

DYNAMIC LEARNING IN HUMAN DECISION BEHAVIOR FOR EVACUATION SCENARIOS UNDER BDI FRAMEWORK

Seungho Lee
Young-Jun Son

Department of Systems & Industrial Engineering
The University of Arizona
Tucson, AZ 85721-0020, U.S.A.

ABSTRACT

A novel approach to represent learning in human decision behavior for evacuation scenarios is proposed under the context of an extended Belief-Desire-Intention framework. In particular, we focus on how a human adjusts his perception process (involving a Bayesian belief network) in Belief Module dynamically against his performance in predicting the environment as part of his decision planning function. To this end, a Q-learning algorithm (reinforcement learning algorithm) is employed and further developed. In this work, the human decision behavior model is implemented in AnyLogic agent-based simulation software, and the constructed simulation is used to test the impact of the proposed learning approach on emergency evacuation performance, and initial results look quite promising.

1 INTRODUCTION

Learning is one of the most important aspects among the intelligent creatures' behaviors, adopting itself into the environment. Thus, machine learning is also a central function in the artificial intelligence (AI) paradigm which aims to create an intelligent machine. The main stream of the machine learning research has focused on developing techniques that intend to yield the best solution. As a result, the characteristics of the widely used learning algorithms targeted for optimal behaviors have become quite distant from those of real humans, which are not always optimal.

Extensive research has been conducted on applying various machine learning algorithms and models into understanding and mimicking human learning. For example, statisticians have introduced Bayesian models as a way to understand how human deals with uncertainty. Learning Bayesian Belief Network (BBN), a widely studied topic in the field of machine learning, generally implies finding an optimal network structure (structural learning) as well as prior distributions between the connected variables (parametric learning). Although many researchers have developed various methods to construct a BBN model (Heck-

man et al. 1994), a major obstacle for its practical implementation is difficulty in constructing an accurate model, especially when training data is limited. As another attempt for developing a human like learning machine, reinforcement learning (RL) has been adopted initially in the domain of psychology of animal learning that concerns learning by trial and error. Later, in the 1980s, RL has been adopted by the AI field as well (Fu and Anderson 2006). As such, the RL technique was successfully demonstrated to mimic human behavior in simple problem solving situations especially when prior knowledge is limited. Also, while BBN training is an NP-hard problem, training in RL is performed relatively easily based on the recursive mathematical formula. However, a major drawback of RL is its difficulty in being applied to complex problems as states (which can be exhaustive for complex problems) and actions need to be clearly defined beforehand. Thus, if the environmental factors (e.g. states and actions) change, they need to be defined accordingly. In addition, RL is more limited to employ prior knowledge than BBN.

In this work, we propose an innovative learning model for human behavior against a dynamically changing complex environment (a terrorist bombing scenario in a public area is considered in this paper), combining BBN and RL techniques and compensating for the deficiency of each method. To this end, we demonstrate the proposed learning model in the context of the extended Belief-Desire-Intention (BDI) human decision-making framework (Zhao and Son 2008), which was developed by the authors earlier.

2 PROPOSED HYBRID LEARNING MODEL IN THE CONTEXT OF BDI FRAMEWORK

The Learning in this work is defined as the evolutionary process of underlying modules which constitute the human decision behavior process when the considered human evolves from a novice to an expert in a certain aspect. In this work, the extended Belief-Desire-Intention (BDI) framework (Zhao and Son 2008) has been employed for

our analysis as its rich and comprehensive framework provides us with various learning aspects.

2.1 Overview of BDI Framework

BDI is a model of a human’s decision-making process, where mental state is characterized by three major components: beliefs, desires, and intentions (Rao and Georgeff 1998). Beliefs represent information a human has about circumstances. Desires correspond to state of affairs that human would wish to be brought about. Intentions represent desires that a human has committed to achieve. Zhao and Son (2008) extended the original BDI model to include detailed sub-modules. Later, Lee et al. (2008) further extended the model (see Figure 1), appending an Emotion Module containing a confidence index and an instinct index to represent more psychological natures of human. In Figure 1, the perceptual processor perceives the environment and generates beliefs. Beliefs, in turn, are used to create desires through the desire generator. Once the deliberator selects an intention from desires, the real-time planner generates plans which will then be executed by the decision executor. The Emotion Module containing confidence index and instinct index affects and is affected by each component throughout the decision making process.

