Chapter

# The Digital Twin of an Organization by Utilizing Reinforcing Deep Learning

*Marko Kesti*

## Abstract

Chapter deals with latest knowledge on deep reinforcement learning in the context of organizational management. Article presents reinforcement learning (RL) as a tool for the manager on the path to learning winning behavior in the complex environment of organization management. Organization management has wicked learning challenges because agents are under biases that prevent understanding the phenomenon of delayed reward. Therefore, the digital simulation with RL is effective forming breakthrough learning results. Human capital management theories provide architecture in creating organization digital twin where agent can practice management actions effect on business economics and staff wellbeing. Utilizing RL algorithms, it is possible to foster behavior for creating sustainable competitive advantage – this means the Nash equilibrium between profit and staff wellbeing. In this digital twin there is AI learning assistant as a teacher that provides demonstrations on how to act so that the delayed reward is good in the future. The article explains game theoretical approach that is the foundation for creating management deep learning AI system. Human agent at the organization is playing the game of Strategic Stochastic Bayesian Nonsymmetric Signaling game in co-operative or non-cooperative way and at zero-sum or general sum game mind-set.

**Keywords:** Reinforcement learning, Digital twin, QWL, Management, Game theory

## 1. Introduction

The state-of-the-art management literature focuses on the qualitative characteristics of management, bringing empirical evidence-based models for improving organization performance. However, the management models that appear in the literature do not consider the individual complexity of organizations, thus limiting the reproducibility of good results. The organization digital twin (ODT) used in the article demonstrates the potential of RL-AI to analyze and quantify complex phenomena related to organizational behavior. In this article we study model-driven reinforcement learning AI as a new method in improving organization performance at complex environment.

There are two main categories of artificial intelligence (AI): data-driven and model-driven. Data-driven AI uses data in finding correlations and forecasting the future. In model-driven AI there is model that simulates the environment. Reinforcement learning (RL) focus is in learning and finding behavior which gives

best value and reward. When RL is utilized at model-driven AI the model simulates the behavior's effect in the value. The agent tries to learn the best behavior by following the model's reward signals. Thus, the behavior of the agent determines the result, not the data of the past.

Game Theory is a branch of mathematics that are used to model the strategic interaction between different players on a context of predefined environment. At management game theory there is predefined organization environment where the players are leaders and team members as workers or employees. Each player has incentives that drives their behavior in the game. Management game is non-symmetric because leader has specific and non-changeable characteristic compared to workers. Workers are motivated in maintaining and improving their work performance and personal self-esteem. Team leaders are motivated in maintaining and improving team performance, which is related to team leader personal profit incentives. Team leader knows that team performance is essential for achieving team profit targets. Workers know that their personal incentives will improve if their work performance is good. Thus, if there are problems at work the rational policy would be to tell the problems to supervisor so that problems can be solved. In addition, solving problems may improve workers self-esteem, having hidden psychological incentive. This organization environment form state space for strategic-Bayesian-stochastic-nonsymmetric-signaling game.

Nash equilibrium is a concept of game theory where optimal outcome is the balance where all players incentives are considered and fulfilled in optimal way. If team leader gives positive feedback for raising the possible problems, it will have positive effect on workers' self-esteem, fostering workers policy to inform the problems by signals. Solving the problems will improve group performance which foster leader's policy to encourage workers signaling game. This way workers and supervisor may find equilibrium of policies (strategies) which lead to general-sum game where optimal and sustainable team profit performance is achieved. However, this article explains why this optimal equilibrium is difficult to achieve in reality. Bersin [1] study reveal that 89% of managers think that leadership is important issue, but current leadership programs bring only minor value in improving leadership quality. This article argues that modern reinforcement learning artificial intelligence gives one solution in solving leadership challenge.

In addition to administrative role, the HR management has important function on adding competitive business value to an organization management (for example see references [2–6]). Managers need predictive measurements that indicate how business is developing and how to improve it. Human assets are essential for creating competitive advantages, thus interest in performance management has increased. Fleetwood and Hesketh [7] argue that researchers should better understand the complexity of the organization environment and seek to open a "black box" of causal relationships between human resources and organizational performance rather than offering simplified solutions.

Several studies indicate that employee psychological well-being has tendency to predict business value of an organization (for example see references [2, 8]). However, management can be confused of how to improve well-being and how much effort should be invested in well-being development at different situations to gain sufficient payback. Research reveals that organizations expect artificial Intelligence to help reducing managerial biases related to human issues and to improve productivity and employee experience [9]. Beside the hopes, researchers are also concerned that artificial intelligence may cause serious harm if the organization context is oversimplified by using data driven machine learning algorithms [10]. This article argues that AI can help solving difficult management problems

related to human biases. One of the most promising new technology is Digital Twin that uses simulation model driven AI. "To build an efficacious Digital Twin, it's important to first agree what problem needs to be solved or what opportunity needs to be explored and how accurate do the predictions need to be" [11].

Human competencies, for example leadership and working skills, have certain causalities to long term productivity. It seems that human competence has three performance-driving characteristics that can be described according to motivation theory as feelings of safety, team culture, and passion for work. It is clear that a passion for work affects a person's performance in a very different way than for example occupational safety issues. In addition, human is a psychophysical entity tied to his own situation. Therefore, the combination of all motivational drivers determines performance [12].

First, we have to study human capital productivity, which includes working time and the utilization of intangible human assets. Human intangible assets refer to performance on how effectively is the working time utilized, and how much value a person produces at each working hour. An employee may work for eight hours a day, but out of that working time, how much is actually used effectively in creating value? This basic understanding of how each employee produces value needs to be recognized before any reliable simulation analytics can be made.

## 2. Concept of organization digital twin

At this article the organization digital twin (ODT) refers to the mathematical environment that simulate organization human capital productivity. To be able to simulate the reality the digital twin must meet following requirements:

- Markov property: The future is independent of the past given the current situation.

- The environment state can be verified from measuring the reality.

