Over the Quantum Rainbow: Explaining Hybrid Quantum Reinforcement Learning

Junghoon Park[1], Jiook Cha[1,2,3,4] Samuel Yen-Chi Chen[5], Shinjae Yoo[6], Huan-Hsin Tseng[6]
[1]Interdisciplinary Program in Artificial Intelligence, Seoul National University
[2] Department of Psychology, Seoul National University
[3] Department of Brain and Cognitive Sciences, Seoul National University
[4] Graduate School of Data Science, Seoul National University
[5] Wells Fargo
[6] Computational Science Initiative, Brookhaven National Laboratory
utopie9090@snu.ac.kr, connectome@snu.ac.kr, yen-chi.chen@wellsfargo.com, {sjyoo, htseng}@bnl.gov

***Abstract*—In the realm of artificial intelligence, deep reinforcement learning (RL) agents struggle with generalizability and require substantial computational resources, unlike humans who easily adapt and generalize across tasks. To address these challenges, we introduce Quantum Rainbow, a hybrid algorithm that leverages the neural mechanisms of human decision-making and the efficiency of quantum computing. Quantum Rainbow combines variational quantum circuits with the Rainbow Deep Q-Network (DQN) model to create a novel approach in reinforcement learning that integrates quantum principles into deep learning paradigms. We evaluate our model using behavioral experiments through the Iowa Gambling Task and 4-Armed Bandit Task. Our investigations reveal a significant relationship between the architecture of quantum circuits and the performance of quantum RL agents. Specifically, using causal discovery methods, we demonstrate the critical role of quantum entanglement in enhancing model performance. These findings not only show promising results but also pave the way for future explorations into optimizing quantum circuit architectures for reinforcement learning applications. This study underscores the potential of quantum-enhanced algorithms to achieve "*quantum advantage*" by addressing fundamental limitations in conventional deep RL methods.**

I. Introduction

Reinforcement learning (RL) may be a key to an organism's adaptation to the environment. RL is defined as a type of learning that the agent, who chooses his/her actions based on sensory inputs of the environmental state, learns to optimize his/her decisions that maximize rewards and minimize punishments. Decades of research has shown similarities between the RL models of artificial and human intelligence [1], [2]. RL theories are in line with the activity of dopamine neurons as reward prediction errors (RPEs) [3]–[5]. They also account for the mechanisms of the brain's mesocorticolimbic system involved in reward-driven learning and decision-making: e.g., dorsolateral prefrontal cortex correlates to task representation, dorsolateral striatum to action selection, ventral striatum, ventral medial prefrontal cortex, and orbitofrontal cortex to reward estimation [6], [7].

In the rapidly evolving landscape of artificial intelligence, the capacity for adaptive learning in complex and dynamic environments remains a pivotal challenge. Current research in RL seeks to bridge the gap between artificial agents and the natural learning capabilities observed in humans and animals. This pursuit is informed by profound insights derived from neurocognitive science, suggesting that learning mechanisms employed by biological systems could guide the development of more efficient and adaptable artificial agents [8]. Traditional deep RL algorithms, inspired by these biological frameworks, perform complex nonlinear mappings from perceptual inputs to actionable outputs, reflecting the intricate decision-making processes of the human brain [9]–[12]. Notably, the advent of Distributional RL has advanced this paradigm by representing RPE signals as vectors, thus capturing a broader spectrum of potential decision-making scenarios [13], [14]. This method enriches the agent's understanding of environmental states, enhancing learning efficacy.

Despite these advancements, the scalability of deep RL to high-dimensional settings remains a formidable challenge. These environments often require vast parameter spaces and extensive training data, which not only escalate computational demands [15] but also impact the models' performance and generalizability [16]. In response, Quantum RL emerges as a groundbreaking approach, offering substantial reductions in parameter complexity while maintaining robust performance across varied scenarios [17]–[19]. This novel approach holds the potential to significantly alleviate the computational burdens of traditional RL systems, paving the way for more scalable and effective solutions in artificial intelligence. Computer simulations have demonstrated that Quantum RL outperforms classical RL in large search spaces, faster learning, and better exploration-exploitation trade-offs [20]. Recent advances in Quantum RL algorithms using variational quantum circuits (VQC) have demonstrated improved model performance

979-8-3315-4137-8/24/$31.00 ©2024 IEEE
DOI 10.1109/QCE60285.2024.00185

through reduced parameter space complexity [17], enhanced reward representation [21], and stability across hyperparameter variations [22].

