\imath}}^l\right)$ for all the states s.

The RL process presents a pre-training stage done in a soft-max layer-wised manner. Initially we have the transition matrix $p_{j|ik}^l$ and utility matrix U_{ijk}^l. The process assumes a softmax action based on a Boltzmann distribution of the strategy $\pi_{k|i}^l$. The players learn by receiving rewards after every selected action. They keep track of these rewards and then select actions that each player believes will maximize the reward.

Then, the agent interacts with its environment and at each stage of the process randomly picks randomly an action $a^l(t)=a_{(\hat{k})}$ (for the estimated value $\hat{k}$) from the vector $\pi_{\hat{k}|\hat{\imath}}^l$ (for a fixed $\hat{\imath}_l$). Next, the player l employs the transition matrix $p_{\hat{\jmath}|\hat{\imath}\hat{k}}^l$ to choose randomly the consecutive state $s^l(t+1)=s_{(\hat{\jmath})}^l$ (for the estimated value $\hat{\jmath}$) from the vector $p_{\hat{\jmath}|\hat{\imath}\hat{k}}^l$ (for a fixed $\hat{\imath}$ and $\hat{k}$). The fixed entry $(\hat{\imath}\hat{\jmath}\hat{k})$ of the transition matrix $\hat{p}_{\hat{\jmath}|\hat{\imath}\hat{k}}^l$ is estimated applying the learning rule

Algorithm 1 Emotional RL process.

Initialization: Set $t = 1$ and observe the initial state s_i.
Compute the error according to

$$e_p(t) = \sum_{k=1}^{K} tr\left((\hat{p_k}^l(t-1) - \hat{p_k}^l(t))^\intercal(\hat{p_k}^l(t-1) - \hat{p_k}^l(t))\right)$$

While $e_p(t-1) > e_p(t)$ **do**
a) Initialize period t:
1. For all (s, a) in $S \times A$ initialize the state-action counts for period t,
2.Further, set the state-action counts prior to period t,

$$\eta^l_{ik}(t) = \sum_{n=1}^{t} \chi\left(X_t = s_i, A_t = a_k\right)$$

and

$$\eta^l_{ijk}(t) = \sum_{n=1}^{t} \chi\left(X_{t+1} = s_j | X_t = s_i, A_t = a_k\right)$$

3. For $s_i, s_j \in S$ and $a_k \in A$ set the observed rewards and the transition counts prior to period t, compute the estimated values of $\hat{U}^l_{\tilde{ij}\hat{k}}$ and $\hat{p}^l_{(\hat{j}|\hat{ik})}$

$$\hat{U}^l_{\tilde{ij}\hat{k}}(t) = \frac{K_{\tilde{ij}\hat{k}}(t_0) + \alpha_{\tilde{ij}|\hat{k}}(n)}{n_{\tilde{ij}\hat{k}}(t_0) + \beta_{\tilde{ij}|\hat{k}}(n)}$$

$$\hat{p}^l_{(\hat{j}|\hat{ik})}(t) = \frac{\eta_{ijk}(t)}{\eta_{ik}(t)}$$

b) Compute the game: policy $\hat{\pi}^l_{ik}$:
Let z^l:=$col\left(z^l_{ik}\right)$

$$\Gamma\left(z, \hat{z}(z)\right) := \sum_{l=1}^{n}\left[f_l\left(\tilde{z}^l, z^{\hat{l}}\right) - f_l\left(z^l, z^{\hat{l}}\right)\right]$$

such that

$$\tilde{z}^l := \arg\max_{z^l \in Z^l_{adm}} f_l\left(z^l, z^{\hat{l}}\right)$$

obtaining

$$\hat{\pi}^l_{k|i} = \frac{z^l_{ik}}{\sum_k^K z^l_{ik}}$$

c) Execute policy $\hat{\pi}^l_{ik}$
Choose action $a^l_k(t) = \hat{\pi}^l_{ik}(t)$
Obtain reward $R^l_t\left(\hat{\pi}^l_{ik}\right)$,
Set $j = i$ and $t = t + 1$.
end

$\hat{p}^l_{\hat{j}|\tilde{ik}}(t)$. Given that $\sum_j \hat{p}^l_{\hat{j}|\tilde{ik}} \neq 1$, at each step the complete row $\hat{\imath}$ is projected to the simplex ($\Pr : \hat{p}_{\hat{ik}}(t) \to S^N$).

Now, the environment moves to a new state $s^l_{(\hat{j})}$ given $a^l_{(\hat{k})}$ and $s^l_{(\hat{i})}$for each player l. The estimated values are updated employing the learning rules for computing $\hat{p}_{(\hat{j}|\hat{ik})}(t)$ and $\hat{U}^l_{(\tilde{ij}\hat{k})}(t)$ (in this case the real-valued $p^l_{(j|ik)}(t)$ and $U^l_{(ijk)}(t)$ are replaced by the approximation $\hat{p}^l_{(\hat{j}|\hat{\imath},\hat{k})}(t)$ and $\hat{U}^l_{(\tilde{ij}\hat{k})}(t)$). We consider that the use of only the value-maximizing action at each state is unlikely in practice, then a specific selection of policies is used to ensure convergence: the computational effort is focused only on those states where a learning rule has an impact in the change of $\hat{p}^l_{\hat{j}|\hat{ik}}$. It employs $\left(e\left[\hat{p}^l(t-1) - \hat{p}^l(t)\right] > 0\right)$ which is the mean squared error that measures the average of the squares of the errors of $\hat{p}^l_{\hat{j}|\hat{ik}}$. As a result, the learning rules are applied only to a subset of states, and they are usually applied (a fixed number of times) until $\hat{p}^l_{\hat{j}|\hat{ik}}$ convergences. If the condition of estimated error e is not s atisfied, then the selection of the random variables $s^l_{(i)}$, $s^l_{(j)}$ and $a^l_{(k)}$ is carried out again. On the other hand, the policy $\hat{\pi}^l_{\hat{k}|\hat{\imath}}$ is computed again using the game theory module. A natural concern is whether the sequence of policies $\hat{\pi}^l_{\hat{k}|\hat{\imath}}$ and the parametric approximation $\hat{p}_{(\hat{j}|\hat{ik})}(t)$ and $\hat{U}^l_{(\tilde{ij}\hat{k})}(t)$ generated by the approximate policy-iteration algorithm converges to the optimal values.

