2023 IEEE International Conference on Development and Learning (ICDL)
November 9-11, Macau, China

Safe Human-Machine Cooperative Game with Level-k Rationality Modeled Human Impact

Junkai Tan, Shuangsi Xue, Hui Cao, *Member, IEEE* and Huan Li

***Abstract*— This paper considers the problem of bounded rational human behavior in the cooperative human-machine game. The cooperation between human and machine is a raising topic for emergency handling, and it is critical to ensure the safety of human. First, a barrier-function-based state transformation is developed to ensure the safety constraints of the human-machine system state. A level-k rationality structure is then exploited by cognitive hierarchy to learn human behavior, and the bounded rational behavior is obtained by using Adaptive Dynamic Programming (ADP). Inspired by behavior modeling from sociology, a softmax probabilistic decision distribution is utilized to model human behavior, which imitates the true impact of human in the cooperative game. Finally, a simulation is implemented to test the effectiveness of the proposed behavior, which demonstrates that the full state constraints and stabilization are guaranteed.**

2023 IEEE International Conference on Development and Learning (ICDL) | 978-1-6654-7075-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICDL55364.2023.10364413

I. Introduction

Human-machine fusion decision is an raising topic in emergency handling [1]–[3]. The sudden occurrence of emergencies in safety-critical systems poses challenges in acquiring sufficient information regarding emergent situations for human decision-making. The machine is able to gather enough information in a relatively short period to effectively manage the crisis. The collaboration of human and machine is essential and significant in safety-critical system. In recent years, game-based system have gained significant interest due to their extensive usage in various field, including economics, robotics, automated driving and cyper-physical-systems [4]–[7]. The objective of collaboration between humans and machines is to attain common benefits. Nevertheless, specific constraints must be complied with to guarantee the safety of the human player.

Safe reinforcement learning is a method involving the interaction between agents and their environment to learn the optimal controller. This approach includes designed mechanisms to guarantee that specific safety constraints are satisfied. The authors of [8], [9] introduce a novel structure that combines actor-critic-identifier to identify system dynamics, resulting in enhanced performance in danger detection. A safe-guaranteed controller is proposed to avoid the state trajectory from causing collisions with non-convex boundary limits in [10]. In [11]–[13], a barrier-function-based transformation is proposed, which converts safety issue to stabilization problem to meet the full state constraints of a rectangular. The authors in [14]–[16] integrate the barrier function and reward function to penalize the behavior of reaching the boundary.

The theory of cognitive hierarchy has gained prominence in recent years. An novel ADP approach of solving the non-equilibrium game is developed to achieve the stabilization without acquiring system dynamic [17]–[19]. The work of [20] propose a cognition modeling game architecture which incorporates unmanned aircraft with the Airspace System. The bounded level reasoning structure is proposed to predict the decision-making process of human-beings, which is constrained by the limited rationality of their beliefs [21]. The cooperation game of human and self-driving vehicles is investigated to achieve collaborative human-vehicle decision-making in [2], [22], [23]. The authors of [17], [24], [25] employed the 'softmax' function to replicate the stochastic distribution of various levels of human behavior, based on the concept of bounded rationality.

The contributions of this paper are threefold: 1) the safe human-machine cooperative game is formulated using barrier-function-based state transformation, which guarantees the full state constraints of the safety-critical system involved human. 2) A level-k rationality architecture is developed base on the theory of Cognitive Hierarchy. the bounded rational behavior is obtained via online learning method of adaptive dynamic programming. 3) Softmax probabilistic distribution model is utilized to simulate the true bounded rationality of human behavior.

This paper is organized as follows. Section II illustrates the basic setup for the cooperative human-machine game. Section III outlines the barrier-function-based transformation system. Section IV developed a bounded rationality level-k architecture via theory of cognitive hierarchy. Section V presented a model-based reinforcement learning structure to implement the online approximation. Section VI modeled the human player impact by a probabilistic approach. Section VII verifies the effectiveness of the proposed method. Section VIII concludes the work of this paper.

II. Preliminaries and Problem Formulation

To investigate the human-machine cooperative game, we consider the continuous-time nonlinear dynamical system with affine input $\forall t \geq 0$,

$$\dot{x} = f(x) + g_h(x)u_h + g_m(x)u_m, \tag{1}$$

where $x = [x_1 \cdots x_n]^{\mathrm{T}} \in \mathbb{R}^n$ is the system state, $u_i \in \mathbb{R}^{o_i}$, for $i = h, m$ is the control input of human and machine

Junkai Tan, Shuangsi Xue, Hui Cao and Huan Li are with the Shaanxi Key Laboratory of Smart Grid, Xi'an Jiaotong University, Xi'an 710049, China, and also with the State Key Laboratory of Electrical Insulation and Power Equipment, School of Electrical Engineering, Xi'an Jiaotong University, Xi'an 710049, China. (e-mails: 15958024@stu.xjtu.edu.cn; xssxjtu@stu.xjtu.edu.cn; huicao@mail.xjtu.edu.cn; lh2000dami@stu.xjtu.edu.cn)(*Corresponding: Shuangsi Xue*)

respectively, $f(x) : \mathbb{R}^n \to \mathbb{R}^n$ represent the nonlinear dynamic of the cooperative system. $g_h(x) : \mathbb{R}^n \to \mathbb{R}^{o_h}$ and $g_m(x) : \mathbb{R}^n \to \mathbb{R}^{o_m}$ represent the input matrix gain of human and machine respectively. Functions $f(x)$, $g_h(x)$ and $g_m(x)$ are Lipschitz and continuous.