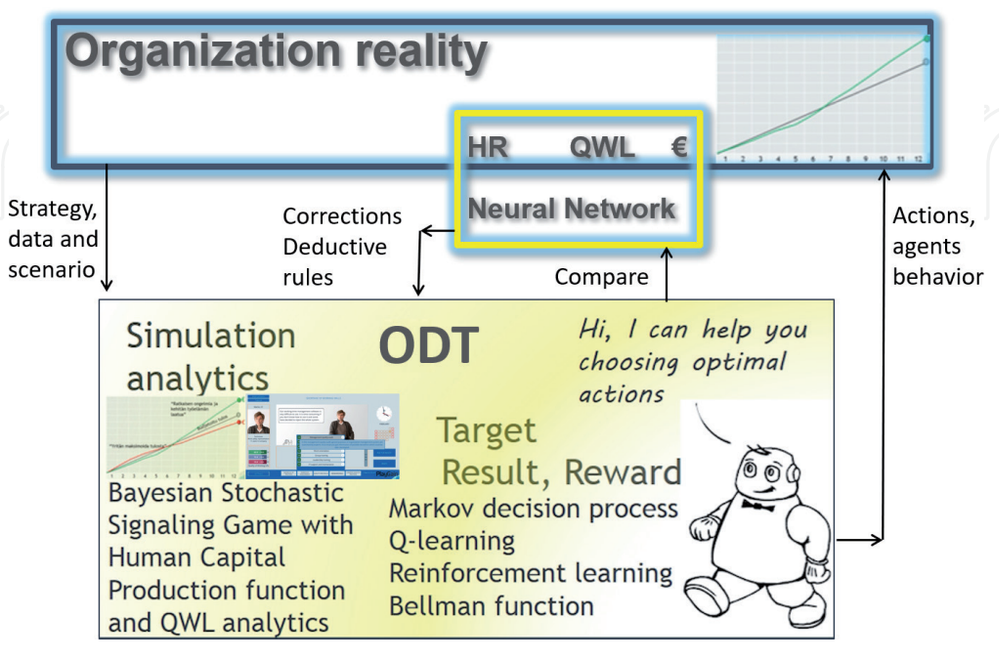


**Figure 1.**
*Concept of organization digital twin.*

Markov property means that the future is not determined by the past data, thus supervised learning regression analytics cannot be solely applied in creating ODT. Markov rule is one backbone for creating ODT digital twin and for utilizing Reinforcement Learning where the behavior of the agents determines the future.

The state transition from state to state follows Markov chain where all necessary information is transferred from past to the present. Therefore, the probability of transition from the current state to the next state depends only on the current data and the activity of the players. In the digital twin, this current data must be able to determine the reality presented by the twin. The data in the twin can be measured and verified from reality, thus creating a feedback loop from the real world. This model verification against reality is also necessary for learning purposes so that ODT can learn to refine the transition functions to match the real world (**Figure 1**).

## 3. Opening the "black-box" of human capital productivity

New science provided theoretical framework for creating ODT for modeling organization human capital productivity. First, there should understand how employees produce economic value. The theory of Quality of Working Life (QWL) determines the effective working time-share from the time spend at work. According the human capital production function the staff effective working time multiplied by K-coefficient produces customer value that is measured by revenue. The coefficient K describes the business branch, tangible investments and business logic. QWL improvement requires HR-development that increase auxiliary working time, thus reducing time for work [12] (**Figure 2**).

The human capital production function can be written in function where revenue is the production volume according the Equation [13]:

$$R = K * L * TWh * (1 - Ax) * QWL \quad (1)$$

and

$$\text{Profit} = R - \text{Variable costs} - \text{Staff costs} - \text{other costs} \quad (2)$$

Where
R = Revenue [$].
K = Coefficient for effective working time revenue relation, HR business ratio [$/h].
L = Labor capacity in full-time equivalent [pcs].
TWh = Theoretical yearly working time [h].
QWL = quality of working life, indicating human capital intangible asset utilization (0–100%).
Ax = The auxiliary working time of the total theoretical working time (vacation, absence, family leave, orientation, training, HR practices, and HRD) [%].
(1 – Ax) = (100% – Ax) = Time available for actual work (time spent at work)
(1 – Ax) * QWL = Effective working time from the theoretical working time.

It should be noted that other working time includes so-called internal error factors such as waiting, searching, correcting and unnecessary work. These are symptoms of different kinds of development needs that the team has noticed either hidden or conceptual.

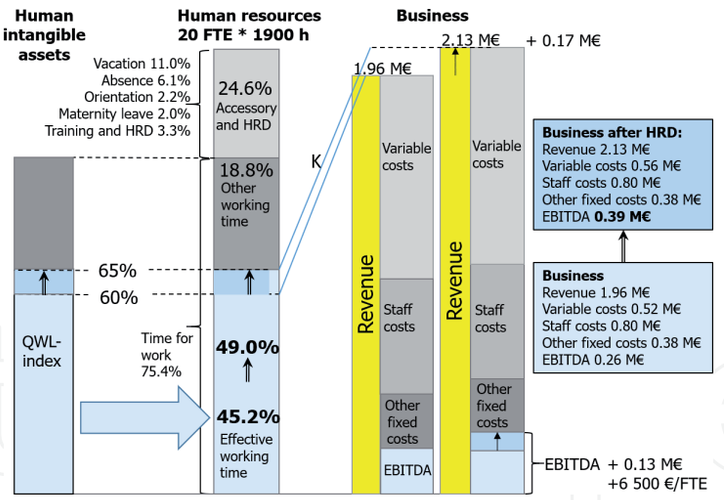


**Figure 2.**
*Illustration of profit team human capital production function.*

When manager does efficient team development there can be increase in effective working time. In addition, if absence and staff turnover is high the development may reduce those, thus increase time for work. This way effective team development will increase effective working time, and have good effect on profit. However, at short notice the development will increase auxiliary working time and reduce time for work, thus will reduce both revenue and profit. The development of human capital involves the phenomenon of investment, which requires some sacrifice in order to gain delayed rewards. When investment phenomenon is involved in the agents' actions there is possible to utilize Q-learning function.

Q-learning is a mathematical method for analyzing behavioral learning points in a simulation model that considers short- and long-term rewards. Nash equilibrium is the result where Q-learning settles to a certain level where the model environment is stable and no player can improve his pay-off [14]. In this case, equilibrium is achieved with a behavior in which both QWL and profit mature to a certain level. Both QWL and profit are management game agents' rewards, which in short-term may be contradictory because improving QWL reduces profit in short-term. This article shows that there are several states of equilibrium in a leadership game.

In most traditional well-being and commitment surveys, scores are averages of factors that are not individually relevant to the whole. Thus, the result is for example engagement index that does not necessarily tell what and how to improve and what impact the improvement would likely have. Traditional well-being surveys with average scores are oversimplified when measuring human performance. For ODT perspective, it is essential that the staff performance is determined realistically. It affects to the rewards and transition functions of agents' behaviors. Therefore, it is essential to describe the theory of QWL.