Nonetheless, the predominance of current research validating quantum algorithms' superiority is confined to simulated, controlled environments. This prevailing focus on simulation raises concerns about the transferability of these results to practical, real-world applications. Establishing a definitive *quantum advantage* necessitates a critical examination of how specific features of quantum circuit design impact the effectiveness of these models. Recent research has underscored the need for a balanced approach to designing quantum circuits, for instance, their expressibility [23]. Although enhancing expressibility of the quantum circuits is commonly pursued to improve performance, it can paradoxically lead to more complex training dynamics and potentially detrimental outcomes. Therefore, a rigorous analysis of the relationship between the architectural intricacies of quantum circuits and their performance is essential.

This paper provides a comprehensive analysis aimed at optimizing Quantum RL systems through a detailed examination of these dynamics. We expand on initial integrations of neurocognitive principles within RL frameworks and assess the transformative potential of quantum computing to surmount extant constraints. Additionally, while preliminary studies have explored hybrid quantum RL strategies combining classical deep RL techniques—like experience replay and double Q-learning—in simulated settings [17], [21], [24], their translation to models reflecting human decision-making remains speculative. To bridge this gap, we introduce Quantum Rainbow, a novel approach merging Rainbow DQN [25], a robust deep RL algorithm, with VQC-based Q-learning [22]. This investigation not only delves into the relationships between quantum circuit architecture and performance across simulated and human behavioral tasks but also seeks to validate the practical utility of Quantum RL in real-world applications. Our second aim is to demystify the relationships between a variety of aspects of quantum circuit architecture (e.g., quantum entanglement, expressibility) and model performance, providing insights into optimal quantum circuit configurations that could realistically enhance computational tasks in everyday applications. Through this exploration, we strive to move beyond theoretical advantages, addressing the practical challenges and opportunities that quantum computing presents.

II. Background

In RL, an agent learns to optimize her/his action to maximize the rewards provided by the given environment. This interaction between the agent and the environment is modeled as a Markov decision process $\langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma \rangle$, with $\mathcal{S}$ and $\mathcal{A}$ being finite sets of states and actions, $\mathcal{R}$ the reward function, $\mathcal{P}$ the transition function, and $\gamma \in [0, 1)$ the discount factor. The agent's value function Q^π of a policy π denote the expected reward from taking the action $a \in \mathcal{A}$ from the state $s \in \mathcal{S}$:

$$Q^\pi(s,a) := \mathbb{E}\Big[G^\pi(s,a)\Big] = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)\right], \quad (1)$$
$$s_t \sim P(s_{t-1}, a_{t-1}), \quad a_t \sim \pi(s_t)$$

where $G^\pi(s,a)$ is the discounted sum of future rewards. For computational reasons, the value function is typically derived by Bellman's equation [26]:

$$Q^\pi(s,a) = \mathbb{E}\Big[\mathcal{R}(s,a)\Big] + \gamma\, \mathbb{E}_{\mathcal{P},\pi}\Big[Q^\pi(s',a')\Big] \quad (2)$$

The agent's goal is to find the optimal policy π^* which maximizes the expected discounted rewards (i.e., $\mathbb{E}_{a\sim\pi^*} Q^*(s,a) = \max_a Q^*(s,a)$). One of the most popular methods to find this optimal policy in a RL task is Q-learning. In Q-learning, the agent maintains an action-value function:

$$Q_\pi(s,a) := \mathbb{E}_\pi\Big[G_t | s_t = s, a_t = a\Big] \quad (3)$$

This function is updated with observations made in the environment using the temporal differences method:

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha \Big(r_{t+1} + \gamma \max_a Q(s_{t+1},a) - Q(s_t,a_t)\Big) \quad (4)$$

where α is the learning rate and r_{t+1} is the reward at time $t+1$.

In Deep Q-Networks (DQN), the Q-function is represented with a deep neural network [12]. DQN algorithms use neural networks to approximate the action values for a given state. That is, it uses the following loss function for the Q-learning updates:

$$L(\theta) = \mathbb{E}_{(s,a,r,s')\sim U(D)} \Big(r + \gamma \max_{a'} Q_{\bar\theta}(s',a') - Q_\theta(s,a)\Big)^2 \quad (5)$$

where $U(D)$ is a replay buffer storing experienced transitions (s,a,r,s'), and θ and $\bar\theta$ are the parameters of the online network and a target network, respectively.