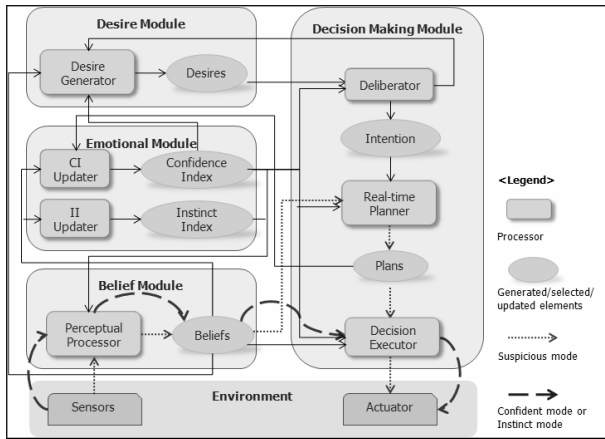


Figure 1: Components of the extended BDI framework

2.2 Proposed Hybrid Learning Model

In this section, we propose a novel hybrid learning algorithm involving BBN and RL for the Belief Module and a Confidence Index (CI) in the Emotion Module of the extended BDI framework.

Bayesian Belief Network for the Belief Module

BBN is a cause and effect, directed acyclic network, where nodes represent considered variables and the direction of arcs encodes the conditional dependencies and cause-effect relationship between the variables. By using BBN, the

probabilistic relationship as well as historical information between variables can be captured via prior and conditional probabilities, which then can be used to infer posterior probabilities given evidence through the Bayes’ theorem. A major advantage of BBN is its ability and flexibility to handle uncertain and dynamic environments. For this reason, we have adopted BBN for the perceptual processor in the Belief module of the BDI framework (Lee et al. 2008), and Bayesian models have become prominent over a broad spectrum of the cognitive science (Griffiths et al. 2008). However, in terms of learning, most of the research works in this area have focused on finding the best model fitting techniques that can accurately represent the given input/output data as opposed to mimicking a dynamic learning process of human (goal of this paper). In this paper, we propose an approach to update the BBN dynamically.

Q-Learning for the Emotion Module

The intent of RL is to obtain an action-value function that gives an expected utility of taking an action in a given state and following a fixed policy thereafter. Q-learning (Watkins 1989) is one of the most actively investigated reinforcement learning techniques. In the extended BDI framework, the Confidence Index (CI) of the Emotion Module affects as well as is affected by all the other modules (e.g. Belief Module, Desire Module, Decision-Making Module). For example, the higher the CI, the longer the planning horizon of the human is in his decision-planning process. Also, the better human’s performance in predicting the environment in his decision-planning process, the higher the CI is. These inter-effects between the CI and other modules evolve as part of the decision maker’s dynamic learning process. In this paper, we investigate such a learning process using the Q-Learning technique (see Equation (1)), in particular for the effect of the CI on the perceptual processor in Belief Module.

Proposed BBN-RL Hybrid Learning Model

As a perceptual processor, BBN takes observed information as inputs and delivers an inferred perception of a decision maker as outputs. The inferred outputs are represented as a set of probability distribution functions $f(x)$ for each of the result factors. Here, as the CI is believed to affect the perceptual processor (inference in BBN), the effect of the CI is considered as an additional step, where the probability distribution for each of the output factors is modified in a way that the probability that positive (optimistic) factors will infer higher values (states) is increased. The positive factor is a relative concept depending on the situation faced by the decision maker. For example, time can be a positive factor when people want to take a rest whereas it can be a negative factor when people travel to a destination. In this work, we propose that the above mentioned effect of the CI on the probability distribution obtained from the BBN is determined and improved (result-

