https://doi.org/10.26599/AIR.2023.9150027 CAAI Artificial Intelligence Research

Game Interactive Learning: A New Paradigm towards Intelligent Decision-Making

Junliang Xing[1], Zhe Wu[2], Zhaoke Yu[2], Renye Yan[2], Zhipeng Ji[2], Pin Tao[1], and Yuanchun Shi[1] ✉

ABSTRACT

Decision-making plays an essential role in various real-world systems like automatic driving, traffic dispatching, information system management, and emergency command and control. Recent breakthroughs in computer game scenarios using deep reinforcement learning for intelligent decision-making have paved decision-making intelligence as a burgeoning research direction. In complex practical systems, however, factors like coupled distracting features, long-term interact links, and adversarial environments and opponents, make decision-making in practical applications challenging in modeling, computing, and explaining. This work proposes game interactive learning, a novel paradigm as a new approach towards intelligent decision-making in complex and adversarial environments. This novel paradigm highlights the function and role of a human in the process of intelligent decision-making in complex systems. It formalizes a new learning paradigm for exchanging information and knowledge between humans and the machine system. The proposed paradigm first inherits methods in game theory to model the agents and their preferences in the complex decision-making process. It then optimizes the learning objectives from equilibrium analysis using reformed machine learning algorithms to compute and pursue promising decision results for practice. Human interactions are involved when the learning process needs guidance from additional knowledge and instructions, or the human wants to understand the learning machine better. We perform preliminary experimental verification of the proposed paradigm on two challenging decision-making tasks in tactical-level War-game scenarios. Experimental results demonstrate the effectiveness of the proposed learning paradigm.

KEYWORDS

decision-making; game interactive learning; human-computer interaction; game theory; machine learning

Real-world systems like automatic driving[1], traffic dispatching[2], information system management[3], and emergency command and control[4], inevitably involve the decision-making procedures within all their running process. The quality of the decision-making results, either accomplished by the human or the machine, thus significantly affects these systems' performance. With the fast development of Artificial Intelligence (AI) in the last decades[5,6], especially the deep learning models and algorithms, we have witnessed significant progress in high-performance intelligent perception models of audio, visual, and text data[7–9]. Due to the newly arising deep reinforcement learning algorithms, researchers have also made a substantial advance in developing human-level intelligent decision models in challenging computer games like Go[10], Pokers[11, 12], and real-time strategy games[13–15]. The MuZero model[16] achieves expert-level performance in Go, Chess, Shogi, and Atrai 2600 games simultaneously. These advances have paved decision-making intelligence as a growing research direction for further artificial intelligence developments in more complex systems.

Although we have witnessed impressive advances in AI algorithms on several decision problems, reaching human-level or even super-human-level performance in various computer game tasks. However, intelligent decision-making in complex real-world systems is still very challenging. Since they often contain multiple coupled distracting features, long-term interference factors, and adversarial environments and opponents. Previous artificial intelligence tried to solve decision-making problems automatically, ignoring humans' irreplaceable function in complex real-world systems. This desire or perspective makes applying existing AI algorithms in complex real-world decision systems infeasible. Machines are super fast at calculating data and accessing storage, while humans are very good at cognitive reasoning and instinctive decision-making. Therefore, a natural way to develop intelligent decision-making techniques for real-world complex systems is to leverage the strengths of both machines and humans to achieve a new form of hybrid human-machine intelligence.

More generally, a complex real-world system constitutes multiple different roles. Here, we provide a unified perspective to categorize these roles into three types: the *human*, the *machine*, and the *environment*. The *human* refers to the human entities involved in the system, the *machine* corresponds to the computer software or intelligent agent that functions in the system to ensure its operation, and the *environment* can be the virtual host or the natural world that the system runs in or the human and machine interact. In an intelligent decision system, human intelligence contributes to coping with tasks that the machine cannot perform well by the machine and provides supervision of the whole system. In contrast, machine intelligence contributes to helping humans fulfill more jobs and makes its role as an essential

1 Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
2 Qiyuan Laboratory, Beijing 100094, China
Junliang Xing and Zhe Wu contribute equally to this work.
Address correspondence to Yuanchun Shi, shiyc@tsinghua.edu.cn

assistant. The environment provides the foundation for tasks to run and the interface for humans and machines to participate in the system. We refer to this human-machine-environment decomposition of a complex system as the Ternary Hybrid Model (THM).

The decision-making system typically characterized by the THM faces the following challenges. (1) **Hard to model**. The interconnection of humans, machines, and the environment in THM makes modeling the interaction relationship difficult. The coexistence of competition and cooperation leads to difficulty modeling the game relationship. The emergence of single-agent and multi-agent intelligence leads to difficulty in modeling the intelligence mechanism. (2) **Hard to compute**. The machine and environment already contain vast interactive data, and THM additionally introduces human initiative. The amount of data to be handled by the intelligent system grows exponentially, and the heterogeneous data and dimensional explosion lead to difficult computation issues. (3) **Hard to explain**. Existing decision intelligence systems have suffered from poor policy interpretability and unreliable results, and the cognitive generation mechanism of human intelligence in THM has yet to be explicit.

To address the hard-to-model, hard-to-compute, and hard-to-explain problems in complex decision-making environments, we propose game interactive learning, a novel paradigm for intelligent decision-making. As shown in the Fig. 1, our novel paradigm relies on three primary and synergistic components: game theoretic approaches, machine learning techniques, and human-computer interaction. This paradigm first models intelligent decision-making systems using game-theoretic approaches to formalize them as machine-representable and mathematically solvable problems. Then, it approximates the possible solutions and evaluates them using knowledge-based and data-driven learning techniques. Finally, it introduces human-computer interaction techniques to improve the intelligibility and credibility of decisions by directly including humans in the decision loop or embedding human experience and knowledge. Based on this paradigm, we further propose a new approach as a concrete implementation of game interactive learning for decision-making problems.

We summarize the main contributions of our work:

- To our best knowledge, we propose the first paradigm of Game Interactive Learning (GIL) for solving complex decision-making problems under a unified human-machine-environment perspective.
- Guided by the GIL paradigm, we further propose a knowledge-based and data-driven approach for the initial integration and co-evolution of human and machine intelligence in decision-making problems.
- We perform a preliminary experimental validation of our methods on two challenging decision-making tasks, demonstrating the potential and scaleability of game interaction learning and methods.

1 Related Work

From the unified perspective of the THM, almost all research on decision-making intelligence can be classified according to the relationship among humans, machines, and environments. Following the history of decision-making intelligence[17], we focus on two main lines: the interaction between the machine and the environment and the hybrid interaction between the human, the machine, and the environment.

The related work on the machine and environment interaction can be classified into competitive and cooperative games according to the scenario characteristics[18]. In competitive environments, two-player zero-sum games are a fruitful area[10–14]. The optimization goal of the Counterfactual Regret Minimization (CFR) algorithm[19,20] matches the Nash Equilibrium and has worst-case guarantees. Reinforcement learning algorithms empower agents to master complex strategies from scratch self-play[21].