For computing the game policy $\hat{\pi}^l_{\hat{k}|\hat{\imath}}$ we follow (Trejo, Clempner, & Poznyak, 2015; 2017). Let us consider a game whose strategies are denoted by $z^l := col\left(z^l_{ik}\right)$ where col is the column operator which transforms the matrix z^l_{ik} into a column. Here $f_l\left(z^l, z^{\hat{l}}\right)$ is the utility-function of the player l which plays the strategy $z^l \in Z^l_{adm}$ and the rest of the players the strategy $z^{\hat{l}} \in Z^{\hat{l}}_{adm}$. For solving the game we employ the iterated proximal/gradient method which employs both the proximal and the gradient method for computing the Nash equilibrium in Markov games. The method transforms the game theory problem in a system of equations, in which each equation itself is an independent optimization problem for which the necessary and efficient condition of a maximum is computed employing a nonlinear programming solver. The proximal/gradient algorithm is as follows

- 1-st step

$$\begin{aligned}\bar{z}^l_n &= \arg\max_{z \in Z_{adm}}\left\{-\tfrac{1}{2}\|z - z_n\|^2 + \gamma_n f\left(z^l, z^{\hat{l}}_n\right)\right\}\\ \bar{z}^{\hat{l}}_n &= \arg\max_{z \in Z_{adm}}\left\{-\tfrac{1}{2}\|z - z_n\|^2 + \gamma_n f\left(z^l_n, z^{\hat{l}}\right)\right\}\end{aligned} \tag{18}$$

- 2-nd step

$$\begin{aligned}z^l_{n+1} &= \Pr{}_{z \in Z_{adm}}\left\{z_n + \gamma_n \nabla_z f\left(\bar{z}^l, \bar{z}^{\hat{l}}_n\right)\right\}\\ z^{\hat{l}}_{n+1} &= \Pr{}_{z \in Z_{adm}}\left\{z_n + \gamma_n \nabla_z f\left(\bar{z}^l_n, \bar{z}^{\hat{l}}_n\right)\right\}\end{aligned} \tag{19}$$

where $\Pr_{z \in Z_{adm}}$ is the projection operator. Both steps the 1-st and the 2-nd are considered together a single step.

The computational algorithm for the emotional RL process is as follows:

5. Numerical example

This application example is related to the selection process of a candidate for a specific position using assessment centers. The goal is to show the effectiveness of the proposed method by: a) measuring the emotional distance among the interacting agents and, b) measuring the "emotional closeness degree" of the interacting agents to an ideal proposed candidate agent. The system is able to simulate the emotional mental state of the agents by the stimuli that inform the RL process about their affects and social attitude in the context of human-agent interaction represented by the game (game theory) that focuses on the emotion dimensions.

5.1. Description of the assessment centers test

A common point of view of the selection process of a candidate for a specific position is that it is imprecise: the selection process can be discussed and judged, but cannot be weighted or measured. As a result, the selection process is based on an intuitive method supported by the intuition and perception of the interviewer. Terms such as "I like him/her", "I think that he/she is good" and similar concepts show how people reason about something uncertain and diffuse. This perception shows the fact that people perceive, understand, interpret and handle people selection

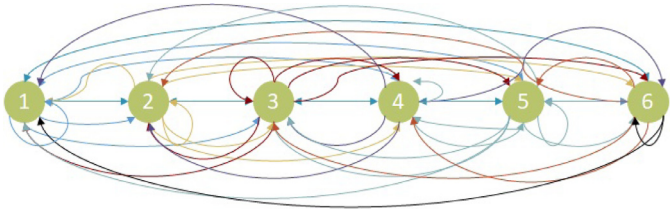

Fig. 1. Emotional Markov Chain.

in different ways. The implication of this perception is that selection cannot be controlled and managed, nor can it be quantified. This view is in contrast to the fact that selecting can and should be defined, measured, and managed (Bouhuys, Geerts, Mersch, & Jenner, 1996; Schmidt & Hunter, 1998). Selecting the ideal or expected employee is crucial for organizational success.

Assessment centers are a series of interaction exercises to test skills of candidates measured to decide their suitability for specific types of employment, especially management or military command. The candidates' personality and aptitudes are evaluated. The assessment center is the final stress test in a recruitment process, and it is where the employer make pressure over the candidates. Designing and running an assessment center require the employment of expensive resources and a lot of time from the recruiter, so they select the candidates who they think have a real chance of being right for the position. Assessment centers are most often used for promotion to managerial positions. They allow applicants to try on senior roles in a simulated environment.

In our case, we will consider three candidates and evaluate six emotions described by the Markov chain represented in Fig. 1. The basic emotions are the following: 1) anger, 2) disgust, 3) fear, 4) joy, 5) sadness, 6) surprise. The actions to transit among the six emotional states are: 1) intimidate, 2) defend and 3) refuse.

5.2. Results of the assessment center

The resulting estimated numerical data for the Markov chain represented in Fig. 1 is as follows.

The estimated values of the transition matrices $\hat{p}^{(1)}_{j|ik}$, $\hat{p}^{(2)}_{j|ik}$, $\hat{p}^{(3)}_{j|ik}$ are given by:

$$\hat{p}^{(1)}_{j|i1} = \begin{bmatrix} 0.6823 & 0.0581 & 0.0678 & 0.0635 & 0.0612 & 0.0671 \\ 0.8181 & 0.0425 & 0.0128 & 0.0527 & 0.0439 & 0.0300 \\ 0.1980 & 0.2006 & 0.1507 & 0.1219 & 0.1534 & 0.1754 \\ 0.1517 & 0.1584 & 0.1648 & 0.1789 & 0.1665 & 0.1798 \\ 0.1746 & 0.1544 & 0.1715 & 0.1590 & 0.1749 & 0.1656 \\ 0.2466 & 0.1449 & 0.1629 & 0.1449 & 0.1425 & 0.1581 \end{bmatrix}$$

$$\hat{p}^{(1)}_{j|i2} = \begin{bmatrix} 0.3373 & 0.1337 & 0.1297 & 0.1237 & 0.1360 & 0.1396 \\ 0.3603 & 0.0949 & 0.1364 & 0.1240 & 0.1563 & 0.1281 \\ 0.4716 & 0.1315 & 0.0848 & 0.0920 & 0.1038 & 0.1164 \\ 0.1716 & 0.1941 & 0.1642 & 0.1376 & 0.1895 & 0.1431 \\ 0.2575 & 0.1405 & 0.1831 & 0.1261 & 0.1321 & 0.1606 \\ 0.4106 & 0.1313 & 0.1331 & 0.1003 & 0.1075 & 0.1172 \end{bmatrix}$$