It seems evident that human performance is rather complex phenomenon, consisting several motivation theoretical aspects that cannot be included at simplified statistical staff survey analytics. Therefore, we have utilized motivation theories of Alderfer [15], Antonovsky [16], Kano [17] and Herzberg [18] in creating advanced human performance theory that meets the contribution of main scientists and forms practical QWL index for performance analytics. QWL index includes three self-esteem categories, which each has unique effect on performance.

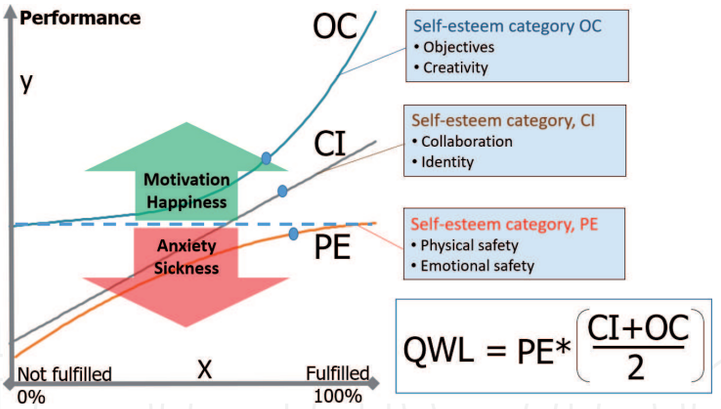


**Figure 3.**
*The theory of QWL.*

The self-esteem categories:

- Physical and emotional safety (PE);
- Collaboration and identity (CI); and
- Objectives and creativity (OC).

Chosen categories and their effect on performance form the theory of QWL index. It is also important to know that in addition that QWL index is production parameter, it has also logical connection to customer satisfaction (see [17]) (**Figure 3**).

Finally, the QWL index is the combination of all three self-esteem factors according the following equation:

$$QWL = PE(x_1) * \left[ \frac{\left( CI(x_2) + OC(x_3) \right)}{2} \right] \tag{3}$$

where

QWL is calculated using the quality of working life index (0 ... 1).

$PE(x_1)$ is the value of the function of physical and emotional safety.

$CI(x_2)$ is the value of the function of collaboration and identity.

$OC(x_3)$ is the value of the function of objectives and creativity.

The functions of the self-esteem categories are adjusted so that the final QWL result is always between 0 and 100% [12].

## 4. Defining management-game for ODT

Bayesian theorem with stochastic game is utilized in defining management-game for ODT. Using game theory, we can model the strategic interaction between different players (agents) in a predefined environment. Our management-game is multi-agent game for the profit unit where the agents are workers and manager. The concept is non-symmetric because manager (team leader) and workers have different roles and their reward characteristics differ. Workers are motivated to maintain and improve their self-esteem (QWL). In addition, there might be some hidden motivation drivers. Team leader motivation drivers are unit profit and possible personal incentives, which may be hidden (e.g. biases). In our game the focus is profit-unit manager's behavior and learning.

At Nash equilibrium the optimal outcome of the game is one where no agent wants to deviate from the chosen policy because that seems to be parallel with opponents' policy. Workplace problems have reducing tendencies on workers' self-esteem, thus decreasing QWL as a production parameter. Management practices have tendencies to improve QWL, but each action will reduce short-term profit. Manager's strategy hypothesis guides the actions at different state events. When the consequence data of action tendencies update the status after each Markov-sequence, the player can update the management strategy, which further controls the next actions. Bayesian probability is related to player subjective behavior, relying on the phenomenon that rational thinking will probably lead to optimal result as the new information comes available [19].

The manager should learn the optimal leadership strategy without knowing the exact reward function or state transition function. This approach is called stochastic model-free reinforcement learning and can be defined with the Nash Q-learning approach. The leader has prior-believe about the state of nature of profit-unit business situation and expected future reward. The uniqueness of the game comes from the fact that it has predictive features that allow for the use of reinforcing learning artificial intelligence for learning Nash equilibrium between staff QWL and organization profit.

Management game is signaling game since workers give essential signals about possible workplace problems that may threaten their self-esteem (QWL) and therefore team performance. Workers preference strategy is to give their leader signals about the problems. In simplified digital team leaders' learning-game the worker's strategy may be stationary, meaning that workers behavior may be chosen in advance when the events scenario is known.

Team leader, as an agent of the management game, is responsible for team profit performance that is the outcome of producing customer value measured by revenue. Agent registers workers' signals and makes own prior belief for the strategy. Agent monitors also scorecards from business outcomes of monthly and cumulative profit, and forms a prior believe policy on how to act to these measures. Agent is rewarded by the profit at each month and cumulative profit at the end of the year. After each state transition the agent will get profit signals and QWL signals from the worker's response from the state change at workers QWL. State-change signals and reward results may cause changes at the preference strategy of the agent for the next sequence (Markov sequence [19]) (**Figure 4**).

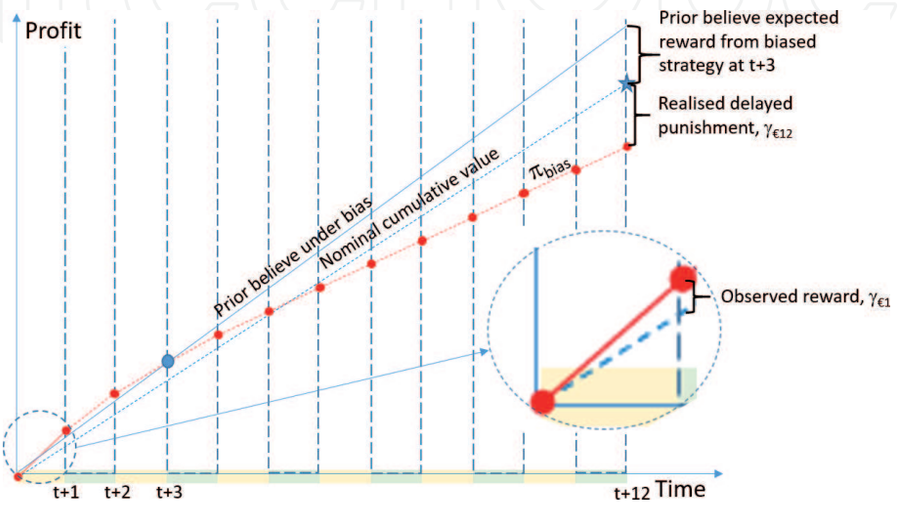


**Figure 4.**
*Leader's prior believe is biased and this strategy leads to delayed punishment.*

Leader reward function is ($\gamma_L$) the combination of monthly profit change, and expected affect to future profit. $\pi_{Li}$ is the leader's strategy at current state (month). It seems that at the beginning, the leader strategy is weighted at the monthly profit and the expected future reward is based on simple linear regression of data achieved so far. This means biased prior believe where the expected reward is not nearly the same as the outcome of the strategy. Thus, the value function under biased strategy is the following:

$$\gamma_L = \gamma_{t€} + \gamma_{12€} \tag{4}$$

where
$\gamma_{t€}$ is the observed state reward.
$\gamma_{12€}$ is the expected future reward.