A. Rainbow DQN

Rainbow DQN combines six algorithmic improvements of the DQN: double Q-learning, prioritized experience replay, dueling networks, multi-step learning, distributional RL, and noisy nets [25].

1) Double Q-Learning

Double Q-learning uses two sets of network weights: the online network to select actions and the target network to estimate the corresponding Q-values. The agent minimizes the loss:

$$\Big(R_{t+1} + \gamma_{t+1}\, q_{\bar\theta}(S_{t+1}, arg\max_{a'} q_\theta(S_{t+1},a')) - q_\theta(S_t,A_t)\Big)^2$$

By separating the action selection and Q-value estimation processes, double Q-learning overcomes the overestimation bias of conventional Q-learning algorithms based on a single network [27], [28].

2) Prioritized Experience Replay

Prioritized experience replay assigns priorities to each experience based on their temporal difference error, indicating the importance of each experience for learning. During training, experiences with higher priorities are sampled more frequently from the replay buffer, allowing the agent to focus on important transitions and learn more efficiently [29].

3) Dueling Networks

Dueling networks decouple the estimation of state value and action advantages. The state value function represents the value of being in a particular state, while the action advantage function represents the advantage of taking each action in that state [30]. By separating these two functions, this neural network architecture can learn the value of each state independently of the chosen action, leading to more stable and efficient learning.

4) Multi-step Learning

Multi-step learning extends this idea by incorporating multiple consecutive steps into the computation of temporal difference errors for updating the Q-values [31]. The multi-step DQN minimizes the following loss:

$$(R_t^{(n)} + \gamma_t^{(n)} \max_{a'} q_{\bar{\theta}}(S_{t+n}, a') - q_\theta(S_t, A_t))^2$$

By considering longer sequences of transitions, the agent can capture more information and improve sample efficiency. It reduces the variance of the updates and helps propagate rewards over longer time horizons.

5) Distributional RL

Rainbow DQN incorporates distributional RL such as the C51 algorithm, where the neural network estimates the entire distribution of returns using distributional Bellman equation [13]. By estimating the reward distribution instead of a single scalar Q-value, this distributional perspective provides a richer representation of the value function and enables the agent to handle environments with stochastic rewards more effectively.

6) Noisy Nets

Noisy nets introduce random noise to the network's weights, allowing the agent to explore in a more targeted manner. By using a factorized noise parameterization, noisy nets can learn to adaptively explore different regions of the state-action space [32].

B. VQCs in RL

VQCs, also known as parametrized quantum circuits, consist of qubits and a series of quantum gates, comprising three essential components: an encoding circuit, a parameterized circuit, and a readout circuit [33]. The encoding circuit is responsible for converting classical data into quantum data, effectively creating parameter-free quantum circuits. Next, the parameterized circuit manipulates the quantum data to generate an approximation of the desired state. Lastly, a readout measurement is performed, typically utilizing one of the Pauli operators (X, Y, Z) to extract relevant information from the circuit. This framework enables the utilization of quantum circuits in various applications, offering a powerful tool for quantum information processing and quantum machine learning research.

VQC represents a prominent form of quantum neural networks, akin to the *quantum* version of deep neural networks. In recent years, there have been research efforts applying these VQC-based models as neural networks in Q-learning. However, these VQC-based Q-learning algorithms and other DQN algorithms have been studied separately [17], [18], [21], [24], and attempts to integrate them have been limited [21], [24], [34].

In the quantum RL model developed by Skolik et al. [22] (Fig. 1), the process begins with the encoding circuit applying R_x rotations to the classical inputs (i.e., environment's state space). These encoded features are then channeled into the variational ansatz of the parametrized circuit. Within this circuit, each layer comprises R_y and R_z rotations and a connected chain of CZ gates. Classical data is reintroduced at the start of each new layer through data reuploading. The readout circuit subsequently computes the expected values of each action, using these values as inputs from the parametrized circuit. These expected values are obtained by measuring Pauli operators.