$$\hat{p}^{(2)}_{j|i1} = \begin{bmatrix} 0.7217 & 0.0597 & 0.0519 & 0.0526 & 0.0593 & 0.0548 \\ 0.1918 & 0.1478 & 0.1970 & 0.1603 & 0.1655 & 0.1377 \\ 0.1770 & 0.1720 & 0.1635 & 0.1529 & 0.1528 & 0.1819 \\ 0.2216 & 0.1315 & 0.1646 & 0.1543 & 0.1756 & 0.1525 \\ 0.2940 & 0.1563 & 0.1265 & 0.1431 & 0.1395 & 0.1406 \\ 0.1585 & 0.1690 & 0.1615 & 0.1822 & 0.1914 & 0.1374 \end{bmatrix}$$

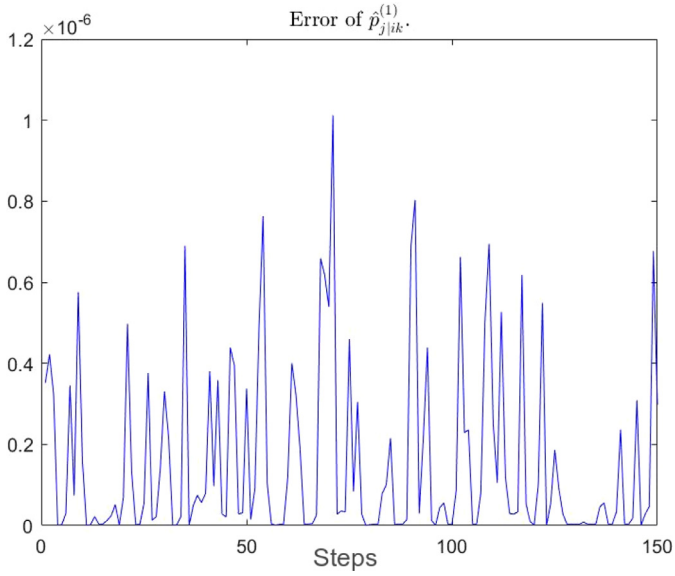

Fig. 2. Error of the estimated transition matrix $\hat{p}^{(1)}_{j|ik}$.

$$\hat{p}^{(2)}_{j|i2} = \begin{bmatrix} 0.6314 & 0.0823 & 0.0751 & 0.0679 & 0.0733 & 0.0700 \\ 0.5340 & 0.0841 & 0.0993 & 0.0879 & 0.0944 & 0.1004 \\ 0.4921 & 0.1002 & 0.1000 & 0.1068 & 0.0934 & 0.1074 \\ 0.2114 & 0.1583 & 0.1466 & 0.1627 & 0.1555 & 0.1655 \\ 0.1684 & 0.1817 & 0.1796 & 0.1670 & 0.1601 & 0.1432 \\ 0.1662 & 0.1681 & 0.1878 & 0.1520 & 0.1877 & 0.1382 \end{bmatrix}$$

$$\hat{p}^{(3)}_{j|i1} = \begin{bmatrix} 0.6656 & 0.0698 & 0.0595 & 0.0634 & 0.0664 & 0.0754 \\ 0.2279 & 0.1559 & 0.1455 & 0.1468 & 0.1710 & 0.1529 \\ 0.2986 & 0.1357 & 0.1586 & 0.1392 & 0.1367 & 0.1313 \\ 0.2510 & 0.1457 & 0.1523 & 0.1487 & 0.1512 & 0.1511 \\ 0.2184 & 0.1310 & 0.1554 & 0.1694 & 0.1721 & 0.1537 \\ 0.3613 & 0.1427 & 0.1151 & 0.1370 & 0.1297 & 0.1142 \end{bmatrix}$$

$$\hat{p}^{(3)}_{j|i2} = \begin{bmatrix} 0.5080 & 0.0949 & 0.1066 & 0.0936 & 0.0970 & 0.1000 \\ 0.3264 & 0.1294 & 0.1486 & 0.1240 & 0.1333 & 0.1383 \\ 0.5911 & 0.0879 & 0.0725 & 0.0844 & 0.0929 & 0.0711 \\ 0.1928 & 0.1483 & 0.1595 & 0.1682 & 0.1664 & 0.1648 \\ 0.5375 & 0.1092 & 0.0841 & 0.0989 & 0.0885 & 0.0818 \\ 0.1993 & 0.1725 & 0.1742 & 0.1721 & 0.1488 & 0.1330 \end{bmatrix}$$

Figs. 2–4 show the estimation function error of the transition matrix. The order of the estimation function error is about 1×10^{-6} and 1×10^{-7}, which in our case is almost 0.

The estimated values of the utilities matrices $\hat{U}^{(1)}_{ijk}$, $\hat{U}^{(2)}_{ijk}$, $\hat{U}^{(3)}_{ijk}$ are given by:

$$\hat{U}^{(1)}_{ij1} = \begin{bmatrix} 4.12 & 1.95 & 6.97 & 16.63 & 3.15 & 32.30 \\ 7.20 & 4.96 & 37.52 & 20.83 & 93.36 & 15.09 \\ 1.44 & 1.50 & 4.56 & 14.78 & 1.71 & 2.44 \\ 1.20 & 4.51 & 38.27 & 1.59 & 0.84 & 11.31 \\ 16.88 & 0.74 & 37.61 & 4.29 & 2.64 & 3.10 \\ 1.06 & 3.77 & 204.91 & 0.56 & 6.47 & 3.00 \end{bmatrix}$$

$$\hat{U}^{(1)}_{ij2} = \begin{bmatrix} 11.48 & 11.48 & 4.22 & 0.77 & 8.60 & 19.54 \\ 3.13 & 61.84 & 8.13 & 1.42 & 22.37 & 3.33 \\ 37.39 & 1.29 & 14.99 & 34.24 & 0.00 & 13.65 \\ 57.17 & 9.56 & 0.53 & 19.58 & 64.05 & 9.50 \\ 4.36 & 7.79 & 2.26 & 1.26 & 4.98 & 66.78 \\ 0.00 & 3.35 & 88.73 & 217.21 & 9.49 & 10.42 \end{bmatrix}$$