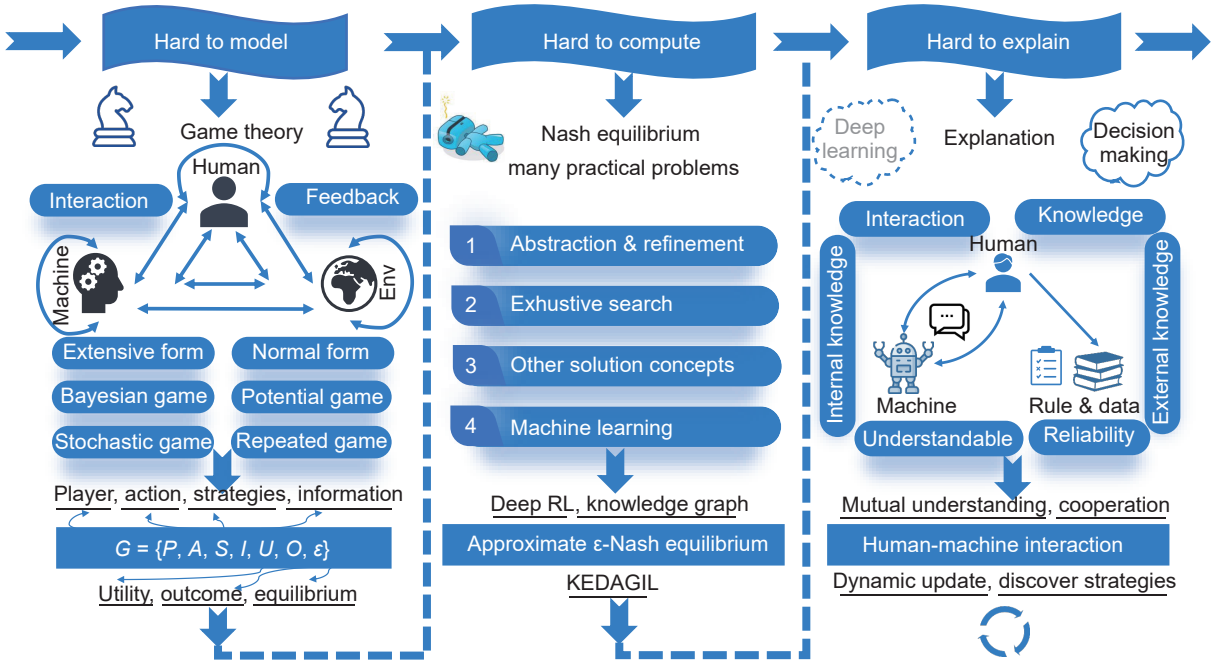

Fig. 1 Illustration of the Game Interactive Learning (GIL) paradigm. The GIL paradigm consists of three parts: game theoretic modeling, machine learning solving, and human-computer interaction explaining. In the modeling phase, we introduce the game-theoretic approach to formalize intelligent decision-making systems into machine-representable and mathematically solvable problems. In the solving phase, we propose a knowledge-based and data-driven learning approach to approximate solutions. Finally, we use human-computer interaction techniques to improve the intelligibility and credibility of the decision.

However, few real-world interactions feature pure conflict, and researchers are increasingly focusing on cooperative domains such as Hanabi[22] and Overcooked[23]. There is also a rapidly growing research interest in mixed motives, such as Poker[24] and Diplomacy[25]. At the same time, public problems involved in almost all scenarios, such as population training[26], policy adaptation[27], communication mechanisms[28], and theory of mind[29], have received much attention.

Although artificial intelligence solves some tasks with impressive performance, there are many problems that machines cannot yet solve alone. The ideal intelligent system leverages the complementary strengths of human and machine intelligence to produce a more intelligent form. The primary research focuses on such hybrid human-machine intelligent systems that integrate humans, machines, and the environment to bring in human experience and knowledge more efficiently and naturally. Most previous work has established direct interaction channels between humans and machines to achieve natural human-machine interaction[30]. The human-in-the-loop[31] incorporation of humans into existing machine learning processes, such as data labeling and model selection, is also a class of approaches that introduce human experience. Imitation learning[32] implicitly embeds human experience into decision-making by fitting human demonstrations. Similarly, some common problems in human-machine hybrid intelligent systems, such as human-machine interaction[33], understanding[34], and co-evolution[35], have received increasing attention.

2 Gaming Interactive Learning Paradigm

2.1 Game theoretical modeling

Modeling decision intelligence systems is the abstraction or representation of the decision-making process using mathematical formulations. Human-machine hybrid intelligence systems suffer from three main modeling difficulties: (1) cross-correlation and dynamic interaction between humans, machines, and the environment, (2) coexistence of competition and cooperation, (3) single-agent intelligence and multi-agent intelligence emerge together. Game theory has shown validity in scenario modeling in diverse fields such as economics, evolutionary biology, and computer science. Game theory provides a natural framework for abstracting complex systems into mathematical models using quantifiable or iterative optimization variables such as players, strategies, payoffs, and risks, which are further optimized and evaluated according to predefined criteria.

We introduce a game-theoretic approach to modeling the hybrid human-machine intelligent systems and possible studies that can be included in the following: (1) Interaction relationship modeling between entities in hybrid human–machine systems. It focuses on modeling and representing interactions between humans, between humans and machines, and between machines and machines. (2) Heterogeneous human-machine intelligence encoding and representation modeling. It studies combining human knowledge and machine computation to form an inferred representation system that integrates human, machine, and environment. (3) Modeling of optimization goals for hybrid human-machine intelligent systems. It mainly follows the setting of the solution concept in game theory to study the solution and evaluation methods under different optimization targets.

2.2 Machine learning to approximate solutions

A game system's solvability means the possible results are computed and evaluated with the given solution concept after representation and modeling. The Nash equilibrium is the most widely used solution concept. Solving the Nash equilibrium in a two-player zero-sum game can be achieved in polynomial time[36] while solving the Nash equilibrium in a two-player constant-sum game is a PPAD-complete problem[37]. When the game participates more than two players or has more decision nodes, solving the exact equilibrium solution, in this case, becomes more complicated, and it is not even possible to identify whether the equilibrium solution exists. There are usually several types of approaches to alleviating this problem: (1) abstraction and refinement of complex games into games with simple structures (zero-sum or normal-form game), (2) exhaustive search of the game via expensive computing power, (3) introducing machine learning techniques to solve approximate solutions efficiently, and (4) moving to alternative solution concepts.

Considering the huge complexity of modeling and the long decision process of human-machine hybrid intelligent systems, abstracting them into simple games or performing brute-force searches on the original game has technical limitations and poor scalability. We focus on the approximate solution using machine learning techniques, and the specific techniques used are described in Section 3.2.

2.3 Human-computer interaction technology guided intelligibility

Explainable is defined as an explanation to humans in human-understandable terms[38]. Where "explanation" is typically translated as logical decision rules, "understandable terms" are presented in natural language, pictures, data, etc., depending on the task context. Most current work focuses on explainability for deep learning[38], which aims to explain neural networks' internal operating and input-output mapping. Explainability in decision intelligence emphasizes generating explainable reasons to explain the model's decisions or unfolding the model for human users to probe and examine[39].