When the leader gets more experience and learns to understand the complexity of the system as well as the meaning of workers' QWL, the prior believe value function changes. QWL change starts to be more interesting, because leader learn to expect more future profit when QWL improves. Thus, along this information the leader adjusts the strategy for optimizing cumulative yearly profit. Here the leadership game stochastic nature is key to learning the Nash general sum equilibrium between the QWL and profit.

$$\gamma_L = \alpha_t \left(\gamma_{t€} + \gamma_{12€}\right) \tag{5}$$

where
$\gamma_{t€}$ is the observed state reward.
$\gamma_{12€}$ is the expected future reward by improving the QWL.
$\alpha_t$ is the learning rate.

QWL is improved by leadership actions that reduce the monthly working time for making the revenue. Thus, improving QWL reduces monthly revenue and profit, but may increase effective working time in the future and so increase the future profit. In monthly basis, this phenomenon may be contradictory and confusing, but by practice, the best reward is achieved where both workers' and leader's payoff functions flourish. This means the Nash equilibrium where yearly QWL is improved with high profit. In Nash equilibrium, leader's choices are the best response to the workers' signals and business cumulative outcome at the end of the year.

Bayesian stochastic strategic non-symmetric signaling learning game follows Markov decision process [20–22]. Management-game forms stochastic game tuple

$$[N,S,C,A,T,P,R] \tag{6}$$

where
N is set of players, i.
S is set of states, s.
C is set of competences at actions a.
A is set of actions, a.
T is set of signals, $\tau$.
P is transition probability function; P: S x A x C thus P(s, c, a), $\rho$:SxA|C $\rightarrow \Delta$ is the transition function, where $\Delta$ is the set of probability distributions over state space S.
R is reward function, R = r1,...rn, $\gamma$:SxA|C $\rightarrow$ R.

There is incomplete but perfect information. The agents (workers and leader) do not know other agents' payoff functions in detail, but they can observe other agents'

immediate payoffs and actions from past months. A leader does not know exactly which actions would be the best but can choose actions that should be good enough. The leader will get workers emotional feedback immediately and information from profit monthly change and cumulative reward. After several game rounds, the player (leader) will learn the optimal actions to improve both the QWL and annual profit. Thus, the player will achieve the Nash equilibrium of stochastic Markov learning game.

## 5. Management-game Markov sequences

Management-game has context specific Markov-sequences. State and state change transition follows the Markov property where the future is independent of the past given the current situation. Once a state is defined, its change is determined by the behavior of the parties. State change is sequential, following the players actions and state transition probability function. Sequences are:

First Month (January)

1. Workers interpret the state situation and give signals based on prior believe ($\tau$).

2. Leader observes the signals and updates the signal-strategy ($\pi_\tau$).

3. Leader updates standard-strategy ($\pi_{st}$). Note: at this first month there is no data to update this year profit strategy.

4. Leader makes actions (or decide doing nothing) (a)

5. Actions leads to state change with possible outside intervention (stochastic)

6. Leader observes immediate ($\gamma_{€1}$) and cumulative ($\gamma_{\Sigma€}$) profit rewards (or sacrifices). From now on, the leader gets also profit outcome, thus updates also profit strategy.

7. According the combination of rewards, the leader upgrades prior believes concerning own behavior

8. Leader upgrades profit-strategy and standard-strategy for choosing actions t + 1

9. Workers give signals to be considered when deciding actions t + 1

10. Leader updates signal-strategy for choosing actions t + 1

11–13. Leader makes actions t + 1 in line with all three strategies.

14. State transition to state t + 1

15... From now on, the supervisor should update all three strategies simultaneously as learning sequences progress (**Figures 5** and **6**).

Leadership game Q-learning function is (7)

$$Q_{t+1}(s,c,a) = (1-\alpha)Q_t(s,c,a) + \alpha\left[\gamma_{\Delta i} + \gamma_{\Delta 12} + \beta \max_a Q_t(s_{t+1},\ c,a)\right] \qquad (7)$$

where
β is [0,1] is discounted reward factor
$\alpha_t$ is [0,1] is the learning rate (1-$\alpha_t$)
$\gamma_{\Delta i}$ is the monthly profit reward
$\gamma_{\Delta 12}$ is the expected cumulative profit reward (floating 12 months)
$\max_a Q_t(s_{t+1},,c,,a)$ is expected maximum equilibrium strategy state change pay from best actions a at competence levels c.