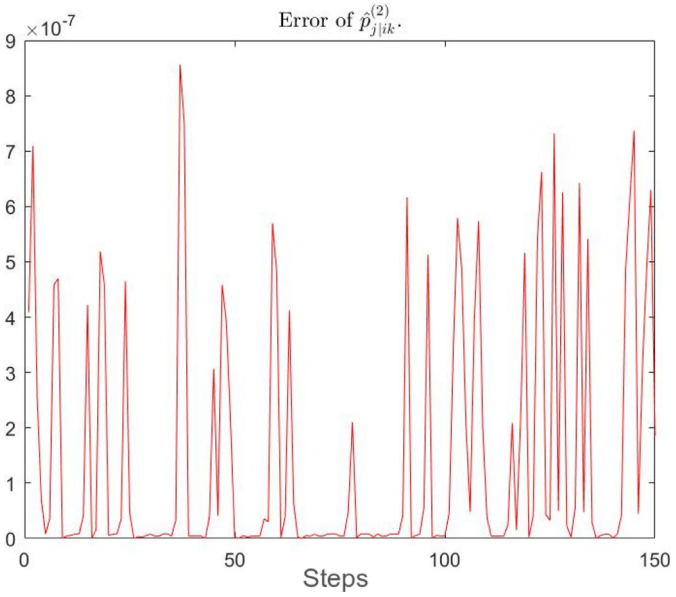

Fig. 3. Error of the estimated transition matrix $\hat{p}^{(2)}_{j|ik}$.

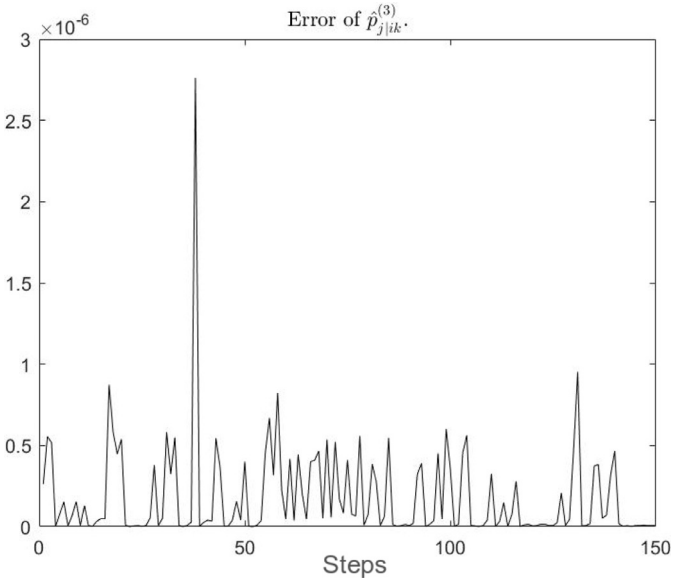

Fig. 4. Error of the estimated transition matrix $\hat{p}^{(3)}_{j|ik}$.

$$\hat{U}^{(2)}_{ij1} = \begin{bmatrix} 2.50 & 5.43 & 5.14 & 2.91 & 5.87 & 1.51 \\ 4.38 & 7.10 & 14.88 & 9.93 & 3.27 & 19.50 \\ 9.07 & 37.72 & 0.88 & 20.19 & 2.50 & 15.92 \\ 9.47 & 5.16 & 46.15 & 9.47 & 2.35 & 1.69 \\ 1.83 & 0.88 & 99.08 & 3.37 & 22.83 & 3.07 \\ 16.48 & 19.23 & 13.71 & 11.52 & 32.94 & 2.88 \end{bmatrix}$$

$$\hat{U}^{(2)}_{ij2} = \begin{bmatrix} 0.65 & 0.27 & 18.15 & 23.26 & 12.92 & 2.57 \\ 1.61 & 2.11 & 17.38 & 11.55 & 0.58 & 5.57 \\ 0.42 & 53.28 & 1.55 & 19.44 & 6.42 & 1.32 \\ 0.78 & 0.53 & 0.10 & 1.94 & 9.21 & 32.35 \\ 21.96 & 9.49 & 0.00 & 9.60 & 0.65 & 1.30 \\ 4.24 & 41.16 & 23.70 & 0.07 & 2.67 & 2.02 \end{bmatrix}$$

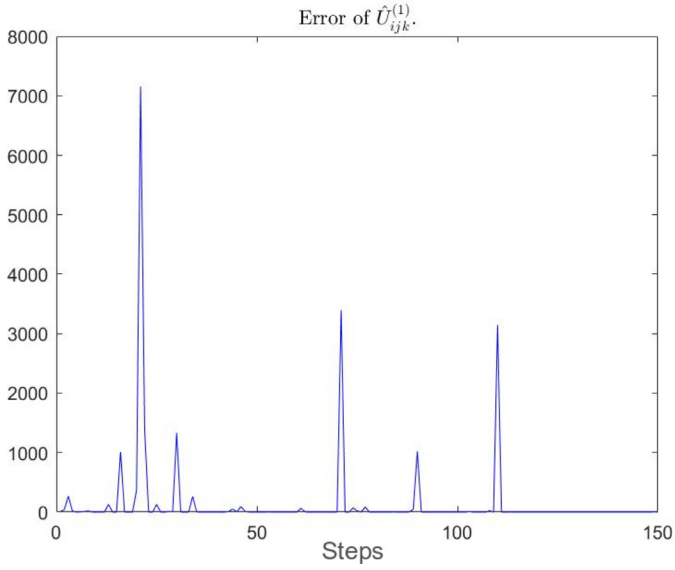

Fig. 5. Error of the estimated utility matrix $\hat{U}^{(1)}_{ijk}$.