In the GIL framework, we do not seek complete explainability but focus on intelligibility and credibility. We integrate human-machine interaction technology into a human-machine hybrid intelligence system to improve intelligibility and credibility. On the one hand, human cognition or knowledge can be naturally embedded into the decision system due to the inclusion of humans into the decision loop, avoiding opaque decisions for black-box machine learning. On the other hand, the machine can interactively exhibit the decision process to humans, satisfying the requirement of system credibility and helping humans discover new decision strategies.

3 KEDAGIL

To provide an realization of the GIL paradigm, we propose KEDAGIL: KnowledgE-based and DatA-driven integrating approach for Gaming Interactive Learning, as shown in Fig. 2, a concrete implementation of our GIL paradigm for decision-making problems. The process of a bi-directional knowledge-data driven update is as follows:

- **Step 1 (Knowledge guiding).** Encode human knowledge and experience into complex decision-making problems to form initial solutions.
- **Step 2 (Data driving).** Agents loaded with the initial solution generate large amounts of adversarial data through gaming interactions (machine-machine or human-machine) and then use data-driven learning techniques to improve performance

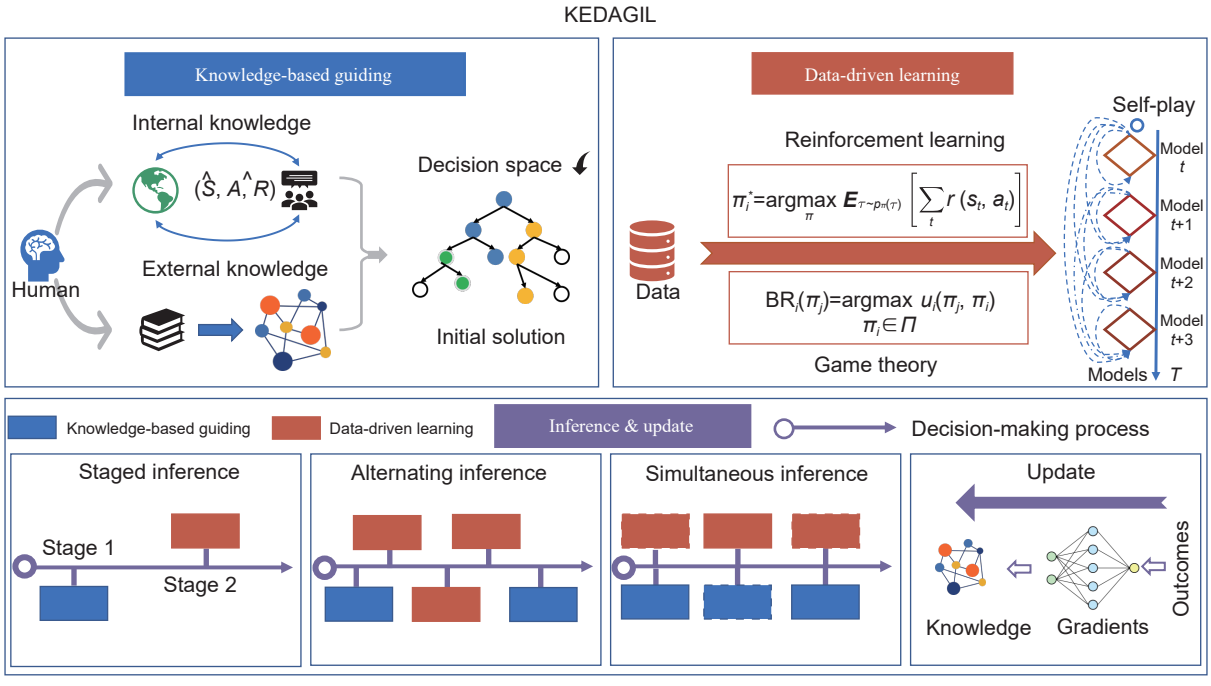

Fig. 2 The KEDAGIL approach is comprised of three steps: (1) knowledge guiding, (2) data driving, and (3) inference & update. More details about the three steps are described in Section 3.

continuously.

- **Step 3 (Inference & update).** Use the updated solution to complete forward inference and analyze the reasons for performance improvement through the interaction process.
- Repeat the above three steps to achieve bi-directional iterative enhancement of knowledge-data driven update.

Formally, we consider the knowledge-based and data-driven processes as a finite-horizon Markov decision process[40] $\mathrm{MDP} = (\hat{\mathcal{S}}, \hat{\mathcal{A}}, \mathcal{R}, \tau, \mathcal{P}, \gamma)$. Intuitively, the MDP jointly models the interaction-evolution process between environment $\mathcal{E}$, human $\mathcal{H}$, and machine $\mathcal{M}$ in a hybrid human-machine intelligence system. $\hat{\mathcal{S}} = (\mathcal{S}, \mathcal{D})$ is a hybrid-state encoding the state of the environment $\mathcal{E}$ and abstracted human knowledge $\mathcal{D}$. $\hat{\mathcal{A}} = (\mathcal{A}^H_{(t=0,1,\ldots,n)}, \mathcal{A}^M_{(t=0,1,\ldots,n)})$ is a hybrid-action containing human behavior $\mathcal{A}^H_t$ and machine action $\mathcal{A}^M_t$ at a given time step t. $\tau = \hat{\mathcal{S}} \times \hat{\mathcal{A}} \to \hat{\mathcal{S}}$ is the transition function, $\mathcal{R} = \hat{\mathcal{S}} \times \hat{\mathcal{A}} \to \mathbb{R}$ is the reward function, $\mathcal{P}$ is the initial hybrid-state distribution, and γ is the discount factor. A policy $\pi(\hat{s}_t) = P(a^M_t \mid \hat{s}_t)$ maps the hybrid-state $\hat{s}$ at the current time t to a distribution over the machine action space A^M.

3.1 Knowledge-based guiding

Human-machine hybrid intelligent systems aim to achieve superior results through interaction and understanding between humans and machines. From the human side, the primary challenge is how to encode human experience and knowledge into the human-machine hybrid system. Previous work has focused more on human-machine interactions. Such studies aim to capture real-time dynamic feedback between humans and machines to accomplish communication and understanding, rarely involving the decision-making process[41]. Some recent work also aims to incorporate the human factor into the decision-making process, proposing a human-in-the-loop reinforcement learning[42] model for policy improvement with the help of valuable features labeled by human experts. However, this approach still relies heavily on the quality of the expert's advice and introduces knowledge that is task-relevant and difficult to transfer and reuse efficiently.

In our setting, we assume that there are two types of sources of human experience and knowledge: *internal knowledge* that is task-related and fed in real-time, and *external knowledge* that contains common sense, principles, or generic decision experience.

Internal knowledge. We design two ways to embed internal knowledge into the decision system: (1) Humans participate explicitly in decision-making through interaction, which may occur when the system requires humans to make high-level decisions or when machine intelligence is weak. (2) Implicit human participation in decisions through behavioral cloning by collecting decision data from human experts in advance and then minimizing the divergences between machine and expert strategies in T time steps over N episodes.

$$\min \sum_{n=1}^{N} \sum_{t=1}^{T} [P(a^H \mid \hat{s}) R(\hat{s}, a^H) - P(a^M \mid \hat{s}) R(\hat{s}, a^M)] \tag{1}$$

External knowledge. We use knowledge graph to encode external knowledge. The inherent connectivity between external knowledge allows graph structure to uncover implicit connections, infer new knowledge, and support scaling to larger data sizes. The knowledge graph's dynamic, scalable, and domain-independent properties also enable humans and machines to understand, infer, and interpret knowledge better.