Due to the cooperation of human-machine system, we consider the performance index of human and machine is exactly the same, which is defined as,

$$\mathcal{J}_{coop}\left(x_0; u_h, u_m\right) = \frac{1}{2}\int_0^\infty r_{coop}\left(x, u_h, u_m\right) d\tau, \quad (2)$$

where $r_{coop}\left(x, u_h, u_m\right) = M(x) + \sum_{j\in\{h,m\}} u_j^T R_{jj} u_j$ is the fully cooperative reward function of human-machine system, $M(x)$ is a quadratic function of state x.

Definition 1. Input pair $(u_h^\star, u_m^\star)$ is the Nash equilibrium, if satisfies, $\mathcal{J}_{coop}\left(x_0; u_h^\star, u_m^\star\right) \le \mathcal{J}_{coop}\left(x_0; u_h, u_m^\star\right), \forall u_h$ and $\mathcal{J}_{coop}\left(x_0; u_h^\star, u_m^\star\right) \le \mathcal{J}_{coop}\left(x_0; u_h^\star, u_m\right), \forall u_m$.

The best response of the fully cooperative human-machine system is the Nash equilibrium, however, the system state involving human beings should always satisfy the safe constraints. To ensure safety, a barrier-function-based transformation system is given in the following subsection.

III. Barrier-Function-based State Transformation Safe-Critical System

A. Barrier Function Transformation

In this subsection, we first consider the problem of state constraints. To simplify the notation of safety limits, the polygonal state constraints set is given as $x \in \mathcal{O}$, where $\mathcal{O} = \{x \in \mathbb{R}^n | a \le Cx + p \le A\}$, $a = [a_1, ..., a_l]^T \in \mathbb{R}^l$, $A = [A_1, ..., A_l]^T \in \mathbb{R}^l$, $p = [p_1, ..., p_l]^T \in \mathbb{R}^l$ and $C \in \mathbb{R}^{l\times n}$. The problem of the safety-critical game can be formulated as follows.

Problem 1. Consider the nonlinear system (1), and given the cooperative performance index (2), find the Nash equilibrium policies $(u_h^\star, u_m^\star)$, with satisfying $x \in \mathcal{O}$.

To simplify the analysis procedure without loss of the generality, we choose the state constraint in the form of $x_k \in (d_k, D_k),\ k = 1, \ldots, n$. The lower constraint and upper constraint satisfy $d_k < 0 < D_k$ and $\|d_k\| \ne \|D_k\|$, which means the state constraints is asymmetric.

To address the state constraint issue, the transformation of the system state using the barrier function is introduced. We transform the safety problem with constraint $x \in \mathcal{O}$ into a stabilization problem.

Based on the constraint $x_k \in (d_k, D_k),\ k = 1, \ldots, n$, we select the barrier function in the form of

$$b(x_k; d_k, D_k) = \log\left(\frac{D_k}{d_k}\frac{d_k - x_k}{D_k - x_k}\right).$$

The inverse function of the barrier function $b(x_k; d_k, D_k)$ on the interval (d_k, D_k) is

$$b^{-1}(y_k; d_k, D_k) = \frac{D_k d_k (e^{\frac{y_k}{2}} - e^{-\frac{y_k}{2}})}{d_k e^{\frac{y_k}{2}} - D_k e^{-\frac{y_k}{2}}}.$$

With the barrier function $b(\cdot)$ and the state $x \in \mathbb{R}^n$ of system (1), the system state transformation via barrier function can be summarized as

$$\begin{aligned} s_k &= b\left(x_k; d_k, D_k\right) = b_k, \\ x_k &= b^{-1}\left(s_k; d_k, D_k\right) = b_k^{-1},\ \forall k = 1, \ldots, n. \end{aligned} \quad (3)$$

By utilizing the chain rule, we obtain the derivative of transformed state s_k with respect to time as

$$\frac{ds_k}{dt} = \left(\frac{dx_k}{ds_k}\right)^{-1}\frac{dx_k}{dt}.$$

The dynamics of transformed state $s = [s_1, ..., s_n]^{\mathrm{T}}$ can be expressed as

$$\dot{s}_k = \frac{\dot{x}_k}{\frac{db^{-1}(s_k; d_k, D_k)}{dy}} = F_k(s) + G_k^h(s)u_h + G_k^m(s)u_m, \quad (4)$$

where $F_k(s_k) = \tau(s_k) \times f\left(\left[b_1^{-1}, ..., b_n^{-1}\right]^{\mathrm{T}}\right)$, $G_k^h(s) = \tau(s_k) \times g_h\left(\left[b_1^{-1}, ..., b_n^{-1}\right]^{\mathrm{T}}\right)$ and $G_k^m(s) = \tau(s_k) \times g_m\left(\left[b_1^{-1}, ..., b_n^{-1}\right]^{\mathrm{T}}\right)$ with $\tau(s_k) = \left(\frac{db^{-1}(y_k; d_k, D_k)}{dy_k}\right)^{-1}$.