ing in a better decision-making performance) via Q-learning. For example, let us suppose that a node X is a positive output factor (e.g. safety measure under an evacuation situation) in the BBN with three discrete states (*High*, *Medium*, and *Low*). Then, the BBN infers the probability distribution of X (i.e. $p(High)$, $p(Medium)$, $p(Low)$) based on an observation. In the proposed model, the CI changes the probability distribution by subtracting δ from $p(Low)$ and adding it to $p(High)$. The altered amount (δ) is determined based on the current CI value. In our work, the relationship between δ and the CI is trained via the Q-learning algorithm. Equation (1) depicts a general Q-learning algorithm, where $Q(s_t, a_t)$ is a discounted reward, $R(s_t, a_t)$ is an observed immediate reward, s_t and a_t are state and action at time t , α_t ($0 \leq \alpha_t < 1$) is a learning rate, and γ ($0 \leq \gamma < 1$) is a discount factor.

$$Q(s_t, a_t) = (1 - \alpha_t)Q(s_{t-1}, a_{t-1}) + \alpha_t [R(s_t, a_t) + \gamma \cdot \text{Max}(Q(s_{t+1}, a))] \quad (1)$$

In the BDI framework, CI and δ values correspond with the state and action terms in Equation (1), respectively. Once the beliefs are updated via the BBN along with the CI, the CI is updated using the true information that is observed after a while. In this work, we employ the CI ($0 \leq CI_0 \leq 1$) (see Equation (2)) that was suggested by Lee et al. (2008), where d_t (>0) denotes the deviation between what is predicted about the environment during the planning stage and the actual observed environment during the execution stage. In this work, d_t is defined as $d_t = \sum_i |m_i(t-1) - m_i(t)|$ where $m_i(t)$ is the inferred prediction of child node i at time t using BBN.

$$CI_t = \alpha \cdot e^{-d_t} + (1 - \alpha)CI_{t-1} \quad (2)$$

For example, we can use the expected value method as follows. If the child node i has the inferred distribution of each state as $p(High) = 0.3$, $p(Medium) = 0.4$, $p(Low) = 0.3$, then $m_i(t)$ can be calculated as $m_i(t) = 0.3 \times 5 + 0.4 \times 3 + 0.3 \times 1 = 3$. In Equation (2), α ($0 \leq \alpha \leq 1$) adjusts the effect of previous confidence to the current confidence, which varies depending on an individual. The initial confidence value (CI_0) has to be given, which will be different for individuals. The change of the CI can be an immediate reward in Q-learning that is represented as R in Equation (1). In this way, we can find a best (involving the most increasing CI value) δ value for a given CI value.

Illustration of Proposed Q-Learning for Effect of CI

In this example, in order to deal with a finite number of states, the continuous CI value is divided into four discrete intervals: $0 \sim 0.25$, $0.25 \sim 0.5$, $0.5 \sim 0.75$, $0.75 \sim 1$. For example, if the current CI value lies between 0 and 0.25,

the current state is 1; similarly, if the current CI value lies between 0.25 and 0.5, the current state is 2. Here, 9 actions are defined, which will alter the probability distribution of positive factors. Actions I, II, III, and IV subtract 90%, 70%, 50%, and 30% of the inferred probability (δ) from the highest state and add it to the probability of the lowest state, respectively. Action V denotes no alteration. Similarly, Actions VI, VII, VIII, and IX subtract 30%, 50%, 70%, and 90% of the inferred probability (δ) from the lowest state and add it to the probability of the highest state, respectively. We set R (immediate reward) as the amount of the CI value change. In other words, the immediate reward R is defined as $R(s_t, CI_t, a_t(\delta)) = CI_t - CI_{t-1}$. Let us suppose that α and CI_0 are set to $\alpha = 0.5$ and $CI_0 = 0.5$, respectively for the CI (see Equation (2)). Moreover, γ is assumed to be 0.1, which means we tend to neglect future rewards. Then, we can establish a state/action Q matrix as

$$Q = \begin{matrix} & I & \dots & IX \\ \begin{matrix} 1 \\ \vdots \\ 4 \end{matrix} & \begin{bmatrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{bmatrix} & & \end{matrix} \quad \text{Exemplary calculations in}$$