By combining internal and external knowledge into the decision system, humans can guide the early strategies of machine learning. Relying on the intuition humans provide, the machine can be given optimal guidance relatively fast, significantly narrowing the machine's search for a better solution in the decision space and giving an initial solution.

3.2 Data-driven learning

After obtaining the initial solution by knowledge guidance, we use a data-driven manner in the second stage to push the initial solution forward to explore the global optimal. Considering the great breakthroughs in decision-making with Deep Reinforcement Learning (DRL), we combine game theory with

DRL to continuously improve the initial solution by leveraging the large amount of data generated during the human-machine and machine-machine interactions.

In competitive scenarios, an agent aims to defeat a human or computer opponent. Defeating the opponent has different solution forms in the context of game theory. We mainly consider an equilibrium solution to maintain payoffs in the worst case and the best response solution to exploit the suboptimal opponent. These two types of solutions could be included in game theory's Nash equilibrium solution concept.

The best response of player i against j can be defined as

$$\mathrm{BR}_i(\pi_j) = \arg\max_{\pi_i \in \Pi} u_i(\pi_j, \pi_i) \tag{2}$$

where j represents any player other than i, u_i denotes gaming payoff. When the opponent's strategy π_j is fixed over time or can be modeled using a period of interaction data, we use the DRL algorithm to maximize the cumulative game reward to find the best response π_i^* against the current opponent:

$$\pi_i^* = \arg\max_{\pi} \boldsymbol{E}_{\tau \sim p_\pi(\tau)} \left[\sum_t r(s_t, a_t) \right] \tag{3}$$

A strategy profile is a Nash equilibrium when either player in the game is the best response against the other players. No player in this strategy profile can obtain higher payoffs by deviating from the Nash equilibrium. When the size of the game is so complex that finding a Nash Equilibrium becomes impossible, approximate Nash equilibrium becomes an optional solution. Approximate Nash equilibrium can be defined as

$$\mathrm{BR}_i^{\varepsilon}(\pi_{-i}) = \left\{ \pi_i \in \sum : u_i(\pi_i, \pi_{-i}) \right\} \geqslant \max_{\pi_i \in \Pi} u_i(\pi_i, \pi_{-i}) - \varepsilon \tag{4}$$

To compute approximate Nash equilibrium, we follow the setting of self-play, where agents learn to master the game by playing against themselves. Numerous studies have shown that self-play with simple game rules can emerge with complex strategies. In a naive self-play implementation, one of the two players is fixed as the opponent's strategy, and the other player updates its parameters as a learner. The performance of both players is gradually improved through continuous alternate learning. However, when the policy space is extensive and contains cycles, the naive self-play suffers from policy forgetting, i.e., forgetting how to defeat earlier strategies. To handle this issue, we freeze the opponent policy's checkpoints that perform competitively and maintain a policy pool. Players seek to perfect a strong and diverse policy by playing mixed matches against opponents in the policy pool.

3.3 Inference & update

KEDGIL comprises two modules: (1) knowledge guiding and (2) data driving. In the knowledge guidance phase, we encode human knowledge and experience into internal and external knowledge. Internal knowledge encodes task-relevant human knowledge through Human-machine interaction or rule guide, and external knowledge organizes task-generic human experience via knowledge graphs. In the data-driven phase, a large amount of data is generated in the interaction that is far beyond the upper limit of what human knowledge can handle. The combination of game theory and DRL can comfortably handle such a scale of data and discover new strategies and relationships from it that humans do not recognize.

The role of humans is crucial in our paradigm. In addition to contributing experience and knowledge, humans need to arrange and combine knowledge-guided and data-driven modules to make inferences for the task. Based on the timing and manner of human involvement in decision-making, we classify the Knowledge-based and data-driven inference model into three categories.

Staged inference. This is a two-stage inference mode. In the first stage, the knowledge-guided module relies on human-provided intuition and experience to give an initial solution. In the second stage, the data-driven learning module continuously improves the initial solution by iterative updates.

Alternating inference. This inference mode divides the decision process according to the task characteristics. The data-driven module completes the tasks suitable for machine processing, and the knowledge-guided module deals with those requiring high-level human decisions and recommendations. The advantage of this inference mode is that it improves the coupling between knowledge-guided and data-driven and benefits from their complementarity.

Simultaneous inference. In this inference mode, the knowledge-guided and data-driven modules give suggestions separately at each decision point. The final decision result is selected after human evaluation. This inference mode guarantees security and interpretability by ensuring human review of the entire decision.

Update. Once we identify the inference mode and complete the forward inference, we can use any DRL algorithm to perform a gradient descent of the data-driven module. However, updating only the data-driven module may be unstable because the task-relevant knowledge in the knowledge-guide module is encoded based on previous experience. To handle this issue, we update the knowledge-guiding module after each discovery of a new policy in the data-driven module.

4 Experiment

In our experimental evaluation, we aim to demonstrate the viability of our proposed game interaction learning paradigm in dealing with decision-making problems and to validate the effectiveness of our proposed approach based on the paradigm on specific tasks. We perform a preliminary experimental validation of the KEDAGIL approach on two challenging decision-making tasks in tactical-level War-game scenarios.

4.1 Tactical-level War-game Scenario I

War-game is played on a map with a hexagonal grid matrix. Players from two camps (Red and Blue) control their agents for military simulation deductions. In a medium undulating terrain scenario, as shown in Fig. 3, Red and Blue each control two tank units around the control point in real-time against each other. The agent can move in either direction of the hexagonal grid. Each tank unit has maximum Health Points (HP) in the initial state, and when HP drops to 0 means destroyed. The win or loss of the simulation is measured by the score, considering the control situation, the agents' survival, and the number of opponents destroyed. The key to winning is to make sensible macro decisions between attacking, saving forces, and capturing control points, combined with the corresponding micro actions.

Referring to human players' experience, we categorize all the possible policies that can be executed in this scenario into four base policies: *Ambush Defence* (AD) policy, *Delaying Defence* (DD) policy, *Highland Offense* (HO) policy, and *Reverse Slope Offense* (RSO) policy. We convert these four policies into executable rule AI and pairwise perform 100 simulated