Then the transformed system dynamics (4) could be written in the following compact form of

$$\dot{s} = F(s) + G^h(s)u_h + G^m(s)u_m, \quad (5)$$

where $F(s) : \mathbb{R}^n \to \mathbb{R}^n$ is the nonlinear dynamic of transformed system. $G^h(s) : \mathbb{R}^n \to \mathbb{R}^{o_h}$ and $G^m(s) : \mathbb{R}^n \to \mathbb{R}^{o_m}$ are the transformed input gain matrix.

B. Nash Equilibrium in Transformed System

Based on the transformed system dynamics of state s, the Nash equilibrium for (5) shall be obtained to achieve optimal control.

The goal of the previous human-machine cooperative game converts to stabilize the transformed system (5), with minimum resource consumption. The minimization problem can be solved by minimizing the value function as

$$V_{coop}\left(s, u_h, u_m\right) = \int_t^\infty r_{coop}\left(s, u_h, u_m\right) d\tau. \quad (6)$$

Definition 2. Consider system (5), a pair of policies $u_{h,m} = \{u_h, u_m\}$ is admissible control pair, if $u_{h,m}$ stabilizes the transformed system (5), and value function V from (6) is finite.

Thus, for the optimality of controlling transformed system (5), an admissible pair of policies $u_{h,m}^* = \{u_h^*, u_m^*\}$ is the Nash equilibrium, which obtained the optimal value function in the form of

$$V_{coop}^*\left(s, u_h, u_m\right) = \min_{u_h, u_m}\int_t^\infty r_{coop}\left(s, u_h, u_m\right) d\tau. \quad (7)$$

Define the Hamiltonian function for the transformed human-machine cooperative system as

$$\begin{aligned} \mathcal{H}\left(s, \nabla V, u_h, u_m\right) \triangleq{}& \left(\nabla V\right)^{\mathrm{T}}\left[F(s) + G^h(s)u_h + G^m(s)u_m\right] \\ &+ r_{coop}\left(s, u_h, u_m\right), \end{aligned} \quad (8)$$

where $\nabla V = \frac{\partial V_{coop}}{\partial s}$ is the gradient of the value function.

By differentiating the Hamiltonian function and applying the stationary conditions, in the form of $\frac{\partial \mathcal{H}_i}{\partial r_i} = 0$, we can obtain the optimal controller pair

$$u_i^\star(s) = -\frac{1}{2} R_i^{-1} \left(G^i(s)\right)^{\mathrm{T}} \nabla V^\star,\ i = h, m. \tag{9}$$

Substituting the optimal controller (9) into the Hamiltonian (8) yields the Hamilton-Jacobi-Isaacs(HJI) equation in the form of

$$\begin{aligned} 0 = (\nabla V^*)^{\mathrm{T}} \left(F(s) - \frac{1}{2} \sum_{j \in \{h,m\}} G^j(s) R_j^{-1} (G^j(s))^{\mathrm{T}} \nabla V^* \right) \\ + Q(s) + \frac{1}{4} \sum_{j \in \{h,m\}} (\nabla V^*)^{\mathrm{T}} G^j(s) R_j^{-1} (G^j(s))^{\mathrm{T}} \nabla V^*. \end{aligned} \tag{10}$$

According to Lemma 1 from [12], given the transformed system (5), we can solve the full-state constraints by finding a pair of Nash Equilibrium policies $u_{h,m} = \{u_h, u_m\}$. In the next section, the cognitive hierarchy will be introduced to obtain a level-k bounded rationality.

IV. Level-k Bounded Rationality

Cognitive hierarchy theory shows that humans think strategically, which means they develop beliefs by predicting which level of rationality others would perform, and subsequently select optimal responses based on those beliefs. In this section, we introduce a level-k bounded rationality structure to obtain different levels of intelligence.

A. Initial Policy (level-0) of Human

For the cooperative human-machine system, level-0 rationality represents an instinctive reaction, which means the behaviors of players are non-cooperative. To prevent the potential stochastic danger, we will obtain human level-0 rationality by solving an optimization problem, which is in the form of minimizing a specific value function

$$V_{u_h}^0(s_0) = \min_{u_h^0} \int_0^\infty \left(M(s) + (u_h^0)^{\mathrm{T}} R_h u_h^0 \right) \mathrm{d}\tau, \tag{11}$$

which is subject to the dynamic of the system $\dot{s} = F(s) + G^h(s) u_h^0$. According to the optimal control theory, the stationary condition for the optimization problem (11) is

$$u_h^0(s) = -\frac{1}{2} R_h^{-1} (G^h(s))^{\mathrm{T}} \nabla V_h^0, \tag{12}$$

where $\nabla V_h^0 = \frac{\partial V_{u_h}^0(s)}{\partial s}$, the value function $V_{u_h}^0$ is known to satisfy the Hamilton-Jacobi-Bellman(HJB) equation, namely

$$\begin{aligned} \mathcal{H}\left(s, \nabla V_h^0, u_h^0\right) = \left(\nabla V_h^0\right)^{\mathrm{T}} \left[F(s) + G^h(s) u_h^0\right] \\ + r_{coop}\left(s, u_h^0, 0\right) = 0. \end{aligned} \tag{13}$$

B. Level-1 Policy of Machine

Assuming the human player always acts the level-0 policy, the level-1 policy of the machine could be solved subsequently, which is the optimal response to the initial level-0 policy from the human.