learning are described here. We set $\alpha = 0.5$, $\alpha_t = 0.7$, $\gamma = 0.1$, and $CI_0 = 0.5$. Then, the current state is 2. Suppose that action III is randomly selected at state 2 with a positive factor's $p(High) = 0.67$, $p(Medium) = 0.18$, and $p(Low) = 0.15$ in the BBN. Then, a modified BBN has $p_1(High) = 0.67 - 0.67 \times 0.5 = 0.335$ and $p_1(Low) = 0.15 + 0.67 \times 0.5 = 0.485$ in the positive factor's probability distribution. Then, using the expected method discussed in the previous section, $m(t)$ of this factor is $m(1) = 0.335 \times 5 + 0.18 \times 3 + 0.485 = 2.7$. Suppose further that we obtain $m(2) = 2.74$ from the next BBN inference. Then $d_t = |m(1) - m(2)| = |2.7 - 2.74| = 0.04$. Using Equation (2), we can calculate CI_1 as $CI_1 = \alpha \cdot e^{-d_t} + (1 - \alpha)CI_0 = 0.5 \cdot e^{-0.04} + 0.5 \cdot 0.5 = 0.7$. Thus $R(CI_1, \delta_1) = 0.7 - 0.5 = 0.2$. Then $Q(CI_1, \delta_1) = (1 - \alpha_t) Q(CI_0, \delta_0) + \alpha_t [R(CI_1, \delta_1) + \gamma \cdot \text{Max}(Q(CI_2, \text{all actions}))] = (1 - 0.7) \cdot 0 + 0.7 [0.2 + 0.1 \cdot \text{Max}(Q(0.75, -0.2), Q(0.75, 0), Q(0.75, +0.2))] = 0.14 + 0.1 \cdot 0 = 0.14$.

Then, the Q matrix is updated as following:

$$Q = \begin{bmatrix} 0 & 0 & 0 & \dots \\ 0 & & 0.14 & \\ \vdots & & & \ddots \end{bmatrix} \quad \text{We repeat the same process until}$$

the Q matrix has converged. Suppose that we have repeated the above training and obtained a converged Q matrix as follows:

$$Q = \begin{bmatrix} 22 & 9 & 8 & \dots \\ 17 & 19 & 12 & \\ 9 & 12 & 18 & \\ 3 & 11 & 23 & \dots \end{bmatrix} \quad \text{Via normaliza-}$$

tion, we can obtain a revised Q matrix,

$$Q = \begin{bmatrix} 0.32 & 0.2 & 0.12 & \dots \\ 0.24 & 0.29 & 0.14 & \\ 0.09 & 0.12 & 0.28 & \\ 0.03 & 0.11 & 0.23 & \dots \end{bmatrix} \quad \text{which is then used for the}$$

operation. For example, the first row of the Q matrix defines the probability distribution of actions in state 1.

Thus, if we are in state 1, the probabilities of selecting actions I, II, and III are 0.32, 0.2, and 0.12, respectively. It is noted that the summation of the elements in each row of the normalized Q matrix is 1.

3 EXPERIMENT UNDER EMERGENCY EVACUATION SCENARIO

In this section, the proposed hybrid learning model is illustrated in a simulated environment for emergency evacuation (bombing attack) scenarios. In particular, evacuation performances (e.g. CI) between learned agents and novice agents are compared, where learned agents update their Q matrix under various emergency scenarios. Also, the effects of various parameters considered in the proposed hybrid model on learning are discussed.