$$\hat{U}^{(3)}_{ij1} = \begin{bmatrix} 6.29 & 1.63 & 2.67 & 20.18 & 1.80 & 6.14 \\ 2.17 & 80.88 & 13.27 & 7.84 & 7.71 & 1.05 \\ 4.16 & 87.26 & 5.32 & 3.95 & 1.50 & 21.37 \\ 7.17 & 0.19 & 48.91 & 3.03 & 2.18 & 2.40 \\ 3.33 & 27.26 & 4.13 & 18.52 & 9.26 & 5.30 \\ 35.52 & 0.56 & 3.56 & 0.77 & 7.64 & 2.88 \end{bmatrix}$$

$$\hat{U}^{(3)}_{ij2} = \begin{bmatrix} 35.37 & 3.15 & 2.10 & 1.33 & 7.88 & 26.52 \\ 4.05 & 0.39 & 21.57 & 0.46 & 1.58 & 0.73 \\ 7.24 & 0.72 & 20.52 & 14.91 & 11.38 & 11.58 \\ 28.47 & 0.35 & 10.93 & 5.03 & 0.33 & 51.40 \\ 3.69 & 2.10 & 39.70 & 0.13 & 8.24 & 20.93 \\ 18.69 & 1.92 & 1.69 & 6.82 & 42.70 & 9.49 \end{bmatrix}$$

The resulting values of the randomize strategies are as follows:

$$\pi^{(1)*}_{i|k} = \begin{bmatrix} 0.4322 & 0.5678 \\ 0.2397 & 0.7603 \\ 0.6460 & 0.3540 \\ 0.5207 & 0.4793 \\ 0.5501 & 0.4499 \\ 0.5936 & 0.4064 \end{bmatrix} \quad \pi^{(2)*}_{i|k} = \begin{bmatrix} 0.4734 & 0.5266 \\ 0.8307 & 0.1693 \\ 0.8040 & 0.1960 \\ 0.4896 & 0.5104 \\ 0.3821 & 0.6179 \\ 0.5119 & 0.4881 \end{bmatrix}$$

$$\pi^{(3)*}_{i|k} = \begin{bmatrix} 0.4627 & 0.5373 \\ 0.5837 & 0.4163 \\ 0.7322 & 0.2678 \\ 0.4519 & 0.5481 \\ 0.7543 & 0.2457 \\ 0.3682 & 0.6318 \end{bmatrix}$$

Figs. 5–7 show the estimation function error of the utility matrices which has also a monotonic decreasing behavior.

5.3. Emotional distance

Employing Eqs. (5) we obtain the joint strategies $z^{(l)*}_{i|k}$, which are given as follows:

$$z^{(1)*}_{i|k} = \begin{bmatrix} 0.1568 & 0.2060 \\ 0.0308 & 0.0976 \\ 0.0838 & 0.0459 \\ 0.0613 & 0.0564 \\ 0.0714 & 0.0584 \\ 0.0781 & 0.0535 \end{bmatrix} \quad z^{(2)*}_{i|k} = \begin{bmatrix} 0.1895 & 0.2108 \\ 0.1015 & 0.0207 \\ 0.0997 & 0.0243 \\ 0.0574 & 0.0598 \\ 0.0468 & 0.0757 \\ 0.0582 & 0.0555 \end{bmatrix}$$

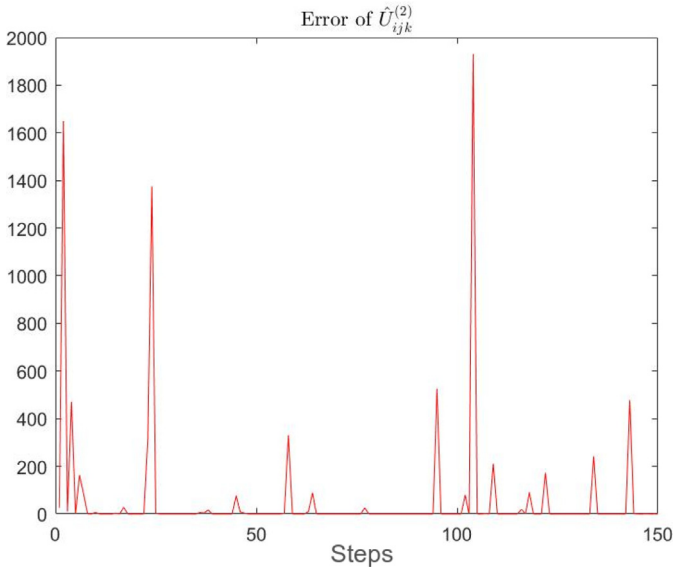

Fig. 6. Error of the estimated utility matrix $\hat{U}^{(2)}_{ijk}$.

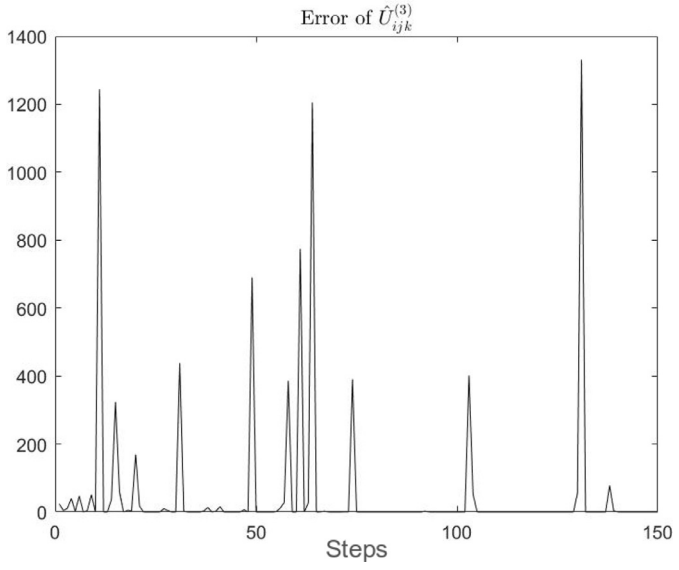

Fig. 7. Error of the estimated utility matrix $\hat{U}^{(1)}_{ijk}$.

$$z^{(3)*}_{i|k}=\begin{bmatrix}0.1871 & 0.2173\\ 0.0684 & 0.0488\\ 0.0886 & 0.0324\\ 0.0540 & 0.0654\\ 0.0911 & 0.0297\\ 0.0431 & 0.0740\end{bmatrix}$$

It is important to note that $\sum_i^N \sum_k^K z^l_{ik} = 1$. Figs. 8–10 show the convergence of the emotional strategies for Player 1, Player 2 and Player 3, respectively.