To acquire the level-1 policy of the machine, an optimization problem is established as follows

$$\begin{aligned} V_{u_m}^1(s_0) = \min_{u_m^1} \int_0^\infty \big(M(s) + (u_m^1)^{\mathrm{T}} R_m u_m^1 \\ + (u_h^0)^{\mathrm{T}} R_h u_h^0 \mathrm{d}\tau, \end{aligned} \tag{14}$$

which is subject to the dynamic of the system $\dot{s} = F(s) + G^h(s) u_h^0 + G^m(s) u_m^1$. The optimal policy for the optimization problem (14) is

$$u_m^1(s) = -\frac{1}{2} R_m^{-1} (G^m(s))^{\mathrm{T}} \nabla V_m^1, \tag{15}$$

where $\nabla V_m^1 = \frac{\partial V_{u_m}^1(s)}{\partial s}$, the value function $V_{u_m}^1$ is known to satisfy the HJI equation, namely $\mathcal{H}\left(s, \nabla V_d^1, u_h^0, u_m^1\right) = 0$.

C. Level-k and Level-$(k+1)$ Policies

An iterative procedure will be employed to formulate higher-level rational policies for the human and machine respectively. This procedure involves iterative optimizations of policies by the human and machine, with a belief that their partner employs lower-level rationality.

Interacting with the machine which employs level-$(k-1)$ rationality, the human player acquires level-k thinking, by solving the following optimization problem

$$\begin{aligned} V_{u_h}^k(s_0) = \min_{u_h^k} \int_0^\infty \big(M(s) + (u_h^k)^{\mathrm{T}} R_h u_h^k \\ + (u_m^{k-1})^{\mathrm{T}} R_m u_m^{k-1} \mathrm{d}\tau, \end{aligned} \tag{16}$$

which is subject to the dynamic $\dot{s} = F(s) + G^h(s) u_h^k + G^m(s) u_m^{k-1}$. The corresponding HJI equation is $\mathcal{H}\left(s, \nabla V_h^k, u_h^k, u_m^{k-1}\right) = 0$.

The stationary condition leads to the formulation of the level-k policy of human

$$u_h^k(s) = -\frac{1}{2} R_h^{-1} (G^h(s))^{\mathrm{T}} \nabla V_h^k. \tag{17}$$

Similarly, the level-$(k+1)$ rationality could be obtained by solving the subsequent minimization problem

$$\begin{aligned} V_{u_m}^{k+1}(s_0) = \min_{u_m^{k+1}} \int_0^\infty \big(M(s) + (u_m^{k+1})^{\mathrm{T}} R_m u_m^{k+1} \\ + (u_h^k)^{\mathrm{T}} R_h u_h^k \mathrm{d}\tau, \end{aligned} \tag{18}$$

which is subject to the dynamic $\dot{s} = F(s) + G^h(s) u_h^k + G^m(s) u_m^{k+1}$. The optimal policy for the optimization problem (18) is

$$u_m^{k+1}(s) = -\frac{1}{2} R_m^{-1} (G^m(s))^{\mathrm{T}} \nabla V_m^{k+1}. \tag{19}$$

The HJI equation satisifies $\mathcal{H}\left(s, \nabla V_d^{k+1}, u_h^k, u_m^{k+1}\right) = 0$.

Lemma 1. [19] Consider the transformed system (5), given the human and machine player bounded level-k and level-$(k+1)$ rationality respectively, the corresponding value functions are positive

definite. If the following conditions hold

$$\begin{aligned} u_h^k(0) = 0,\ u_m^{k+1}(0) =& 0, \\ \dot{V}_{u_h}^k(\mathrm{s}) < 0, \dot{V}_{u_m}^{k+1}(\mathrm{s}) <& 0, \forall \mathrm{s} \neq 0, \\ \mathcal{H}\left(s, \nabla V_h^k, u_h^k, u_m^{k-1}\right) =& 0, \forall \mathrm{s}, \\ \mathcal{H}\left(s, \nabla V_d^{k+1}, u_h^k, u_m^{k+1}\right) =& 0, \forall \mathrm{s}, \\ \mathcal{H}\left(s, \nabla V_h^k, u_h, u_m^{k-1}\right) \geqslant& 0, \forall \mathrm{s}, u_h, \\ \mathcal{H}\left(s, \nabla V_d^{k+1}, u_h^k, u_m\right) \leqslant& 0, \forall \mathrm{s}, u_m. \end{aligned} \tag{20}$$

Then, the pair of bounded rational policies $u_{h,m} = \{u_h^k, u_m^{k+1}\}$ is the Nash equilibrium.

In the next section, the adaptive dynamic programming(ADP) method would be used to approximate bounded rationality online.

V. Online Learning via Adaptive Dynamic Programming

In this section, two critic networks are utilized to obtain the value functions $V_i,\ i \in \{h, m\}$ of the human and machine respectively. The corresponding policies of the human and machine are denoted as u_i^j for simplification. Up to level-k, we select the single-layer network to approximate the value function as

$$V_i^j = (W_i^j)^T \phi^j(s) + \epsilon_i^j(s), \tag{21}$$

where $W_i \in \mathbb{R}^{p_i}$ represent the ideal neuron weight of the single-layer network and $\phi(x) \in \mathbb{R}^{n \times p_i}$ is the corresponded activation function, p_i is the number of hidden layer neuron and $\epsilon_i(x)$ is the approximation error of single-layer network. The gradient for the value function is

$$\nabla V_i^j = (\nabla \phi^j(x))^T W_i^j + (\nabla \epsilon_i^j)^T(s). \tag{22}$$

The estimated value function $\hat{V}_i$ is expressed as

$$\hat{V}_i^j = (\hat{W}_i^j)^T \phi^j(x), \tag{23}$$

where $\hat{W}_i^j \in \mathbb{R}^{p_i}$ is the estimated weight of the single network.