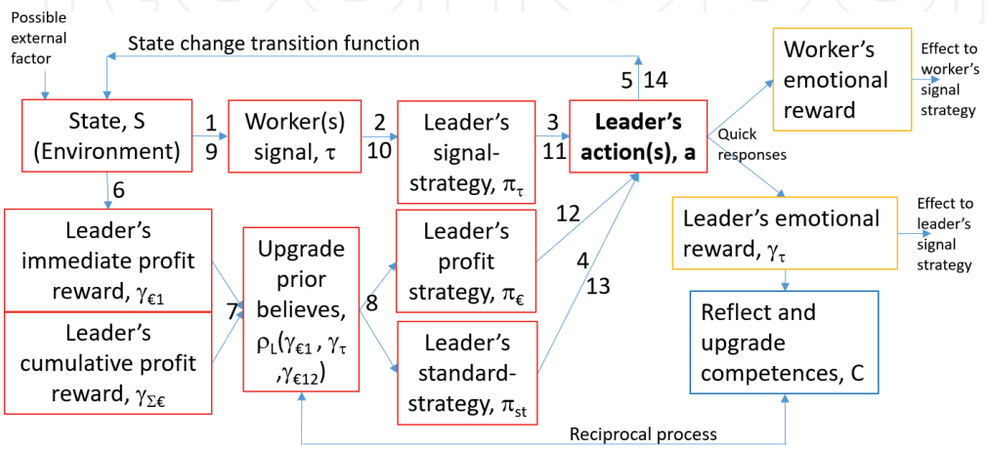


**Figure 5.**
*Management-game Markov sequences.*

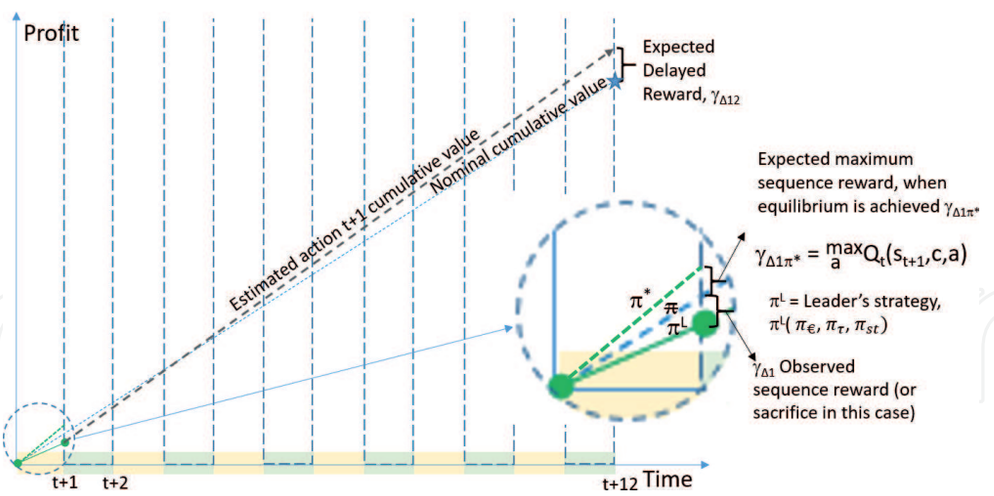


**Figure 6.**
*Management game learning phenomenon for finding equilibrium.*

With expected equilibrium strategy pay ($\max_a Q_t(s_{t+1},,c,,a)$) you can calibrate the Q-learning points so that it gives approximately 0-points when no actions are done, thus no learning was achieved. In our Q-learning function this value is 221 € monthly improvement value per employee. This corresponds the costs of one absence day per month for each worker or one working day more in work efficiency. Using this value, Q-learning gives 0 points regardless of what the supervisor's skills are.

## 6. Management learning strategies

There are three different strategic areas of prior-believes that forms the manager's learning context. These strategies are influenced by the supervisor's interaction skills (competences), which tend to either promote or hinder learning in the area. Every manager has personal competences, which seems to form personal Nash equilibrium and corresponding Q-learning results. According to this article, it seems that Nash equilibrium is different for each combination of manager's competences. In addition, the leader's strategic mind-set defines the equilibrium. Indeed, management equilibrium seems to be evolving phenomenon, depending on organization and its' players change of characteristics (**Figure 7**).

The focus of the signal-strategy ( $\pi_{\hat{o}}$ ) is to learn to understand employee signals and utilize them to achieve best reward. This strategy is strongly related to the psychological agreement between workers and supervisor. When working team members learn to play general-sum-game, the signals are provided early and in constructive way, which foster optimal actions. In case signal-strategy turn to 0-sum-game the signals tend to be hided or used to harm other members of the team. Thus, creating best foundation for signal-strategy is grounded on continuous fostering of psychological agreement at the working society.

Profit-strategy ( $\pi_{€}$ ) focus is to learn from experience how target profit is achieved at anticipated time span. Economical profit indicators are usually constantly monitored, giving them a lot of attention. In addition, organization profit target time span is determined at management system, which create certain predefined attitude towards achieving profit. From a strategic point of view, there is a big difference between focusing on maximum result this month or aiming for the maximum profit with delay of several months. If a management system requires maximum results over a short period of time, then it reinforces the detrimental profit-maximization bias. In this bias the team-leader tend to push workers performance too much, which lead to maximizing performance that is declining. In addition, a manager under this bias neglect employee signals because the signals pose a risk that short-term profits are threatened when scarce working hours are used to solve the problem. Clearly, this behavior damages the signal game, as employees learn that problems are not worth reporting.

The focus of the standard-strategy ( $\pi_{st}$ ) is to learn how to plan actions in advance to secure the reward in the future. Usually this strategy comes from the

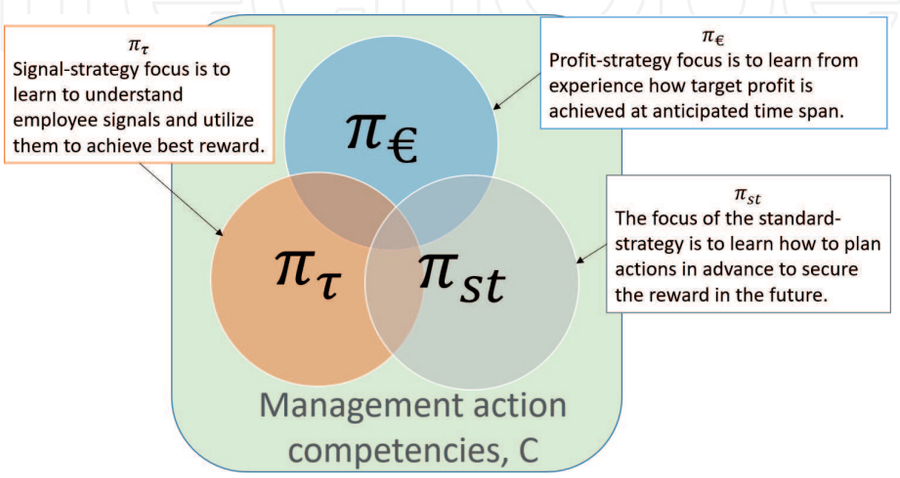


**Figure 7.**
*Management learning strategies.*

organization's human resources management, which recommends the implementation of certain management practices according to the annual plan. In practice it is common that this recommended plan is followed in various ways – some managers follow the plan while others do not. Those who do not follow the plan are likely to have learned good reasons why the recommended measures are not be implemented. Approved defense excuses may be related to the lack of time, because profit target needed all the focus. Clearly, this behavior damages the benefits of good standard-strategy.

All of these supervisor strategies are built on the supervisor's personal and ever-evolving managerial skills. In this management game theoretical approach there are personal leadership action competencies that determine the effect of each action. There is interaction between management competencies and learning strategies. The supervisor reflects the effectiveness of his or her own leadership behavior and changes personal management strategies accordingly.