a) The assessment center evaluation shows the emotional compatible among players. Employing the Kullback–Leibler distance $D_{KL}(z^l_{ik}|z^{l'}_{ik})$ given in Eq. (10) we have that

$$\begin{aligned} D_{KL}(z^{(1)}_{ik}|z^{(2)}_{ik}) &= 0.1316 \quad D_{KL}(z^{(2)}_{ik}|z^{(1)}_{ik}) = 0.1161\\ D_{KL}(z^{(1)}_{ik}|z^{(3)}_{ik}) &= 0.0663 \quad D_{KL}(z^{(3)}_{ik}|z^{(1)}_{ik}) = 0.0624\\ D_{KL}(z^{(2)}_{ik}|z^{(3)}_{ik}) &= 0.0624 \quad D_{KL}(z^{(3)}_{ik}|z^{(2)}_{ik}) = 0.0617 \end{aligned}$$

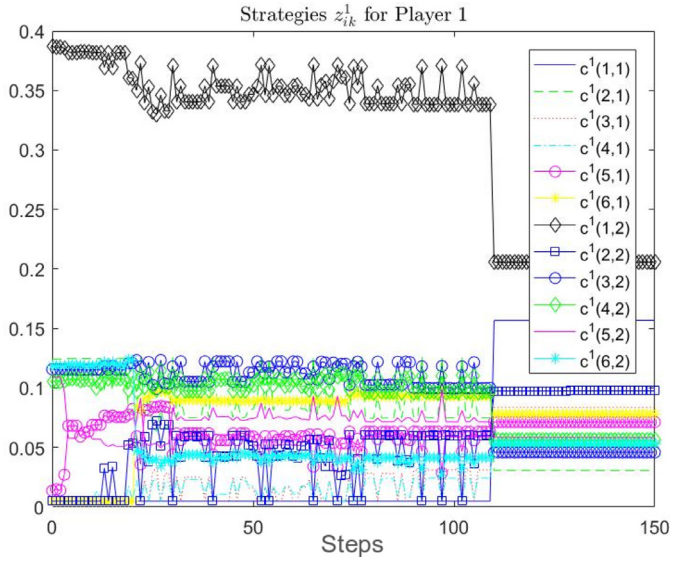

Fig. 8. Convergence of the emotional strategies for Player 1.

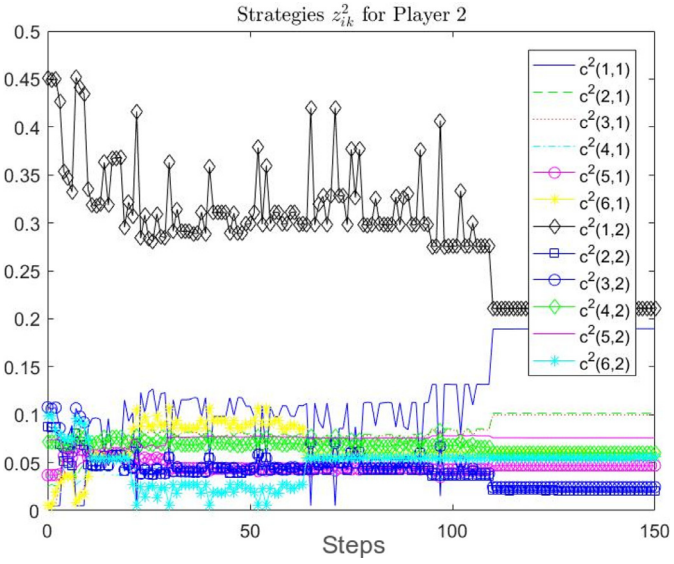

Fig. 9. Convergence of the emotional strategies for Player 2.

We conclude that Player 3 is emotional compatible with Player 1 and Players 2. However, Player 1 and Players 2 are less compatible. Interesting is to note that the Kullback–Leibler distance is asymmetric, then the feelings of one player for another are different.

b) We will define a "ideal prototype" (ideal or expected candidate for the position) in terms of an emotional strategy as follows:

$$z^{(4)*}_{i|k}=\begin{bmatrix}0.1871 & 0.2173\\ 0.0684 & 0.0488\\ 0.0886 & 0.0324\\ 0.0540 & 0.0654\\ 0.0911 & 0.0297\\ 0.0431 & 0.0740\end{bmatrix}$$

Remark 1. The ideal or expected candidate for the position $z^{(4)*}_{i|k}$ is a proposal given as a result of the evaluation of the competencies model.

We want to produce the "closeness degree" of how close each candidate (Player 1, Player 2 and Player 3) is to what is expected

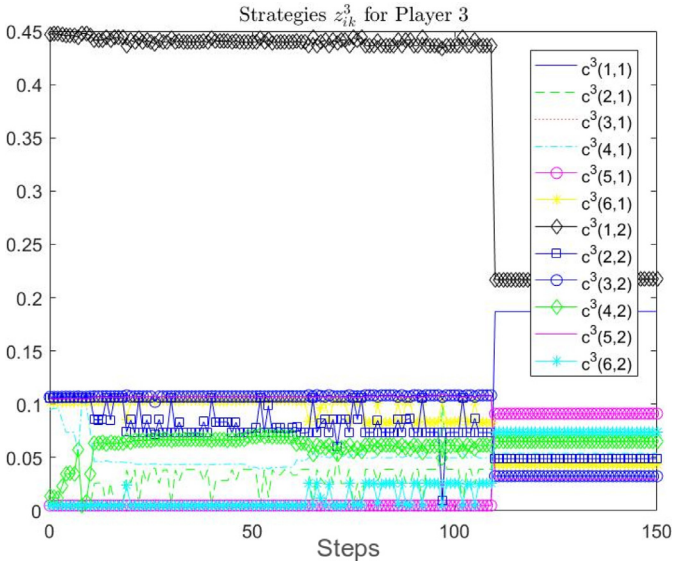

Fig. 10. Convergence of the emotional strategies for Player 3.

from an "ideal prototype" (Player 4) (Clempner, 2010). For evaluating each candidate we will employ the Kullback–Leibler distance $D_{KL}(z_{ik}^{l}|z_{ik}^{l'})$ given in Eq. (10).

$$D_{KL}(z_{ik}^{(4)}|z_{ik}^{(1)}) = 0.0736 \quad D_{KL}(z_{ik}^{(1)}|z_{ik}^{(4)}) = 0.1129$$
$$D_{KL}(z_{ik}^{(4)}|z_{ik}^{(2)}) = 0.1443 \quad D_{KL}(z_{ik}^{(2)}|z_{ik}^{(4)}) = 0.1445$$
$$D_{KL}(z_{ik}^{(4)}|z_{ik}^{(3)}) = 0.0736 \quad D_{KL}(z_{ik}^{(3)}|z_{ik}^{(4)}) = 0.0765$$

As a result, we have that Player 3 has the closeness degree to the ideal prototype.