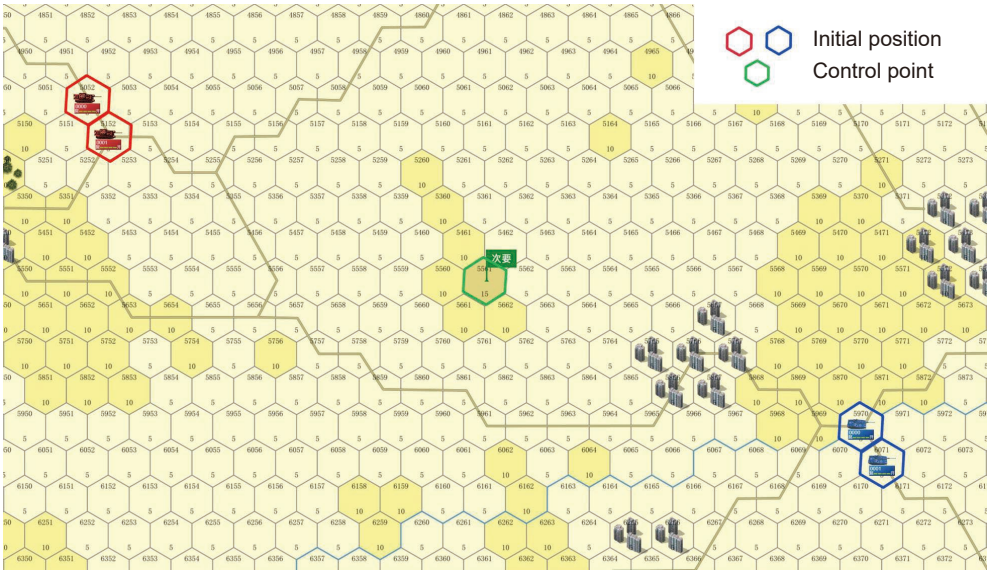

Fig. 3 Scenario I: medium undulating terrain.

deductions, and the win rate matrix calculated according to the simulation rules is shown in Table 1.

Based on the matrix, we know that the RSO policy and AD policy are the dominant strategies for Red, and the DD policy and RSO policy are the dominant strategies for Blue. We can calculate the Mixed Strategy Equilibrium. Assume that Red's mixed strategy is $\sigma_r = (x, 1-x)$, and Blue's mixed strategy is $\sigma_b = (y, 1-y)$. The expected utility functions of Red and Blue can be represented by $u_r(\sigma_r, \sigma_b)$ and $u_b(\sigma_r, \sigma_b)$. Let $\frac{\partial u_r(\sigma_r, \sigma_b)}{\partial x} = 0$ and $\frac{\partial u_r(\sigma_r, \sigma_b)}{\partial y} = 0$, and we can solve to get the mixed equilibrium strategy as shown in Table 2.

Based on the analysis drawn from game theoretic models, we use a data-driven approach to improve strategy performance continuously. We train the agents using the self-play framework introduced in Section 3.2. We use a PPO implementation that integrates a dual-clip technique to guarantee that the policy gradient decreases monotonically in the trust region. The main objective of the PPO is the following:

$$\hat{E}_t\left[\max\left(\min\left(r_t(\theta)\hat{A}_t \text{clip}\left(r_t(\theta), 1-\varepsilon, 1+\varepsilon\right)\hat{A}_t\right), c\hat{A}_t\right)\right] \quad (5)$$

where $r_t(\theta)$ denotes the probability ratio between the new policy $\pi_\theta(a_t \mid s_t)$ and old policy $\pi_{\text{old}}(a_t \mid s_t)$ at timestamp t, $\hat{A}_t$ is an estimator of the advantage function, and ε and c are hyperparameters to clip the probability ratio.

Table 1 Win rate matrix of the base policy.

Red	Blue			
	HO	RSO	AD	DD
HO	50.0%	29.0%	32.4%	50.0%
RSO	71.0%	50.0%	57.0%	69.0%
AD	72.0%	90.0%	48.2%	46.4%
DD	50.0%	31.0%	53.6%	50.0%

Table 2 Mixed strategy equilibrium in Scenario I.

Camp	p_1	p_2
Red	0.304 (AD policy)	0.696 (RSO policy)
Blue	0.639 (DD policy)	0.361 (RSO policy)

The reward function contains two parts: intrinsic and extrinsic rewards. Since the state action space of the War-game is enormous and the rewards are sparse, it is challenging to explore valid strategies by relying on extrinsic rewards alone. We introduce intrinsic rewards to guide the exploration of the agents in the early stage of training. Specifically, the agent receives an intrinsic reward upon achieving any subgoal from the set G = {reaching ambush points, reaching defense points, reaching high ground, reaching reverse slopes}. These subgoals correspond to the four base policies outlined in Table 1. As previously mentioned, these policies encapsulate all potential strategies within this scenario. Consequently, the intrinsic rewards derived from these base policies are dense and effectively guide the agent's exploration. The redesigned reward function is as follows:

$$r = r_{\text{ext}} + \tau \cdot r_{\text{int}} \quad (6)$$

where r_{ext} is the extrinsic reward based on the adjudication rule, and r_{int} is the intrinsic reward based on human experience and knowledge, τ decays to 0 as the number of steps increases.

We further integrate human experience to optimize the self-play training process. Specifically, the opponent pool $\mathcal{O}$ contains checkpoints from the training history and style-specific opponent models selected based on base policies.

We save the model's checkpoints every 12 000 timestamps and evaluate it against 9 expert knowledge AIs, computing and visualizing the average win rate. As shown in Fig. 4, during the early stages of training, intrinsic rewards guide the strategy to

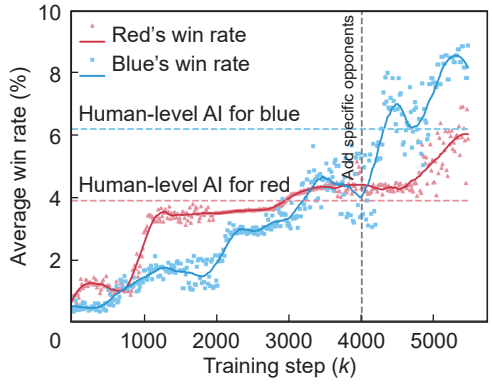

Fig. 4 Training curve in the tactical-level War-game Scenario I.

explore efficiently, and performance improves quickly. However, human empirical knowledge has limitations, and the strategy performance curve flattens. We add specific strategies that match the base policies' style to the opponent pool as fixed opponents for training. The strategy performance breaks through the bottleneck, eventually reaching an average win rate of 69% for the red side and 85% for the blue side.

4.2 Tactical-level War-game Scenario II

Figure 5 shows that the scenario of the encounter at the riverine and paddy field is loaded in the War-game's engine, introduced in Section 4.1. In this scenario, red and blue control six operators around the primary and secondary capture points against each other, and the force disposition of both sides is shown in Table 3. Compared to the medium undulating terrain scenario, this scenario has several major challenges: (1) More number of operators and heterogeneity. In addition to Medium Tanks (MT), Additional Medium Vehicles (MV), Heavy Vehicles (HV), Heavy Tanks (HT), Infantry (IN), Loitering Munition (LM), Unmanned Ground Vehicles (UGV), and other types of operators have been added. (2) More complex terrain and larger operational (mobility) areas. (3) Asymmetric scenarios where the opposing sides differ in the operators' number, type, and initial point conditions.

The division of forces between red and blue is based on the distribution of terrain and the initial positions. The blue side has the advantage in terrain and is closer to the secondary control point. By analyzing the force distribution and equipment advantages of red and blue sides, combined with human experts' experience and knowledge, we formed a strategy set Π containing three types of knowledge-based AI: conservative, aggressive, and equilibrium. Conservative strategy priority is approaching the battlefield's center, occupying favorable terrain, and then waiting for the opportunity to seize control. The aggressive strategy outflanked the battlefield, taking advantage of the view and prioritizing the elimination of the enemy before seizing control. The equalization strategy starts with a circuitous to scout and strike the enemy at the start, and then quickly moves into the center of the battlefield to take control point as the main battlefield.