## 7. Digital twin AI advisor using Bellman function

Digital twin advisor uses Bellman [20] expectation function in finding optimal actions for achieving Nash equilibrium. Bellman expectation function for strategy π is

$$v_{\eth}(s) = E_{\pi}\left[R_{t+1} + \beta v\left(S_{t+12}\right)\right] \tag{8}$$

where

$R_{t+1}$ = immediate reward

$\beta v\left(S_{t+12}\right)$ = discounted future value (12 months estimation).

Optimal policy forms from the actions that result in optimal value function, thus

$$q_{\eth *}(s,,c,,a) = R^{s}_{\eth *} + \beta^{arcmax}_{a\in A} v_{\pi *}\left(S_{t+12}\right) \tag{9}$$

where

$R^{s}_{\eth *}$ = immediate state reward from strategy π*

$\beta^{arcmax}_{a\in A} v_{\pi *}\left(S_{t+12}\right)$ = discounted maximum future value (12 months estimation).

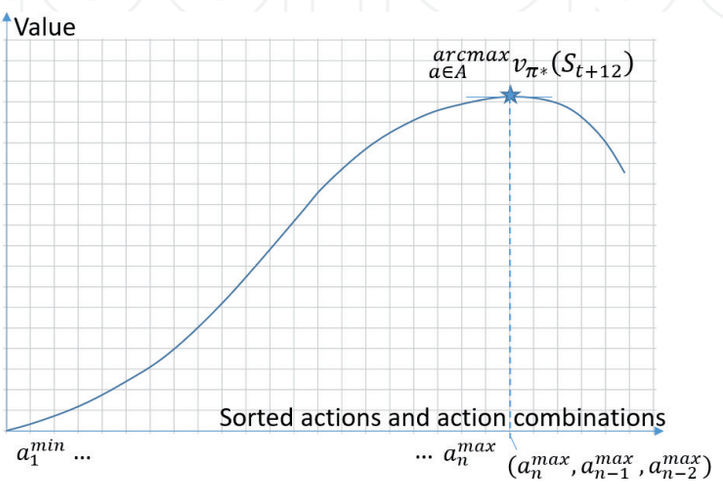


**Figure 8.**
*Bellman function principle of marginal productivity value.*

In our digital twin AI assistant is using Bellman function. It returns the combination of actions that gives the best value after floating 12 months. This is achieved so that first each action value is analyzed and sorted in magnitude of the value. Then the combinations of best actions are evaluated until marginal productivity of the value is achieved, see example at **Figure 8**. One simulation episode is 12 months; thus, the Bellman function maximize future reward even when the episode is coming to end.

Simulation game is done using UNITY 3D, for making possible to play the learning game episodes. Each episode is 12 months, consisting several workplace challenges. In the test runs we used Cash Cow episode where problems are easy, the market situation is steady, and the company does not seek special increase in revenue. State space problems are signaled by the workers that comes meeting the team-leader (agent). In this ODT there is so far 25 workplace challenges which reduce QWL according situational probability matrix. Leader has 32 best management practices (action space) that may be used as the leader prefers. Each action reduce profit and may improve QWL according state space situation specific probability function [23] (**Figure 9**).

We tested simulation using three different competence values; 30%, 60% and 90%. **Table 1** contains the results of three simulation rounds as follows:

- BIAS = human simulation episode (round) with bias to maximize short term profit. Only problem-solving actions are made. In BIAS episode the focus is on maximizing short-term profit.

- Learning = human simulation episode where leader has learned to maximize best result in QWL and profit. Agent execute best learning strategy (see **Figure 7**) with long-term profit mind-set, problems solving as good as possible and following yearly management-plan of actions.

- Bellman = artificial intelligence episode where all actions are chosen according Bellman function (see **Figure 8**).

It seems that with management competence levels 30% there are difficulties to achieve budgeted target result in profit. If QWL is sacrificed for short term wins, the cumulative profit result at the end of the year will be poor. It seems that in