6. Conclusion

This paper presented a new method for measuring the emotional state among interacting agents in a given environment. For solving the problem employed a non-cooperative game theory approach for representing the interaction between agents solving the game using an innovative two-step approach. In addition, we developed a RL process for introducing the stimuli to the environment. Such RL process presented a pre-training stage done in a soft-max layer-wised manner. We used the Kullback–Leibler divergence of the resulting strategies for measuring the emotional distance between the interacting agents. Because it is a distribution-wise asymmetric measure the feelings of one player for another are different. We showed the effectiveness of the proposed method presenting an application example related to assessment centers considering the six primary emotions.

In terms of future work, there exist a number of challenges left to address. One interesting technical challenge is that of investigating different measuring techniques for the different emotions. A different interesting empirical challenge would be to run a long-term controlled experiment for assessment centers, complementary to the one we present in this paper. Nonetheless, if this could be done, such an experiment would provide some very valuable insight on the application of our approach in the real-world.

References

Almahdi, S., & Yang, S. Y. (2017). An adaptive portfolio trading system: A risk-return portfolio optimization using recurrent reinforcement learning with expected maximum drawdown. *Expert Systems with Applications, 87*, 267–279.

Anderson, J., Bothell, D., Byrne, M., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review, 111*, 1036–1060.

Aylett, R., & Louchart, S. (2008). If i were you: Double appraisal in affective agents. In *Proceedings of the 7th international joint conference on autonomous agents and multiagent systems* (pp. 1233–1236). Estoril, Portugal.

Becker-Asano, C., & Wachsmuth, I. (2010). Affective computing with primary and secondary emotions in a virtual human. *Autonomous Agents and Multi-Agent Systems, 20*, 32–49.

Belkaid, M., & Sabouret, N. (2014). A logical model of theory of mind for virtual agents in the context of job interview simulation. In *2nd international workshop on intelligent digital games for empowerment and inclusion idgei* (pp. 83–90). Haifa, Israel.

Bickmore, T., & Picard, R. (2005). Establishing and maintaining long-term human–computer relationships. *ACM Transactions on Computer-Human Interaction, 12*, 293–327.

Bosse, T., Jonker, C. M., van der Meij, L., Sharpanskykh, A., & Treur, J. (2009). Specification and verification of dynamics in agent models. *International Journal of Cooperative Information Systems, 18*, 167–193.

Bosse, T., & Zwanenburg, E. (2014). Do prospect-based emotions enhance believability of game characters? A casestudy in the context of a dice game. *IEEE Transactions on Affective Computing, 5*(1), 17–31.

Bouhuys, A. L., Geerts, E., Mersch, P. P., & Jenner, J. A. (1996). Nonverbal interpersonal sensitivity and persistence of depression: Perception of emotions in schematic faces. *Psychiatry Research, 64*(3), 193–203.

Champandard, A. J. (2003). *AI game development: synthetic creatures with learning and reactive behaviors*. Indianapolis, IN: New Riders Publishing.

Chatzakou, D., Vakali, A., & Kafetsios, K. (2017). Detecting variation of emotions in online activities. *Expert Systems With Applications, 89*, 318–332.

Clempner, J., & Poznyak, A. S. (2011). Convergence properties and computational complexity analysis for lyapunov games. *International Journal of Applied Mathematics and Computer Science, 21*(2), 49–361.

Clempner, J. B. (2010). A pattern model for assessing work competencies using petri nets. *International Journal of Computer Science and Applications, 7*(4), 50–78.

Clempner, J. B., & Poznyak, A. S. (2014). Simple computing of the customer lifetime value: A fixed local-optimal policy approach. *Journal of Systems Science and Systems Engineering, 23*(4), 439–459.

Clempner, J. B., & Poznyak, A. S. (2016a). Analyzing an optimistic attitude for the leader firm in duopoly models: A strong stackelberg equilibrium based on a lyapunov game theory approach. *Economic Computation And Economic Cybernetics Studies And Research, 4*(50), 41–60.

Clempner, J. B., & Poznyak, A. S. (2016b). Convergence analysis for pure and stationary strategies in repeated potential games: Nash, lyapunov and correlated equilibria. *Expert Systems with Applications, 46*, 474–484.

Clempner, J. B., & Poznyak, A. S. (2017). Multiobjective markov chains optimization problem with strong pareto frontier: Principles of decision making. *Expert Systems With Applications, 68*, 123–135.

Core, M., Traum, D., Lane, H., Swartout, W., Gratch, J., Lent, M. V., & Marsella, S. (2006). Teaching negotiation skills through practice and reflection with virtual humans. *Simulation, 82*(11), 685–701.

Cruz, D. L., & Yu, W. (2017). Path planning of multi-agent systems in unknown environment with neural kernel smoothing and reinforcement learning. *Neurocomputing, 233*, 34–42.

Deng, S., Wang, D., Li, X., & Xu, G. (2015). Exploring user emotion in microblogs for music recommendation. *Expert Systems with Applications, 42*(23), 9284–9293.

Dias, J., Mascarenhas, S., & Paiva, A. (2014). In T. Bosse, J. Broekens, J. Dias, & J. van der Zwaan (Eds.), *Fatima modular: Towards an agent architecture with a generic appraisal framework* (pp. 44–56)). Cham: Springer International Publishing.

Ding, N., Sethu, V., Epps, J., & Ambikairajah, E. (2012). Speaker variability in emotion recognition - an adaptation based approach. In *Ieee international conference on acoustics, speech, and signal processing* (pp. 5101–5104). Prague, Czech Republic.

Ekman, P., Friesen, W. V., & Ellsworth, P. (1982). *What emotion categories or dimensions can observers judge from facial behavior?*. New York: Cambridge University Press.

El-Nasr, M. S., Yen, J., & Ioerger, T. R. (2000). Flame-fuzzy logic adaptive model of emotions. *Autonomous Agents and Multi-Agent Systems, 3*(3), 219–257.

Fisher, C., Schoenfeldt, L., & Shaw, J. (2003). *Human resource management*. Boston: Houghton Mifflin Company.