3.1 Simulation Model of Emergency Evacuation

In this work, the proposed hybrid learning model is tested and illustrated using agent-based simulation developed for emergency evacuation scenarios (Lee et al. 2008). In the simulation, we observe the crowd behaviors under a terrorist bomb attack in the Washington D.C. National Mall area. In our simulation, three types of agents are considered, including 1) commuter, 2) novice, and 3) police agent, whose defined behaviors are different. The number of each agent type in the simulation can be adjusted. Figure 2 depicts a snapshot of the AnyLogic simulation, where bomb explosion is shown in the middle of the map and agents are evacuating from the area. The constructed simulation allowed us to observe agents' behaviors that mimic human in the given scenario, using which we were able to evaluate various evacuation policies (Lee et al. 2008). Figure 3 depicts a BBN that is used by the agents to perceive the environmental information and convert them into their own belief. As shown in Figure 3, the agents consider fire, smoke, police, crowd, and distance to exit as the environmental information. Then, via the BBN, the information is translated into the agent's belief about the risk and evacuation time for each of the alternative paths.



Figure 2: Emergency Evacuation simulation in AnyLogic

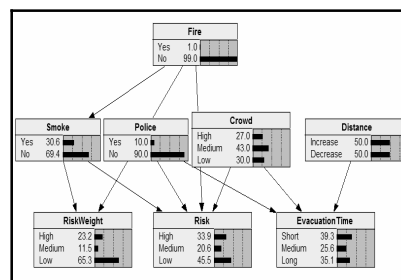


Figure 3: BBN used for perceptual processor of BDI agent

3.2 Experimental Results

In this section, we discuss preliminary simulation results obtained from different sets of parameters considered in the proposed hybrid learning model such as α (see Equation (2)), α_t , and γ (see Equation (1)). In particular, we compare the results obtained from the BBN-RL hybrid method with those from the BBN method only. In order to construct a Q matrix, we first create 500 instances of normal agent and one special type of agent that updates the Q matrix. Every agent adjusts its CI according to Equation (2), and only the learning agent updates the Q matrix using Equation (1) and the algorithm in Figure 2. Since an agent's speed of movement and his planning horizon depend on the CI value, the evacuation performances (e.g. average evacuation time, finding the best evacuation path) are closely related with the CI values. Thus in this paper, we measured the CI as the performance index. To obtain a converged Q matrix, we made the learning agent to update the Q matrix under various situations of emergency evacuation during 200 replications of the simulation. Then we have normalized the matrix so that values in each row (state) are summed to 1. Table 1 depicts the normalized Q matrix using $\alpha = 0.5$, $\alpha_t = 0.7$, and $\gamma = 0.5$. In this case ($\alpha = 0.5$, $\alpha_t = 0.7$, and $\gamma = 0.5$), action V ('Do nothing' action) has the lowest value, which means action V returns the least reward (increment of CI). We have repeated this Q matrix training process using different parameter α , α_t , and γ values varying 0.3 to 0.9 with increment of 0.2.

Table 1: The normalized state (1 ~ 4) and action (I ~ IX) Q matrix using $\alpha = 0.5$, $\alpha_t = 0.7$, and $\gamma = 0.5$

	I	II	III	IV	V	VI	VII	VIII	IX
1	0.16	0.11	0.06	0.04	0.00	0.05	0.09	0.18	0.31
2	0.16	0.13	0.05	0.02	0.01	0.05	0.11	0.14	0.32
3	0.16	0.11	0.06	0.04	0.02	0.05	0.09	0.20	0.27
4	0.17	0.12	0.05	0.04	0.03	0.06	0.11	0.17	0.26

Once we obtain the Q matrix, we ran the model with letting the agent to use each Q matrix for the BBN adjustment. Each simulation with a different Q matrix is replicated 100 times. CI_0 is set to 0.5 and CI_t is updated ac-

ording to Equation (2) thereafter. In addition to the effect of each of the parameters as mentioned earlier, we compared the CI results from two different action/selection policies which are greedy and softmax policies. The greedy policy is a special case of ϵ -greedy policy where $\epsilon = 0$. Thus the agent selects the action that has the biggest utility value at each state. In the other hand, the softmax policy is selecting the action a_i on state s_j randomly based on the probability $P(s_j, a_i) = \frac{Q(s_j, a_i)}{\sum_{a_i} Q(s_j, a_i)}$ where $Q(s_j, a_i)$ is the element in j^{th} row and i^{th} column of the Q matrix.