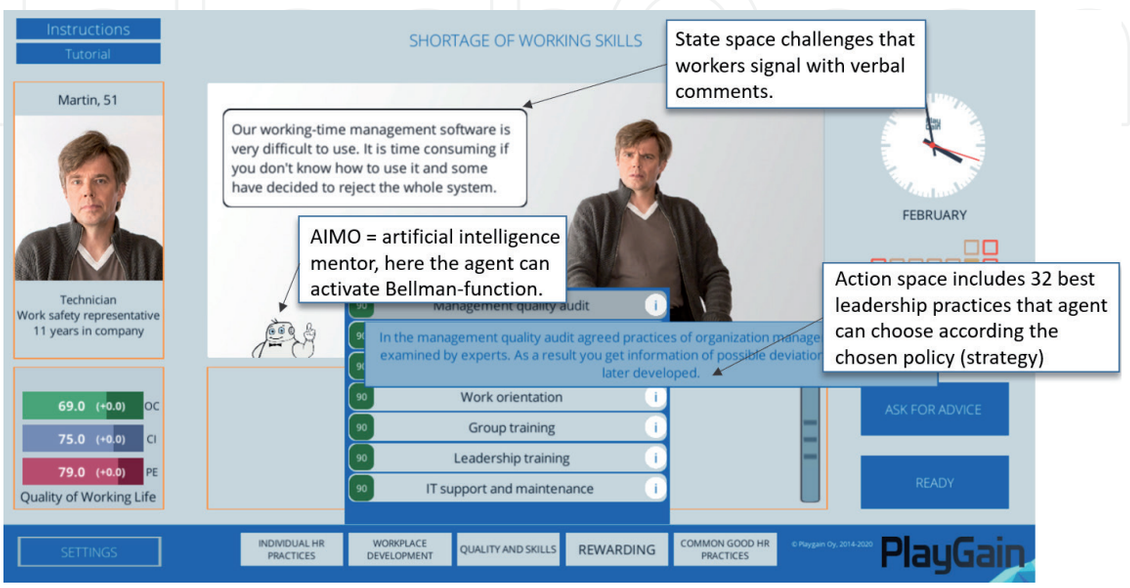


**Figure 9.**
*Simulation game user-interface.*

| | Q-learning | QWL start | QWL end | QWL difference | Cumulative | | Profit difference | Equilibrium in 1 y. |
|---|---|---|---|---|---|---|---|---|
| | | | | | Budj. € | EBITDA | | |
| Competence 30%, BIAS | 3 310 | 60,2% | 57,9% | -2,3% | 254 923 | 244 921 | −10 002 | — |
| Competence 30%. Learning | 5 370 | 60,2% | 64,6% | 4,4% | 254 923 | 243 650 | −11 273 | yes |
| Competence 30%, Bellman | 21412 | 60,2% | 67,5% | 7,3% | 254 923 | 257 070 | 2 147 | yes |
| Competence 60%, BIAS | 5 854 | 60,2% | 59,0% | −1,2% | 254 923 | 263 284 | 8 361 | yes |
| Competence 60%, Learning | 20 425 | 60,2% | 68,3% | 8,1% | 254 923 | 287 083 | 32 160 | yes |
| Competence 60%, Bellman | 35 931 | 60,2% | 70,4% | 10,2% | 254 923 | 293 442 | 38 519 | no |
| Competence 90%, BIAS | 7 737 | 60,2% | 59,8% | −0,4% | 254 923 | 276 828 | 21 905 | yes |
| Competence 90%, learning | 31 240 | 60,2% | 69,9% | 9,7% | 254 923 | 305 604 | 50 681 | no |
| Competence 90%, Bellman | 38 446 | 60,2% | 70,1% | 9,9% | 254 923 | 312 003 | 57 080 | no |

**Table 1.**
*Test episode values.*

one-year simulation episode there is achieved equilibrium where Q-learning points and QWL values are not exceeding. At 30% competence levels the BIAS episode Q-learning points varies between 0 and 3000 points. It seems as if the agent has no idea of how to achieve sustainable development where both QWL and profit improves. With low competence levels only with Bellman decisions will the profit slightly exceed the target value.

Manager's competence levels 60% are quite realistic, representing average line-managers leadership-action skills. In one-year simulation both BIAS and Learning strategies achieve Nash equilibrium, however in different profit outcome. At BIAS strategy the QWL is set at level 60%, which actually corresponds workforce medium QWL value in Finland [24]. When equilibrium is achieved, it may be difficult to change the behavior (see **Figure 10**).


**Figure 10.**
*BIAS strategy Q-learning points.*

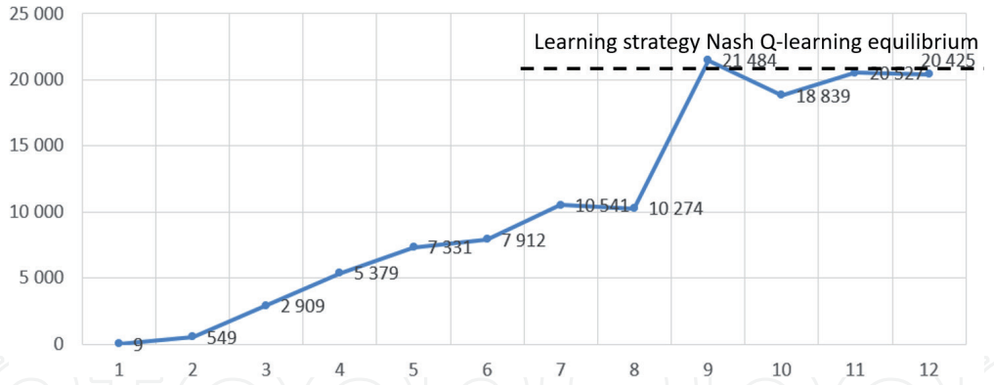


**Figure 11.**
*Learning strategy Nash Q-learning equilibrium.*

Learning strategy has also equilibrium at competence level 60%, but higher QWL and profit values than in BIAS (see **Figure 11**). In our practical simulation studies this type of results are usually learned when simulation episodes are practiced over ten times. Must bear in mind that management systems have tendency to press maximizing short-term profits, thus remaining in BIAS mind set. Learning to be excellent leader requires several years practice in organization system that allows investing in people. This phenomenon may explain why some leaders learn to be excellent team-leaders while majority remains at lower level.

There is interesting phenomenon at 90% competence level BIAS strategy. Even with very high leadership skills the QWL is set at 60% where equilibrium remains. This is due to the behavior where leadership actions are implemented only when problems arise, thus there are no proactive investments in team development. In competence levels 90% it seems that one-year simulation episode is not enough time to achieve perfect equilibrium at Learning and Bellman strategies, since Q-learning points and QWL seems to continue improving throughout the episode. It would need longer time period to achieve equilibrium.

BIAS strategy seems to achieve equilibrium where QWL is no longer improved and the Q-learning points finds management cultural maximum value. The lower the competence, the lower the level of QWL, however the difference is not so big, varying from 57% to 60%. This is interesting because in Finland the workforce medium QWL is around 60% [24]. One could argue that the profit maximization bias is common and not depending on line-managers leadership competences, and therefore most employees feel the QWL is around 60%. Moreover, the reason for profit maximization bias is not necessarily a lack of leaders' skills, but a management system that forces leaders to focus on short-term profit rather than people.

## 8. Conclusions and discussion

Organizational management research has typically focused on qualitative behavioral factors that have a complex relationship to organizational success, and in addition, impacts often come with a delay. Each organization is a unique system with certain same laws, but also a unique context of its own. Therefore, repeating the empirical research results has proven to be challenging, which also makes it difficult to draw generalizable conclusions [7]. This article examines the utilization of model-based artificial intelligence in management development. ODT can be used to assess the impact of management behavior on an organization's success, considering situational data and the impact of management culture. ODT helps to explore

the fundamental nature of an organization, which means a metaphysical essence in where everything affects everything.

The article uses artificial intelligence to illustrate how leadership behavior can create a so-called QWL glass roof that invisibly prevents teams from growing to the top performing category. The management system forms the behavior of supervisors in such a way that harmful biases of management thinking may occur, in which case people's performance does not develop favorably. These harmful biases of thought are very complex as they include phenomenon of delayed effects on an organization's competitiveness. Model-driven reinforcement learning artificial intelligence reveals a variety of human and complex mechanisms that hinder the development of competitiveness.

Reinforcement learning is following rational learning phenomenon, where learning take place gradually, according the experience. Simulation model provides learning platform where person can learn without fear of remorse. This is essential especially for managers, because in real life there is hardly room for learning from mistakes. The ODT models the situation with the organization's own data. The simulation can be designed according to the company's own strategy, allowing future challenges to be practiced. This allows management and supervisors to adapt in advance and prepare for future challenges. More proactive management reduces the realization of personnel and business risks and adds value to performance. For example, adapting to a recession can be practiced, as can market growth, both of which require a different way of managing. Artificial intelligence combined with the digital twin helps to emphasize leadership skills and practices that lead to sustainable development.