To reduce the computational load, the general policy of the human and machine, u_i^j, is approximated by a neural network respectively as

$$u_i^j = -\frac{1}{2} R_i^{-1} (G^i(s)^T ((\nabla \phi_i^j(s))^T W_i^j + (\nabla \epsilon_i^j(s))^T). \tag{24}$$

With the estimated value's gradient using the weights W_i in (22), the actual controller can be expressed in the form of

$$\hat{u}_i^j = -\frac{1}{2} R_i^{-1} G^i(s)^T (\nabla \phi_i^j(s))^T \hat{W}_i^j, \tag{25}$$

Based on the estimation of the value function (23), and policies (25), the approximation error of the HJB equation could be defined as

$$\begin{aligned} \mathcal{H}\left(s, \nabla \phi_i^j, u_i^j\right) = & \sum_{l=h,m} (u_l^j)^T R_i u_l + M(s) \\ & + [(W_i^j)^T \nabla \phi_i^j + (\nabla \epsilon_i^j)^T](F + \sum_{l=h,m} G^l u_l^j). \end{aligned} \tag{26}$$

For the simplification of notation, denote that $\mathcal{H}\left(s, \nabla \phi_i^j, u_i^j\right) = e_{H,i}^j$, $\mathcal{H}\left(s, \nabla \phi_i^j, \hat{u}_i^j\right) = e_i^j$ and $\omega_i^j = \nabla \phi_i (F + G^h u_h^j + G^m u_m^j)$.

To obtain the approximated admissible policy $u_{h,m}^j$, an optimization procedure is established. To facilitate the optimization, we construct the energy-like objective E_i by combining the historical and instantaneous data, which is could be defined as follows

$$E_i^j = \frac{1}{2} \sum_{k=0}^{M} \frac{(e_{i,k}^j)^2}{\left(1 + (\omega_{i,k}^j)^T \omega_{i,k}^j\right)^2}, \tag{27}$$

where $\omega_{i,k}^j, k = 1, ..., M$ is the historical data of ω_i^j, $\omega_{i,0}^j$ is the current record of ω_i^j. M is the length of the historical stack. Define $\overline{\omega}_i^j = \left[\omega_{i,1}^j ... \omega_{i,M}^j\right]$ as the historical data stack.

Based on the property of the objective function E_i^j, by utilizing the least-square method, the adaptive learning law for the estimated critic network weight $\hat{W}_i$ can be derived as

$$\dot{\hat{W}}_i^j = -a_i^j \frac{\partial E_i^j}{\partial \hat{W}_i^j} = -a_i^j \sum_{k=0}^{M} \frac{\omega_{i,k}^j e_{i,k}^j}{\left(1 + (\omega_{i,k}^j)^T \omega_{i,k}^j\right)^2}, \tag{28}$$

where the learning rate of each bounded rational level, denoted as a_i^j, plays a crucial role in determining the convergence speed of network weights W_i.

In order to examine the stability of the proposed regular ADP controller, the Lyapunov stability analysis is presented. First, the error dynamic of $\tilde{W}_i^j$ is given as:

$$\dot{\tilde{W}}_i^j(t) = -a_i^j \sum_{k=0}^{M} \frac{\omega_{i,k}^j}{(\omega_{i,k}^j)^T \omega_{i,k}^j + 1} \left[\frac{(\omega_{i,k}^j)^T \tilde{W}^j + e_{H,i}^{k,j}}{(\omega_{i,k}^j)^T \omega_{i,k}^j + 1}\right]. \tag{29}$$

Lemma 2. [12] Critic Weights W_i^j is uniformly ultimately bounded (UUB) under the following assumption of 1. $rank(\overline{\omega}_i^j) = p_i$; 2. $e_{H,i}^j$ is upper bounded by $e_{Hmax,i}^j$

VI. Human Impact Modeling and Interacting with Machine

Within this section, an algorithmic framework will be proposed, in which a machine cooperates with a human possessing varying levels of cognitive ability to pursue the same goal. In order to adopt our result in the real human-machine scenario, the fixed-level policy throughout the course of the interaction should be refrained, which imposes limitations on humans with regard to their utilization of rationality and ignores the variability of human behavior.

As mentioned in the previous section, the machine player calculates the level-k rational policies through the implementation of an ADP algorithm that cooperates with human policies. As a result, rather than using a precise level of human behavior, a probabilistic distribution of human policies is utilized to model the stochastic and dynamic effects resulting from human behavioral decision-making.

Assuming that the machine is capable of precisely measuring human policies' impact on the system. The error of optimism is defined as the difference between the measured human behavior denoted $\mathcal{U}_h(\tau)$, and the human policy of level-k.

$$r^k = \int_{T_{\text{int}}} \left\| \mathcal{U}_h(\tau) + \frac{1}{2} R_h^{-1} (G^h)^T (\nabla \phi_h^j)^T \hat{W}_h^j \right\| \mathrm{d}\tau, \tag{30}$$

where $j \in \{1, \ldots, k_m\}$, k_m is the maximum level of human rationality been computed. Remark that (30) is the norm of the measured human policies' distance from each cognitive level computed by the given ADP algorithm.