Importantly, Skolik et al. [22] incorporate trainable weights at both the input and output stages of the VQC. At the encoding layer, trainable weights are applied to the state inputs from the environment, and at the readout stage, weights modify the expected values derived from the computations. The final Q-value of a state s and action a is expressed as follows:

$$Q(s,a) = \langle 0^{\otimes n_Q} | \, U_\theta(s)^\dagger \, O_a \, U_\theta(s) \, | 0^{\otimes n_Q} \rangle \cdot \omega_{o_a} \qquad (6)$$

where O_a is an observable, n_Q the number of qubits, $U_\theta(s)$ the quantum neural network of state s parametrized by θ, and ω_{o_a} the trainable weight for action a.

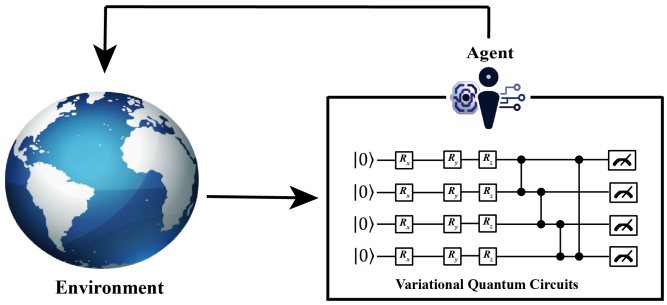

Fig. 1. Architecture of Variational Quantum Circuits in Skolik et al. [22]. VQCs comprise three parts: encoding circuits, parametrized circuits, and readout circuits. Initially, environmental states are encoded into the VQC using parameterized X-axis rotations (encoding circuits). Subsequent rotations along the Y and Z axes are executed using CZ gates (parametrized circuits). The expected values for each action are computed by the readout circuits.

Experimental evaluations conducted on OpenAI Gymnasium environments demonstrated that the VQC-based Q-learning model exhibits quantum advantage compared to other classical DQN models [22].

C. Metrics of VQC architecture

We considered quantum metrics of VQCs in terms of quantum entanglement, expressibility, and effective dimension.

1) Quantum Entanglement

a) Logarithmic Negativity

The logarithmic negativity of a bipartite state ρ is defined as

$$E_N(\rho) := \log_2 ||\rho^{T_A}||_1. \tag{7}$$

where ρ^{T_A} is the partial transpose of the density matrix with respect to the subsystems A. Despite the fact that logarithmic negativity is not a convex function, it has been proven that it does not increase under local operations and classical communication [35], [36]. Since entanglement manipulation protocols in quantum communication and computation often involve such operations, logarithmic negativity reflects the operational reality of quantum systems. Additionally, a quantum state with higher logarithmic negativity is guaranteed to remain more entangled under any local operations and classical communication operation, which makes it a useful tool for comparing the degree of entanglement between different quantum states. Logarithmic negativity also provides the upper bound for the amount of entanglement distillation contained in quantum state ρ, i.e., a process where less pure entangled states are transformed into more pure (maximally entangled) states, which can be used for tasks like quantum teleportation or dense coding [35], [36]. A recent study has shown that the estimation of logarithmic negativity can be applied to quantifying quantum entanglement in variational quantum algorithms on near-term quantum devices such as noisy intermediate-scale quantum (NISQ) devices [37].

b) Coherent Information

The coherent information for a bipartite quantum state ρ is defined as

$$I_C(\rho) = S(\rho_A) - S(\rho), \tag{8}$$

where $S(\rho)$ is the von Neumann entropy of ρ and ρ_A is the subsystem A. Coherent information, which measures the capacity of quantum information conveyed in a noisy quantum channel, cannot be increased by quantum information processing [38], [39]. In terms of quantum entanglement, coherent information holds the following:

$$E_F(\rho) \geq E_D(\rho) \geq I_C(\rho) \tag{9}$$

where $E_F(\rho)$ is the entanglement of formation and $E_D(\rho)$ is the entanglement distillation [40], [41]. $E_F(\rho)$ quantifies the minimum amount of entanglement required to create a given state ρ via local operations and classical communication (LOCC). It indicates the amount of entanglement that can be created from the quantum state ρ. $E_D(\rho)$ is the process of transforming multiple copies of a quantum state ρ into fewer copies of a more entangled state, with a high success probability, using LOCC. High $E_D(\rho)$ indicates the quantum state ρ can be effectively transformed into highly entangled states, making it useful in quantum protocols.