ODT has been used in college students' leadership studies. Learning outcomes have been monitored through self-assessments, and the results are encouraging. Gamified simulation-learning is based on reinforcement learning, where progress takes place through experiential adaptation according to the student's capabilities and learning ability [25]. ODT is also used in managerial trainings for companies and municipal organizations. Perhaps the biggest challenge in coaching supervisors in working life is unlearning the biases that prevent leadership success. Traditional teaching is largely based on sharing best knowledge, where the teacher shares information on how to act and why to behave in a certain way. The power of digital simulation teaching is based on the fact that it adapts the brain through experiential learning. When a supervisor has to change the prevailing leadership attitude, he or she kind of adapts the brain to another frequency where listening and caring for employees rises higher in priorities. In this way, the supervisor becomes interested in developing herself in interaction practices where she may not have previously felt the need to learn.

The architecture of the digital twin models the reality of an organization with relatively good accuracy, which is important in building trust in an artificial intelligence solution. The core of the model is in the Human Capital Production Function of and in the scientific research of the Quality of Working Life index [26]. The architecture lays the foundation for a neural network that has been fine-tuned with the probabilities of empirical research as well as correlations created through supervised learning. For example, the physical and emotional safety (PE) of the QWL index correlates with sickness absence, so that when the PE factor falls, sick leaves increases. The correlation is brought into the digital twin, which makes the model more accurate because it also models sick leaves. In addition to research data, the digital twin can be calibrated with data from the organization. ODT learning can be extended in the organizational hierarchy to the level of an individual supervisor. In this way, artificial intelligence learns the strengths and weaknesses of a leader, so that the advice given by artificial intelligence is targeted at each supervisor.

Supervised learning AI that is based on data alone is unable to "understand" organizational complexity and phenomenon of delayed impact relationships. In fact, there is a word of warning in using simple data-driven AI in complex organization environment, because it may strengthen the harmful behavioral biases. Article indicates that ODT with Bellman algorithm can be used in finding organization specific optimal behavioral patterns and measures which will form sustainable competitiveness. The article suggests that in the future, top-tier companies will use RL artificial intelligence to support management decision-making.





## Author details

Marko Kesti
University of Lapland, Rovaniemi, Finland

*Address all correspondence to: marko.kesti@ulapland.fi

## References

[1] Bersin J., Geller J., Wakefield N. and Walsh B. (2016). The new organization: Different by design. In Global Human Capital Trends study available at https://www2.deloitte.com/us/en/insights/focus/human-capital-trends/2016/human-capital-trends-introduction.html

[2] Kesti, M. (2012). The tacit signal method in human competence-based organization performance development. Rovaniemi: Lapland University Press.

[3] Pfeffer, J. (1994). Competitive advantage through people: Unleashing the power of the workforce. Boston: Harvard University Press.

[4] Ulrich, D. (1997). Human resource champions, the next agenda for adding value and delivering results. Boston: Harvard Business School Press.

[5] Guest, D. (1997). Human Resource Management and Performance: A Review and Research Agenda. International Journal of Human Resource Management, 8(3), 263-276.

[6] Becker B., & Huselid M. (2006). Strategic human resources management: Where do we go from here? Journal of Management, 32, 898-925.

[7] Fleetwood S., & Hesketh, A. (2010). Explaining the performance of human resource management. Cambridge University Press. Wright & Bonett, 2007.

[8] Wright, T. A., & Cropanzano, R. (2004). The role of psychological well-being in job performance: A fresh look at an age-old quest. Organizational Dynamics, 33(4), 338-351.

[9] Bourne Vanson (2018). Future of Work study of 4,600 business leaders from companies with 250+ employees, across 40+ countries and 12 industries, Dell Technologies. Available at https://www.delltechnologies.com/en-us/perspectives/future-of-work.htm

[10] Natarajan Balasubramanian, Yang Ye and Mingtao Xu (2020). Substituting Human Decision-Making with Machine Learning: Implications for Organizational Learning, Academy of Management Review VOL. 0.

[11] DellTechnologies (2020). The Future of Digital Twins at https://www.delltechnologies.com/en-us/perspectives/the-future-of-digital-twins/, December 1, 2020

[12] Kesti, M., Leinonen, J. and Syväjärvi, A. (2016). A Multidisciplinary Critical Approach to Measure and Analyze Human Capital Productivity. In Russ, M. (ed.). Quantitative Multidisciplinary Approaches in Human Capital and Asset Management (pp 1-317). Hershey, PA: IGI Global. (1-22).

[13] Kesti, M. and Syväjärvi, A. (2015). Human Capital Production Function in Strategic Management. Technology and Investment, 6, 12-21.

[14] Nash J. (1951). Non-Cooperative Games. The Annals of Mathematics, Second Series, Vol. 54, No. 2, pp. 286-295.

[15] Alderfer, C. P. (1967). Convergent and discriminant validation of satisfaction and desire measures by interviews and questionnaires. Journal of Applied Psychology, 51(6), 509-520.

[16] Antonovsky A. (1979). Health, stress and coping. San Francisco: Jossey-Bass.

[17] Kano, N. (1984). Attractive quality and must-be quality. Journal of the Japanese Society for Quality Control, April, 39-48.

[18] Herzberg, F., Mausner, B. and Snyderman, B., B. (1959). The motivation to work, Second edition, John Wiley & Sons, New York.

[19] Watkins C. and Dayan P. (1992). Q-learning. Machine learning, 3: 279-292.

[20] Bellman, R. (1957). "A Markovian decision process". Journal of mathematics and mechanics. 6(5), 679-684.

[21] Littman M.L. (1994). Markov games as a framework for multi-agent reinforcement learning, Proc. 11th international conference on machine learning, New Brunswick, pp. 157-163.

[22] Sutton R. S. and Barto A. G. (1998) Reinforcement learning: an introduction, MIT Press/Bradford Books.

[23] Leadermind simulation learning game introduction. (2020). Playgain Inc. Available at www.leadermind.fi

[24] Savukoski T. (2017). Työelämän laadun indeksin yhteys työkyvyttö-myyseläkeriskiin. ProGradu, Lapin Yliopisto.

[25] Kesti, M. , Ylitalo, A.-I. and Vakkala H. (2019). Management Game: Gamifying Leadership Learning, International Journal of Innovation in the Digital Economy, Volume 10, Issue3.

[26] Kesti M. (2019) Architecture of Management Game for Reinforced Deep Learning. In: Arai K., Kapoor S., Bhatia R. (eds) Intelligent Systems and Applications. IntelliSys 2018. Advances in Intelligent Systems and Computing, vol 868. Springer, Cham.