The sample size of expert experience and data is limited, has cognitive limitations, and cannot be generalized to more complex or new environments. We follow KEDAGIL's setting and use machine learning techniques to enhance the decision-making effects based on three rule-based base strategies after game theoretical modeling. We initialize the neural network π_θ, which integrates the three types of rule-based AI according to Eq. (1). The performance of the optimized strategy π_θ can approximate the strategy set Π. Although π_θ can integrate the advantages of various policies, the sample data generated by the base policy alone needs to be more diverse to break the bottleneck of human decision-making ability. To address this issue, we use the improved self-play in Section 3.2 to continue to improve the algorithm's effectiveness. In the early stage of training, we initialize the opponent pool $\mathcal{O}$ using the base policy set Π, and add newly generated opponent strategies to the opponent pool every τ step.

Each opponent model maintains an initial score q to determine the probability p that the current model is selected:

$$p_i = \frac{q_i \times s_i}{\sum_{q_j \in \mathcal{O}} q_j \times s_j} \tag{7}$$

where s_i is the number of times the current opponent model i is selected, and the score q_i of model i is updated according to the match result:

$$q_i \leftarrow q_i - \frac{\eta}{\mathcal{N} p_i} \tag{8}$$

where $\mathcal{N}$ is the number of models in the opponent pool $\mathcal{O}$.

During training, we set a statistic z to implement a fallback and replacement mechanism to prevent the model from falling into a local optimum. Specifically, if the model fails in the current round,

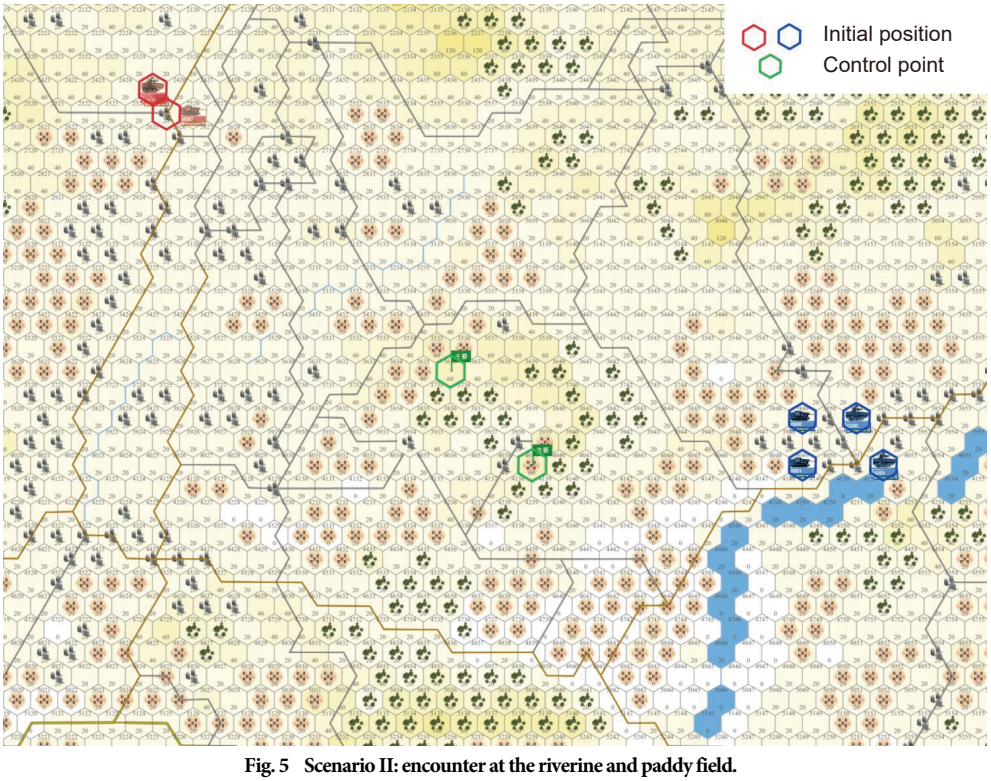

Fig. 5 Scenario II: encounter at the riverine and paddy field.

Table 3 Division of forces between red and blue in tactical-level War-game Scenario II.

Camp	Operator	HP	Initial position
Red	MT	4	2526
	MV	4	2425
	LM	4	2425
	LM	4	2425
	IN	4	2425
	UGV	4	2425
Red	HT	3	3849
	HT	3	4049
	HV	3	3851
	IN	3	3851
	HV	3	4052
	IN	3	4052

z increases by 1; otherwise, it is set to 0. When z reaches a threshold, the fallback mechanism is triggered, returning the current model parameters to the previous version. If z reaches the threshold multiple times, we replace the model with the top performer from the opponent pool $\mathcal{O}$.

Table 4 shows the win rate of the base policies Π, the integrated policy π_θ, and the KEDAGIL policy $\pi_{\theta'}$ against the eight built-in AIs. The π_θ integrates the respective advantages of the base policies and can approximate the best performance against different built-in AIs. However, the base and integrated policies' win rates are low. Figure 6 shows the results of continuing the training using self-play after using the integrated policy as the initialization model. Early in training, the performance of the policy improves faster due to suitable initialization. In Fig. 6, the win rate dropped twice due to the strategy's cyclic and non-transitive properties. However, the model's win rate improved and stabilized with the activation of fallback and replacement mechanisms. Table 4 shows that the KEDAGIL policy has the highest win rate against all types of built-in AI, with an average win rate of 93.5%.

5 Discussion

A generalized decision-making intelligence system should comprise humans, machines, and an environment. Different learning paradigms exist considering the combinations and interactions between different parts: human-machine competition, human-machine collaboration, human-environment interaction, reinforcement learning, and human-in-the-loop. We propose a novel paradigm of gaming interactive learning driven by a mixture of knowledge and data. As shown in Fig. 7, the connections and differences between gaming interactive learning and existing paradigms are as follows.

Comparison with human-machine competition. The human-machine competition focuses on how machine intelligence beats human intelligence, which is a hot topic in artificial intelligence. Before the era of deep learning, the key techniques in human-computer competition revolved around the framework of "search-evaluate-optimize", in which typical algorithms include Mini-max search and alpha-beta pruning. However, such algorithms rely on brute force exhaustion and can only conquer games of limited complexity. In recent years, machine intelligence has achieved far better performance than humans in complete information game scenarios with the rise of deep learning and deep reinforcement learning techniques. It is fast approaching the level of top human players in multi-player incomplete information games. The human-machine competition focuses on competitive scenarios but does not provide opportunities for agents to learn how to cooperate.

Comparison with human-machine collaboration. Human-machine collaboration studies how humans and AI agents can work together to accomplish a common goal. Recently, there has been a growing interest in designing machine learning agents capable of collaborating with humans, such as Hanabi, Overcooked, and team games where competition and cooperation coexist. Research on human-machine collaboration currently focuses on 1. understanding other agents and humans, their beliefs, motivations, and goals. 2. communication between humans and agents, including establishing a bidirectional communication channel to overcome value misalignment. The current focus of the human-machine collaboration is still on improving machine intelligence. It rarely involves bringing human experience and knowledge into the solution process to achieve the co-evolution of humans and machines.