The inequality above shows that coherent information incorporates these aspects, emphasizing the importance of both fundamental entanglement and the practical usability of entanglement in quantum information processing. Thus, coherent information contains valuable properties of quantum entanglement and is also easy to calculate, making it a suitable measure of quantum entanglement in hybrid quantum-classical machine learning framework [42].

c) Entangling Capability

The entangling capacity of a variational quantum circuit quantifies the circuit's proficiency in effectively delineating the solution space of the machine learning task and capturing non-trivial correlations within the quantum dataset [43], [44]. The entangling capability can be obtained by sampling the circuit parameters and calculating the sample average of the Meyer-Wallach measure [45] for the resulting states [46]. More precisely, we take the estimate of the entangling capability to be

$$Ent = \frac{1}{|S|} \sum_{\theta_i \in S} Q(|\psi_{\theta_i}\rangle), \tag{10}$$

where $S = \{\theta_i\}$ is the set of sampled circuit parameter vectors and Q is the Meyer-Wallach measure. This measure with $n-$qubits is defined as

$$Q(|\psi\rangle) = \frac{4}{n} \sum_{j=1}^{n} D(\iota_j(0)|\psi\rangle, \iota_j(1)|\psi\rangle) \tag{11}$$

where D is the generalized distance:

$$D(|u\rangle, |v\rangle) = \frac{1}{2} \sum_{i,j} |u_i v_j - v_i u_j|^2. \tag{12}$$

Here, $\iota_j(b)$ represents the linear mapping which acts on a computational basis (i.e., quantum state) with $b_j \in \{0, 1\}$:

$$\iota_j(b)\,|\,b_1 \cdots b_n\rangle = \delta_{bb_j} |b_1 \cdots b_{j-1} b_{j+1} \cdots b_n\rangle \tag{13}$$

where the qubit b_j is absent and δ denote the Kronecker-Delta operator.

An entangling capability score of 0 indicates that the quantum circuit exclusively generates product states, whereas a score of 1 denote that the circuit consistently produces highly entangled states. It has been used in recent studies as a measure of the variational quantum circuit's ability to generate entangled states [19], [47], [48].

2) Expressibility

Expressibility refers to a quantum circuit's capacity to generate states within the Hilbert space effectively [46]. It can be measured by comparing the states generated by varying the parameters of a variational quantum circuit with the uniform distribution of states, specifically, the ensemble of Haar-random states, renowned for its expressiveness. The use of Haar ensemble properties enables the derivation of an efficient and problem-independent measure of expressibility. Although expressiveness is not obligatory for favorable algorithm performance, this definition allows for the identification of limited VQC structures, such as those generating product states, offering valuable insights into their capabilities.

One first samples two sets of parameters from the variational quantum circuits and derives distribution of fidelities from the

corresponding quantum states $|\psi_i\rangle, |\psi_j\rangle$. By defining F as fidelity and N as the dimension of the Hilbert space [49], the analytic Haar ensemble becomes $P_{Haar} = (N-1)(1-F)^{N-2}$. Finally, the Kullback–Leibler (KL) divergence [50], between the estimated probability distribution of fidelities $\hat{P}_{VQC}(F;\theta)$ and the Haar random states ensemble can be computed to quantify expressibility:

$$Expr = D_{KL}\left(\hat{P}_{VQC}(F;\theta)||P_{Haar}(F)\right). \tag{14}$$

Expressibility, when defined through KL divergence, quantifies the information loss incurred when approximating the distribution of state fidelities produced by a variational quantum circuit with that of Haar random states [46].

It has been used in recent studies as a measure of the variational quantum circuit's ability to learn a target function [19], [47], [48].

3) Effective Dimension

Effective dimension serves as a pertinent metric to estimate the space occupied by a model within the model space, a domain encompassing all conceivable functions pertinent to a specific model class. Here, the Fisher information matrix serves as the key metric for quantifying the range of functions a model can fit [51]. A pivotal determinant in these computations is the number of data observations, which inherently establishes a natural scale or resolution for observing the model space. This approach holds practical significance, especially in scenarios where data is scarce. Additionally, it provides valuable insights into the interplay between data availability and the accurate assessment of model complexity [52].