Funge, J. (1999). Representing knowledge within the situation calculus using interval-valued epistemic fluents. *Reliable Computing, 5*, 35–61.

García-Magariño, I., & Plaza, I. (2017). Absem: An agent-based simulator of emotions in mindfulness programs. *Expert Systems with Applications, 84*, 49–57.

Gratch, J., & Marsella, S. (2004). A domain independent framework for modeling emotion. *Journal Cognitive Systems Research, 5*, 269–306.

Kazemitabar, S., Taghizadeh, N., & Beigy, H. (2017). A graph-theoretic approach toward autonomous skill acquisition in reinforcement learning. *Evolving Systems*. doi:10.1007/s12530-017-9193-9. To be published.

Laird, J. E., Newell, A., & Rosenbloom, P. S. (1987). Soar: An architecture for general intelligence. *Artificial Intelligence, 33*, 1–64.

Lim, H., & Kim, H.-J. (2017). Item recommendation using tag emotion in social cataloging services. *Expert Systems with Applications, 89*, 179–187.

Lim, M. Y., Dias, J., Aylett, R., & Paiva, A. (2012). Creating adaptive affective autonomous npcs. *Autonomous Agents and Multi-Agent Systems, 24*, 287–311.

Luo, B., Zeng, J., & Duan, J. (2016). Emotion space model for classifying opinions in stock message board. *Expert Systems with Applications, 44*, 138–146.

Marinier, R., Laird, J., & Lewis, R. (2009). A computational unification of cognitive behavior and emotion. *Journal Cognitive Systems Research, 10*, 48–69.

Marsella, S., Gratch, J., & Petta, P. (2010). *Blueprint for affective computing: A sourcebook and manual* (pp. 21–46)). New York: Oxford Univ. Press.

Marsella, S. C., & Gratch, J. (2009). Ema: A process model of appraisal dynamics. *Cognitive Systems Research, 10*(1), 70–90.

Mill, A., Allik, J., Realo, A., & Valk, R. (2009). Age-related differences in emotion recognition ability: A cross-sectional study. *Emotion, 9*(5), 619–630.
Najim, K., & Poznyak, A. S. (1994). *Learning automata: Theory and applications*. Pergamon Press, Inc. Elmsford, NY, USA.
Newell, A. (1990). *Unified theories of cognition*. Harvard.
Paiva, A., Dias, J., Aylett, R., Woods, S., Hall, L., & Zoll, C. (2005). Learning by feeling: Evoking empathy with synthetic characters. *Applied Artificial Intelligence, 19*, 235–266.
Perez-Gaspar, L.-A., Caballero-Morales, S.-O., & Trujillo-Romero, F. (2016). Multimodal emotion recognition with evolutionary computation for human-robot interaction. *Expert Systems With Applications, 66*, 42–61.
Picard, R. W. (1997). *Affective computing*. Cambridge. MIT.
Poznyak, A. S., Najim, K., & Gomez-Ramirez, E. (2000a). *Self-learning control of finite Markov chains*. Marcel Dekker, New York.
Poznyak, A. S., Najim, K., & Gómez-Ramírez, E. (2000b). *Self-learning control of finite Markov chains*. Marcel Dekker, Inc.
Radac, M.-B., & Precup, R.-E. (2017). Data-driven model-free slip control of anti-lock braking systems using reinforcement q-learning. *Neurocomputing*. doi:10.1016/j.neucom.2017.08.036. To be published.
Radac, M.-B., Precup, R.-E., & Roman, R.-C. (2017). Model-free control performance improvement using virtual reference feedback tuning and reinforcement q-learning. *International Journal of Systems Science, 48*(5), 1071–1083.
Reisenzein, R., Hudlicka, E., Dastani, M., Gratch, J., Hindriks, K., Lorini, E., & Meyer, J.-J. (2013). Computational modeling of emotion: Toward improving the inter- and intradisciplinary exchange. *IEEE Transactions on Affective Computing, 4*(3), 246–266.
Sánchez, E. M., Clempner, J. B., & Poznyak, A. S. (2015). A priori-knowledge/actor-critic reinforcement learning architecture for computing the mean–variance customer portfolio: the case of bank marketing campaigns. *Engineering Applications of Artificial Intelligence, 46, Part A*, 82–92.
Schmidt, F., & Hunter, J. (1998). The validity and the utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262–274.
Smith, C. A., & Lazarus, R. (1990). *Handbook of personality: Theory and research* (pp. 609–637). New York: Guilford Press.
Solis, C. U., Clempner, J. B., & Poznyak, A. S. (2016). Modeling multi-leader-follower non-cooperative stackelberg games. *Cybernetics and Systems, 47*(8), 650–673.
Sutton, R. S., & Barto, A. (1998). Reinforcement learning: An introduction. *Introduction, MIT Press, Cambridge, MA*.
Trejo, K. K., Clempner, J. B., & Poznyak, A. S. (2015). A stackelberg security game with random strategies based on the extraproximal theoretic approach. *Engineering Applications of Artificial Intelligence, 37*, 145–153.
Trejo, K. K., Clempner, J. B., & Poznyak, A. S. (2016). Adapting strategies to dynamic environments in controllable stackelberg security games. In *Ieee 55th conference on decision and control (cdc)* (pp. 5484–5489). Las Vegas, USA
Trejo, K. K., Clempner, J. B., & Poznyak, A. S. (2017). Computing the lp-strong nash equilibrium for markov chains games. *Applied Mathematical Modelling, 41*, 399–418.
Vamvoudakis, K. G., Modares, H., Kiumarsi, B., & Lewis, F. (2017). Game theory-based control system algorithms with real-time reinforcement learning: How to solve multiplayer games online. *IEEE Control Systems, 37*(1), 33–52.
Vogt, T., & Andre, E. (2006). Improving automatic emotion recognition from speech via gender differentiation. In *International conference on language resources and evaluation* (pp. 1123–1126). Genoa, Italy.
Zhang, H., Jiang, H., Luo, Y., & Xiao, G. (2017). Data-driven optimal consensus control for discrete-time multi-agent systems with unknown dynamics using reinforcement learning method. *IEEE Transactions on Industrial Electronics, 64*(5), 4091–4100.
Zhou, Y., Haoa, J.-K., & Duval, B. (2016). Reinforcement learning based local search for grouping problems: A case study on graph coloring. *Expert Systems with Applications, 64*, 412–422.