Figure 4 depicts the CI values over simulation time t for different α when we did not adjust BBN using the Q-Learning algorithm. α decides the reflection of wrong prediction into CI. Thus as α decreases the variance of CI decreases and CI value itself decreases slowly.

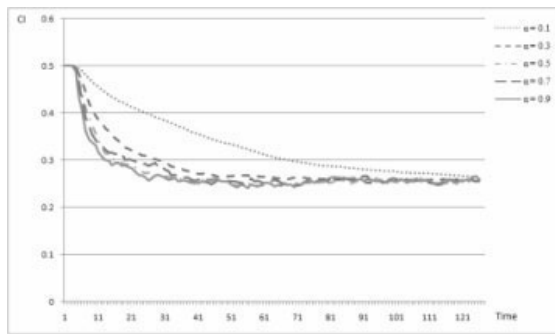


Figure 4: CI over time without applying Q learning for each α

4 CONCLUSIONS

We have proposed a promising hybrid learning model integrating BBN and RL techniques. The proposed model has been implemented in emergency evacuation agent-based simulation. The developed simulation allowed us to observe the effect of learning under various conditions. The simulation results demonstrated that the proposed model effectively adjusted itself to an inexperienced situation without any prior knowledge. The inexperienced agent inferring environment based only on BBN progressed to an expert by adjusting the inferred perception via RL.

ACKNOWLEDGMENTS

This work was supported by Air Force Office of Scientific Research under AFOSR/MURI F49620-03-1-0377.

REFERENCES

Fu, W. and J.R. Anderson. 2006. From Recurrent Choice to Skill Learning: A Reinforcement-Learning Model.

Journal of Experimental Psychology: General 135(2): 184-206.

Griffiths, T. L., Kemp, C., and Tenenbaum, J. B., 2008, "Bayesian models of cognition," In Ron Sun (ed.), *Cambridge Handbook of Computational Cognitive Modeling*. Cambridge University Press.

Heckerman, D., D. Geiger, and D. Chickering. 1994. Learning Bayesian networks: The combination of knowledge and statistical data. *In Proceedings of 10th Conference on Uncertainty in AI*.

Lee, S., Y. Son, and J. Jin. 2008. Integrated Human Decision Making and Planning Model under Extended BDI Framework. *ACM Transactions on Modeling and Computer Simulation* (submitted).

Rao, A.S. and M.P. Georgeff. 1998. Decision procedures for BDI logics. *Journal of Logic and Computation* 293-342.

Watkins, C. J. C. H., 1989, "Learning from delayed rewards," Ph. D. thesis, Cambridge University.

Zhao, X. and Y. Son. 2008. BDI-based Human Decision-Making Model in Automated Manufacturing Systems. *International Journal of Modeling and Simulation* 28(3):347-356.

AUTHOR BIOGRAPHIES

SEUNGHO LEE is a Ph.D. student in the Department of Systems and Industrial Engineering at the University of Arizona. He received his Bachelor of Engineering degree in Industrial Engineering from Korea University in Korea in 1999 and his M.S. degree in Industrial Engineering from Texas A&M University in 2005. His research focuses on application of distributed simulation and simulation of human decision making. He is a student member of IIE and INFORMS. He can be reached by email at <mountlee@email.arizona.edu>.

YOUNG-JUN SON is an associate professor in the Department of Systems and Industrial Engineering and Director of Center for Advanced Integration of Manufacturing Systems and Technologies at The University of Arizona. He is an associate editor of the *International Journal of Modeling and Simulation* and the *International Journal of Simulation and Process Modeling*. He has received several research awards such as the SME 2004 Outstanding Young Manufacturing Engineer Award, the IIE 2005 Outstanding Young Industrial Engineer Award, the Industrial Engineering Research Conference Best Paper Award (in 2005 and 2008), and the Best Paper of the Year Award (2007) in *International Journal of Industrial Engineering*. His research focuses primarily on distributed and hybrid simulation for the analysis and control of large-scale dynamic systems such as automated manufacturing system, integrated enterprise, power grid, and homeland security. He can be reached by email at <son@sie.arizona.edu>.