Consider the machine player performing at the optimal response, namely the Nash equilibrium $\mathcal{U}_h(t) = -\frac{1}{2} R_h^{-1} (G^h)^T (\nabla \phi_h)^T W_h^*,\ \forall t \geq 0$.

According to Theorem 1, it is achievable to train any given level-k to attain convergence with the optimal response strategy of a human, which in the form of $u_h^j(t) = -\frac{1}{2} R_h^{-1} (G^h)^T (\nabla \phi_h^j)^T \hat{W}_h^j$. Moreover, given Lemma 1 holds, the level-k rationality will approach infinity and ultimately converge to the Nash solution. This implies that the Nash solution represents the limit of the level-k cognitive hierarchy, i.e., $\lim_{j \to +\infty} \left\| V_h^j - V_h^\star \right\| = 0$, it provides

$$\begin{aligned} \lim_{j \to +\infty} u_h^j(t) &= \lim_{j \to +\infty} \left(-\frac{1}{2} R_h^{-1} (G^h)^T (\nabla \phi_h^j)^T \hat{W}_h^j \right) \\ &= -\frac{1}{2} R_h^{-1} (G^h)^T (\nabla \phi_h)^T \hat{W}_h^* = \mathcal{U}_h(t). \end{aligned} \tag{31}$$

Consequently $\lim_{k\to+\infty} r^k = 0$, the following probabilistic distribution model would be established based on the error r^k.

During each interval of interaction T_{int}, the error r^k will be organized in a vector $\mathbf{r}$ of the form $\mathbf{r} = \left[r^1, r^2, \ldots, r^{k_m}\right]$. To formulate the bounded rational human player's behavior, the softmax function is utilized to transform the error vector $\mathbf{r}$ into a bounded value vector, i.e., $\sigma = \left[\sigma^1, \sigma^2, \ldots, \sigma^{k_m}\right]^{\mathrm{T}}$, the basic element σ^k could be expressed as

$$\sigma^k = \frac{e^{-r^k}}{\sum_{i=1}^{k_m} e^{-r^i}}. \tag{32}$$

Then, the policy of the human player could be represented by the proposed probabilistic distribution as follows

$$\mathcal{P}(\mathcal{U}_h = \hat{u}_h^k) = \sigma^k. \tag{33}$$

It is notable that the choice map known as "softmax" is a routine selection for the purpose of modeling human decision-making [24]. Minor errors are indicative of greater chances to choose the suitable bounded rational behavior.

For the purpose of determining the safe and stabilizing policies of level-k human-machine cooperation, an online ADP algorithm is formulated in the following Algorithm 1.

Algorithm 1: The ADP algorithm for the Human-Machine Cooperative game

Given the initial state x_0, gain matrix R_i, learning rate a_i^j.
for k = 0,...,k_m **do**
  for i=h,m **do**
    Initialize weights $\hat{W}_i^k$*, set the behavior of the cooperator as* $u_{i'}^{k-1}$.
    Learn the optimal behavior u_i^k*, with system* (5) *and update law* (28).
  end
end
for k = 0,...,k_m **do**
  Obtain the probabilistic distribution $\mathcal{P}(\mathcal{U}_h)$ *from* (33) *by interacting with each rational level* u_m^j.
end
Apply modeled human behavior $\mathcal{U}_h$*, learn the optimal machine behavior* $\mathcal{U}_m$*, with system* (5) *and update law* (28).

VII. Simulation Results

A. System Setup

Consider the following nonlinear affine-input system from [12]

$$\begin{aligned}\dot{x}_1 =& x_2,\\ \dot{x}_2 =& -x_2 - \frac{x_1}{2} + \frac{x_2\left(\cos\left(2x_1\right)+2\right)^2}{4} + \frac{x_2\left(\sin\left(2x_1\right)+2\right)^2}{4}\\ &+ \left(\sin\left(4x_1^2\right)+2\right)u_h + \left(\sin\left(4x_1^2\right)+2\right)(x)u_m.\end{aligned} \tag{34}$$

where $x(t) \in \mathbb{R}^2$ is the original state, $u_i \in \mathbb{R}, i = h, m$ is the policy of the human and machine player.

To stabilize the original system (34), the objective of our proposed controller is to guarantee that the state $x(t)$ converges to zero while making sure that the state does not move out of the arbitrary safe boundary, namely $x \in \mathcal{O}$, we give the following exact numerical form expressed as

$$\mathcal{O} = \{(x_1, x_2) | x_i \in (a_i, A_i), \forall i \in \{1, 2\}\},$$

where $a_1 = -1.3, a_2 = -3.1, A_1 = 0.5$, and $A_2 = 0.5$.

The initial state is selected as $x_0 = [-1, -3]$. The learning rates for the human and machine player are selected as $a_1 = 1$, $a_2 = 1$ respectively, and the weights are initialized as $\hat{W}_i^j(t_0) = [2, 2, 2]^T,\ i = h, m,\ j = 1, ..., k_m$.

The cooperative reward function is defined as

$$r(x, u_h, u_m) = M(x) + \sum_{j\in\{h,m\}} u_j^T R_j u_j, \tag{35}$$

where $M(x) = x^T x, R_h = 2I_2$ and $R_m = I_2$.