Comparison with human-in-the-loop. Rapid machine intelligence advances show better operational efficiency and statistical accuracy than humans. Nevertheless, humans can rely on experience and knowledge accumulation to quickly learn and generalize explainable and scalable patterns from a few samples. The concept of human-in-the-loop was proposed in recognizing the necessity of interdependence between humans and machines, hoping to train better models with minimal cost by integrating human knowledge and experience. However, the technical framework of human-in-the-loop focuses on introducing human guidance using human-machine interaction, such as labeling and selecting data or models through interaction in data processing and during model training, and needs a focus on decision-making issues. Human-in-the-loop RL attempts to address this issue but still needs an approach that systematically integrates human knowledge and experience.

Figure 7 shows that gaming interactive learning focuses on human-machine interaction, competition, and cooperation. Compared to the human-in-the-loop system, gaming interactive learning introduced a game theoretical modeling approach and an empirical knowledge integration approach based on human-

Table 4 Win rate of the base policies Π, the integrated policy π_θ, and the KEDAGIL policy $\pi_{\theta'}$ against the eight built-in AIs.

Policy	AI-1	AI-2	AI-3	AI-4	AI-5	AI-6	AI-7	AI-8	Average
Conservative policy	0.410	0.392	0.352	0.426	0.532	0.422	0.312	0.247	0.386
Aggressive policy	0.272	0.366	0.350	0.466	0.250	0.328	0.426	0.540	0.374
Equilibrium policy	0.148	0.424	0.184	0.240	0.510	0.528	0.362	0.388	0.348
Integrated policy	0.400	0.378	0.348	0.440	0.528	0.516	0.412	0.536	0.444
KEDAGIL policy	**0.822**	**0.864**	**0.820**	**0.948**	**0.950**	**0.852**	**1.000**	**0.840**	**0.935**

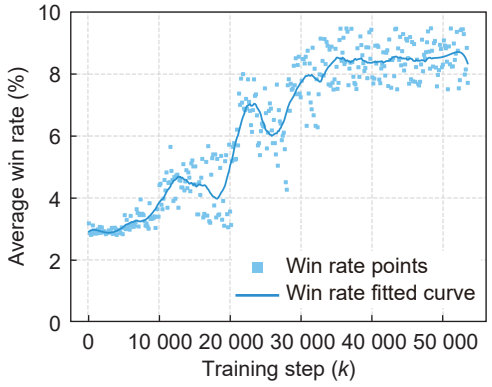

Fig. 6 Training win rates and win rate fitting curve of KEDAGIL policy in the tactical-level War game Scenario II.

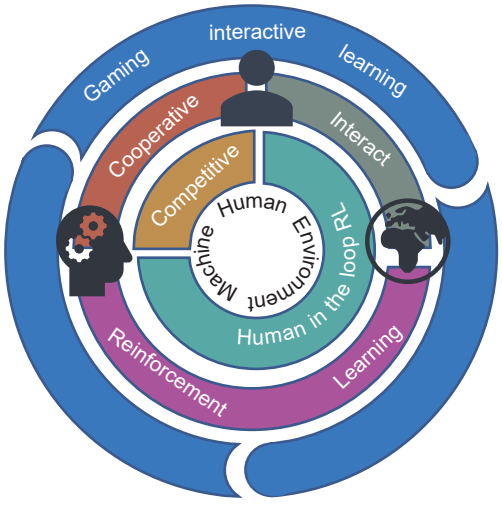

Fig. 7 Various learning paradigms in decision systems involving humans, machines, and environments.

machine interaction. Compared to several other paradigms, gaming interactive learning covers a broader range and achieves a closed loop of humans, machines, and the environment.

6 Conclusion

In this work, we first provide a unified view of the composition in a complex decision-making system by decomposing it into human, machine, and environment. We propose game interactive learning, a new paradigm, by fully considering the role of each component and the relationship between them. Based on the game interactive learning paradigm, we propose a hybrid knowledge-based and data-driven method to solve the hard-to-model, hard-to-compute, and hard-to-explain problems of complex decision-making systems. By integrating the advantages of humans and machines, we can achieve superior and more explanatory results than existing artificial intelligence, facilitating the progress of machines and humans together. Preliminary experimental verification of the proposed paradigm on two challenging decision-making tasks demonstrates promising results of the proposed paradigm and approach. In the future, we plan to extend the proposed paradigm and method to more decision-making tasks to facilitate further development in the field of human-machine hybrid intelligence.

Article History

Received: 2 August 2023; Revised: 11 October 2023; Accepted: 10 December 2023

References

[1] D. S. Gonzalez, M. Garzon, J. S. Dibangoye, and C. Laugier, Human-like decision-making for automated driving in highways, in *Proc. IEEE Intelligent Transportation Systems Conf.* (*ITSC*), Auckland, New Zealand, 2019, 2087–2094.

[2] A. Tazoniero, R. Gonçalves, and F. Gomide, Decision making strategies for real-time train dispatch and control, in *Analysis and Design of Intelligent Systems Using Soft Computing Techniques*, P. Melin, O. Castillo, E. Gomez Ramírez, J. Kacprzyk, W. Pedrycz Eds. Heidelberg, Germany: Springer, 2007, pp. 195–204.

[3] S. Nowduri, Management information systems and business decision making: Review, analysis, and recommendations, *J. Manag. Mark. Res.*, vol. 7, pp. 1–8, 2011.

[4] W. L. Waugh Jr., Mechanisms for collaboration in emergencymanagement: ICS, NIMS, and the problem with command and control, in *The collaborative public manager: New ideas for the twenty-first century*, R. O'Leary and L. B. Bingham Eds. Washington, DC, USA: Georgetown University Press, 2009, pp. 157–175, 2009.

[5] G. E. Hinton and R. R. Salakhutdinov, Reducing the dimensionality of data with neural networks, *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[6] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[7] H. Purwins, B. Li, T. Virtanen, J. Schluter, S. Y. Chang, and T. Sainath, Deep learning for audio signal processing, *IEEE J. Sel. Top. Signal Process.*, vol. 13, no. 2, pp. 206–219, 2019.

[8] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, Deep learning for visual understanding: A review, *Neurocomputing*, vol. 187, pp. 27–48, 2016.

[9] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, Deep Learning: Based Text Classification, *ACM Comput. Surv.*, vol. 54, no. 3, pp. 1–40, 2022.

[10] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot et al., Mastering the game of Go with deep neural networks and tree search, *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[11] M. Bowling, N. Burch, M. Johanson, and O. Tammelin, Heads-up limit hold'em poker is solved, *Science*, vol. 347, no. 6218, pp. 145–149, 2015.

[12] D. Zha, J. Xie, W. Ma, S. Zhang, X. Lian, X. Hu, and J. Liu, DouZero: Mastering DouDizhu with self-play deep reinforcement learning, in *Proc. 38th Int. Conf. Machine Learning*, virtual, 2021, pp. 12333–12344.

[13] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev et al., Grandmaster level in StarCraft II using multi-agent reinforcement learning, *Nature*, vol. 575, no. 7782, pp. 350–354, 2019.

[14] C. Berner, G. Brockman, B. Chan, V. Cheung, P. Dębiak, C. Dennison, D. Farhi, Q. Fischer, S. Hashme, C. Hesse, et al., Dota 2 with large scale deep reinforcement learning, arXiv preprint arXiv: 1912.06680, 2019.