In the simulation, the rationality of human and machine is up to level-5, which is smart enough to imitate the variety of human behavior. After the learning procedure, the human impact modeling algorithm mentioned in previous section will be utilized to model a bounded rational human policy map.

B. Simulation Results

The learning process of the machine critic weights is shown in Fig. 1, which is obtained by interacting with the modeled human impact in the human-machine cooperative game. The approximated value function of cooperative game is convergent. The probabilistic

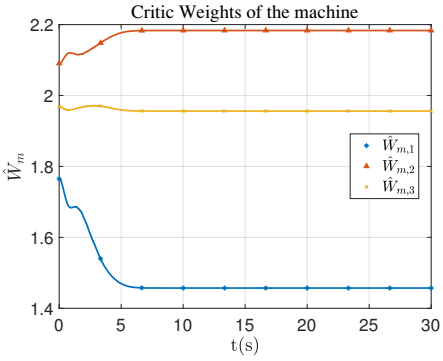

Fig. 1. machine player's value function weights

distribution of modeled human behavior is presented in Fig. 2, including intelligence from level-1 to level-5. The policy of level-3 has the highest probability, while level-1 has the lowest probability and the other levels have the highest probability. The probability of the other levels is about the same, and such a distribution pattern is consistent with the intuition of the variety of human action.

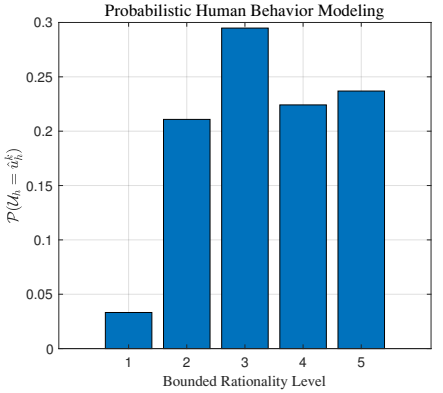

Fig. 2. Modeled human impact's distribution

The main result is presented in Fig. 3, where the state trajectories of the transformed system and non-transformed system are presented. The trajectory of the non-transformed system is reaching out the safe boundary of rectangle state constraint $\mathcal{O}$, which may cause great damage to the human player due to the vulnerability of human-beings. With the transformed system, the human-machine cooperation finally stabilize the state without violating the safety constraint.

978-1-6654-7075-9/23/$31.00 ©2023 IEEE

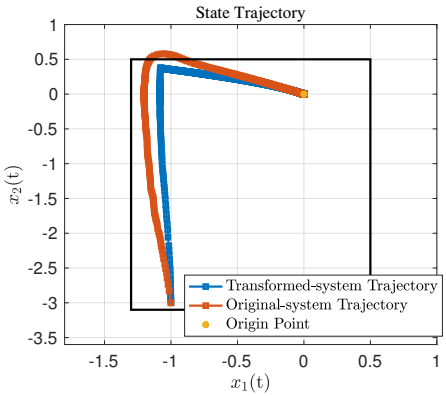

Fig. 3. State trajectories of transformed system and non-transformed system

VIII. Conclusions

This research investigates the issue of bounded rationality in human behavior for the setting of a cooperative game involving both humans and machines. Cooperation between humans and machines is an emerging subject in emergency management, and it is necessary to guarantee human safety. Initially, a state transformation based on barrier function is formulated to guarantee the safety constraints of the state of the human-machine system. The cognitive hierarchy utilizes a level-k Rationality Framework to develop knowledge of human behavior, while Adaptive Dynamic Programming is employed to attain bounded rational behavior. Drawing inspiration from sociological behavior modeling, a softmax probabilistic decision distribution is employed to mimic human behavior. Finally, a simulation has been implemented to evaluate the efficacy of the proposed framework, revealing that full state limitations and stabilization are ensured.

References

[1] Z. Li, Q. Li, P. Huang, H. Xia, and G. Li, "Human-in-the-Loop Adaptive Control of a Soft Exo-Suit With Actuator Dynamics and Ankle Impedance Adaptation," *IEEE Transactions on Cybernetics*, pp. 1–13, 2023.

[2] S. Li, N. Li, A. Girard, and I. Kolmanovsky, "Decision making in dynamic and interactive environments based on cognitive hierarchy theory, Bayesian inference, and predictive control," in *2019 IEEE 58th Conference on Decision and Control (CDC)*, pp. 2181–2187, Dec. 2019.

[3] Y. Yildiz, A. Agogino, and G. Brat, "Predicting Pilot Behavior in Medium-Scale Scenarios Using Game Theory and Reinforcement Learning," *Journal of Guidance, Control, and Dynamics*, vol. 37, pp. 1335–1343, July 2014.

[4] C. F. Camerer, T.-H. Ho, and J.-K. Chong, "A Cognitive Hierarchy Model of Games," *The Quarterly Journal of Economics*, vol. 119, pp. 861–898, Aug. 2004.

[5] Y. Zheng, H. Zhao, J. Zheng, C. He, and Z. Li, "Stackelberg-Game-Oriented Optimal Control for Bounded Constrained Mechanical Systems: A Fuzzy Evidence-Theoretic Approach," *IEEE Transactions on Fuzzy Systems*, vol. 30, pp. 3559–3573, Sept. 2022.

[6] J. Li, L. Yao, X. Xu, B. Cheng, and J. Ren, "Deep reinforcement learning for pedestrian collision avoidance and human-machine cooperative driving," *Information Sciences*, vol. 532, pp. 110–124, Sept. 2020.

[7] Y. Zhou, K. G. Vamvoudakis, W. M. Haddad, and Z.-P. Jiang, "A Secure Control Learning Framework for Cyber-Physical Systems Under Sensor and Actuator Attacks," *IEEE Transactions on Cybernetics*, vol. 51, pp. 4648–4660, Sept. 2021.

[8] R. Kamalapurkar, H. Dinh, S. Bhasin, and W. E. Dixon, "Approximate optimal trajectory tracking for continuous-time nonlinear systems," *Automatica*, vol. 51, pp. 40–48, Jan. 2015.

[9] S. M. N. Mahmud, S. A. Nivison, Z. I. Bell, and R. Kamalapurkar, "Safe Model-Based Reinforcement Learning for Systems With Parametric Uncertainties," *Frontiers in Robotics and AI*, vol. 8, p. 733104, Dec. 2021.

[10] M. H. Cohen and C. Belta, "Safe exploration in model-based reinforcement learning using control barrier functions," *Automatica*, vol. 147, p. 110684, Jan. 2023.

[11] Y. Yang, Y. Yin, W. He, K. G. Vamvoudakis, H. Modares, and D. C. Wunsch, "Safety-Aware Reinforcement Learning Framework with an Actor-Critic-Barrier Structure," in *2019 American Control Conference (ACC)*, pp. 2352–2358, July 2019.

[12] Y. Yang, K. G. Vamvoudakis, and H. Modares, "Safe reinforcement learning for dynamical games," *International Journal of Robust and Nonlinear Control*, vol. 30, pp. 3706–3726, June 2020.

[13] Y. Yang, D.-W. Ding, H. Xiong, Y. Yin, and D. C. Wunsch, "Online barrier-actor-critic learning for H control with full-state constraints and input saturation," *Journal of the Franklin Institute*, vol. 357, pp. 3316–3344, Apr. 2020.

[14] Z. Marvi and B. Kiumarsi, "Safe reinforcement learning: A control barrier function optimization approach," *International Journal of Robust and Nonlinear Control*, vol. 31, no. 6, pp. 1923–1940, 2021.

[15] N.-M. T. Kokolakis and K. G. Vamvoudakis, "Safe Finite-Time Reinforcement Learning for Pursuit-Evasion Games," in *2022 IEEE 61st Conference on Decision and Control (CDC)*, (Cancun, Mexico), pp. 4022–4027, IEEE, Dec. 2022.

[16] J. Xu, J. Wang, J. Rao, Y. Zhong, and H. Wang, "Adaptive dynamic programming for optimal control of discretetime nonlinear system with state constraints based on control barrier function," *International Journal of Robust and Nonlinear Control*, vol. 32, pp. 3408–3424, Apr. 2022.

[17] A. Kanellopoulos and K. G. Vamvoudakis, "Non-equilibrium dynamic games and cyberphysical security: A cognitive hierarchy approach," *Systems & Control Letters*, vol. 125, pp. 59–66, Mar. 2019.

[18] K. G. Vamvoudakis, F. Fotiadis, A. Kanellopoulos, and N.-M. T. Kokolakis, "Nonequilibrium dynamical games: A control systems perspective," *Annual Reviews in Control*, vol. 53, pp. 6–18, 2022.

[19] N.-M. T. Kokolakis and K. G. Vamvoudakis, "Bounded rational Dubins vehicle coordination for target tracking using reinforcement learning," *Automatica*, vol. 149, p. 110732, Mar. 2023.

[20] N. Musavi, D. Onural, K. Gunes, and Y. Yildiz, "Unmanned Aircraft Systems Airspace Integration: A Game Theoretical Framework for Concept Evaluations," *Journal of Guidance, Control, and Dynamics*, vol. 40, pp. 96–109, Jan. 2017.

[21] C. O. Yaldiz and Y. Yildiz, "Driver Modeling Using Continuous Reasoning Levels: A Game Theoretical Approach," in *2022 IEEE 61st Conference on Decision and Control (CDC)*, pp. 5068–5073, Dec. 2022.

[22] N. Li, D. W. Oyler, M. Zhang, Y. Yildiz, I. Kolmanovsky, and A. R. Girard, "Game Theoretic Modeling of Driver and Vehicle Interactions for Verification and Validation of Autonomous Vehicle Control Systems," *IEEE Transactions on Control Systems Technology*, vol. 26, pp. 1782–1797, Sept. 2018.

[23] K. Liu, N. Li, H. E. Tseng, I. Kolmanovsky, A. Girard, and D. Filev, "Cooperation-Aware Decision Making for Autonomous Vehicles in Merge Scenarios," in *2021 60th IEEE Conference on Decision and Control (CDC)*, pp. 5006–5012, Dec. 2021. ISSN: 2576-2370.

[24] R. S. Sutton and A. G. Barto, *Reinforcement Learning, second edition: An Introduction*. MIT Press, Nov. 2018.

[25] R. Tian, N. Li, I. Kolmanovsky, and A. Girard, "Beating humans in a penny-matching game by leveraging cognitive hierarchy theory and Bayesian learning," in *2020 American Control Conference (ACC)*, pp. 4652–4657, July 2020.