[15] D. Ye, G. Chen, W. Zhang, S. Chen, B. Yuan, B. Liu, J. Chen, Z. Liu, F. Qiu, H. Yu, et al., Towards playing full MOBA games with deep reinforcement learning, in *Proc. 34th Int. Conf. Neural Information Processing Systems*, virtual, 2020, pp. 621–632.

[16] J. Schrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, A. Guez, E. Lockhart, D. Hassabis, T. Graepel et al., chess and shogi by planning with a learned model, *Nature*, vol. 588, no. 7839, pp. 604–609, 2020.

[17] Y. Duan, J. S. Edwards, and Y. K. Dwivedi, Artificial intelligence for decision making in the era of Big Data–evolution, challenges and research agenda, *Int. J. Inf. Manag.*, vol. 48, pp. 63–71, 2019.

[18] A. Tampuu, T. Matiisen, D. Kodelja, I. Kuzovkin, K. Korjus, J. Aru, J. Aru, and R. Vicente, Multiagent cooperation and competition with deep reinforcement learning, *PLoS One*, vol. 12, no. 4, p. e0172395,

2017.

[19] M. Zinkevich, M. Johanson, M. Bowling, and C. Piccione, Regret minimization in games with incomplete information, in *Proc. 21st Annu. Conf. Neural Information Processing Systems*, Vancouver, Canada, 2007, pp. 1729–1736.

[20] Z. Wang, C. Mu, S. Hu, C. Chu, and X. Li, Modelling the dynamics of regret minimization in large agent populations: a master equation approach, in *Proc. 31st Int. Joint Conf. Artificial Intelligence*, Vienna, Austria, 2022, pp. 23–29.

[21] J. Heinrich, M. Lanctot, and D. Silver, Fictitious self-play in extensive-form games, in *Proc. 32nd Int. Conf. Machine Learning*, Lille, France, 2015, pp. 805–813.

[22] N. Bard, J. N. Foerster, S. Chandar, N. Burch, M. Lanctot, H. F. Song, E. Parisotto, V. Dumoulin, S. Moitra, E. Hughes, et al., The Hanabi challenge: A new frontier for AI research, *Artif. Intell.*, vol. 280, p. 103216, 2020.

[23] D. J. Strouse, K. R. McKee, M. M. Botvinick, E. Hughes, and R. Everett, Collaborating with humans without human data, in *Proc. 35th Annu. Conf. Neural Information Processing Systems*, 2021, virtual, pp. 14502–14515.

[24] N. Brown and T. Sandholm, Superhuman AI for multiplayer poker, *Science*, vol. 365, no. 6456, pp. 885–890, 2019.

[25] Meta Fundamental AI Research Diplomacy Team (FAIR), A. Bakhtin, N. Brown, D. E., G. Farina, C. Flaherty, D. Fried, A. Goff, J. Gray, H. Hu, et al., Human-level play in the game of *Diplomacy* by combining language models with strategic reasoning, *Science*, vol. 378, no. 6624, pp. 1067–1074, 2022.

[26] M. Jaderberg, V. Dalibard, S. Osindero, W. MCzarnecki, J. Donahue, A. Razavi, O. Vinyals, T. Green, I. Dunning, K. Simonyan, et al., Population based training of neural networks, arXiv preprint arXiv: 1711.09846, 2017.

[27] Z. Wu, K. Li, H. Xu, Y. Zang, B. An, and J. Xing, L2E: Learning to exploit your opponent, in *Proc. Int. Joint Conf. Neural Networks* (*IJCNN*), Padua, Italy, 2022, pp. 1–8.

[28] J. N. Foerster, Y. M. Assael, N. de Freitas, and S. Whiteson, Learning to communicate with deep multi agent reinforcement learning, in *Proc. 30th Conf. Neural Information Processing Systems* (*NIPS 2016*), Barcelona, Spain, pp. 2145–2153, 2016.

[29] N. Rabinowitz, F. Perbet, F. Song, C. Zhang, S. M. Ali Eslami, and M. Botvinick, Machine theory of mind, in *Proc. 35th Int. Conf. Machine Learning*, Stockholm, Sweden, 2018, pp. 4218–4227.

[30] Y. J. Liu, M. Yu, G. Zhao, J. Song, Y. Ge, and Y. Shi, Real-time movie-induced discrete emotion recognition from EEG signals, *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 550–562, 2018.

[31] F. M. Zanzotto, Viewpoint: Human-in-the-loop artificial intelligence, *J. Artif. Intell. Res.*, vol. 64, pp. 243–252, 201.

[32] A. Hussein, M. M. Gaber, E. Elyan, and C. Jayne, Imitation learning: A survey of learning methods, *ACM Comput. Surv.*, vol. 50, no. 2, p. 21.

[33] D. Wang, E. Churchill, P. Maes, X. Fan, B. Shneiderman, Y. Shi, and Q. Wang, From human-human collaboration to human-AI collaboration: Designing AI systems that can work together with people, in *Proc. Extended Abstracts of the 2020 CHI Conf. Human Factors in Computing Systems*, Honolulu, HI, USA, 2020, pp. 1–6.

[34] L. Yuan, X. Gao, Z. Zheng, M. Edmonds, Y. N. Wu, F. Rossano, H. Lu, Y. Zhu, and S. C. Zhu, *In situ* bidirectional human-robot value alignment, *Sci. Robot.*, vol. 7, no. 68, p. eabm4183, 2022.

[35] A. Dengel, L. Devillers, and L. M. Schaal, Augmented human and human-machine co-evolution: Efficiency and ethics, in *Reflections on Artificial Intelligence for Humanity*, Cham, Switzerland: Springer, 2021, pp. 203–227.

[36] J. Perolat, B. Scherrer, B. Piot, and O. Pietquin, Approximate dynamic programming for two-player zero-sum Markov games, in *Proc. 32nd Int. Conf. Machine Learning*, Lille, France, 2015, pp. 1321–1329.

[37] R. Mehta, Constant rank two-player games are PPAD-hard, *SIAM J. Comput.*, vol. 47, no. 5, pp. 1858–1887, 2018.

[38] F. Doshi-Velez and B. Kim, Towards a rigorous science of interpretable machine learning, arXiv preprint arXiv: 1702.08608, 2017.

[39] M. T. Ribeiro, S. Singh, and C. Guestrin, "why should I trust you?": Explaining the predictions of any classifier, in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 1135–1144.

[40] M. Mundhenk, J. Goldsmith, C. Lusena, and E. Allender, Complexity of finite-horizon Markov decision process problems, *J. ACM*, vol. 47, no. 4, pp. 681–720, 2000.

[41] C. Yu, Y. Gu, Z. Yang, X. Yi, H. Luo, and Y. Shi, Tap, dwell or gesture?: Exploring head-based text entry techniques for HMDs, in *Proc. 2017 CHI Conf. Human Factors in Computing Systems*, Denver, CO, USA, 2017, pp. 4479–4488.

[42] K. Lee, L. M. Smith, and P. Abbeel, Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training, in *Proc. 38th Int. Conf. Machine Learning*, virtual, 2021, pp. 6152–6163.