As described in Abbas et al. [52], the effective dimension of a statistical model $\mathcal{M}_\Theta := \{p(\cdot,\cdot;\theta) : \theta \in \Theta \subset \mathbb{R}^d\}$ with respect to $\gamma \in (0,1]$ is defined as:

$$d_{\gamma,n}(\mathcal{M}_\theta) := 2\frac{\log\left(\frac{1}{V_\Theta}\int_\Theta \sqrt{\det\left(I_d + \frac{\gamma\, n}{2\pi\log n}\hat{F}(\theta)\right)}\, d\theta\right)}{\log(\frac{\gamma\, n}{2\pi\log n})} \tag{15}$$

where $V_\Theta := \int_\Theta d\theta \geq 0$ is the volume of the d-dimensional parameter space of Θ, and $n \in N, n > 1$ is the number of data samples. The normalised Fisher information matrix $\hat{F}_\theta \in \mathbb{R}^{d\times d}$ is defined as:

$$\hat{F}_{ij}(\theta) := d\frac{V_\Theta}{\int_\Theta tr(F(\theta))d\theta}F_{ij}(\theta), \tag{16}$$

where the normalization ensures that $\frac{1}{V_\Theta}\int_\Theta tr(\hat{F}(\theta))d\theta = d$.

It is shown that quantum neural networks (i.e., variational quantum circuits) achieves higher effective dimension than classical neural networks, which indicates that the quantum model has greater capability to perform well on new data than its classical counterpart [52].

A recent study in quantum reinforcement learning [19] has used effective dimension as one of the metrics to characterize the properties of variational quantum circuits.

III. Quantum Rainbow

In brief, our Quantum Rainbow algorithm can be seen as a hybrid model of human decision-making, which implements Q-function approximation in two consecutive processes: first using VQC and then using the Rainbow DQN (Fig. 2). Using the VQC-based Q-learning with data reuploading proposed by Skolik et al. [22], the agent first draws quantum representations of possible states and actions within the environment. Next, the agent utilizes the six improvements of the Rainbow algorithm (i.e., double Q-learning, dueling networks, prioritized experience replay, multi-step learning, distributional RL, noisy nets) to estimate the reward distribution and choose the optimal policy.

The proposed VQC part serves to encode classical inputs from the environment. These inputs undergo processing within parametrized circuits that incorporate data reuploading. Initial Q-values are obtained for each action, employing Equation 6 as the basis for computation. These initial Q-values are then fed into both the value stream and the advantage stream of the Rainbow DQN architecture.

In the classical Rainbow layers, the value and advantage streams' outputs are combined for each quantile i of the reward distribution. Noisy linear layers with factorized Gaussian noise are employed within each stream. Subsequently, a softmax layer is applied to estimate the normalized reward distributions using the following equation:

$$p^i_\theta(q(s,a),a) = \frac{\exp(v^i_\eta(q) + a^i_\psi(q,a) + \bar{a}^i_\psi(q))}{\sum_j \exp(v^j_\eta(q) + a^j_\psi(q,a) + \bar{a}^j_\psi(q))} \tag{17}$$

Here, $q = q(s,a)$ represents the initial Q-values of state s and action a obtained from the VQC part, v^i_η and a^i_ψ the value stream and advantage stream for quantile i, respectively, and $\bar{a}^i_\psi(q) = \frac{1}{N_{\text{actions}}}\sum_{a'} a^i_\psi(q,a')$. The update process incorporates multi-step learning and prioritized experience replay buffer. Both the VQC and the Rainbow DQN are used as target networks for double Q-learning. Notably, we used quantile regression DQN (QR-DQN) instead of C51 for distributional RL of the Rainbow part. We did this because of QR-DQN's ability to estimate the return quantile values for N fixed, uniform probabilities, which enables more accurate distributional estimation, particularly in scenarios with infrequent and episodic rewards [15].

IV. Experiment

To evaluate the Quantum Rainbow, we utilized two human behavioral task environments: Iowa Gambling Task and 4-Armed Bandit Task. The rationale behind the selection of these environments was threefold: first, to investigate the algorithm's ability to learn based on human decision-making behavior; second, to study the potential real-world applications of the algorithm by training on more complex decision-making tasks than computer-simulated RL environments; and third, to ensure that the environments could be executed efficiently using a small number of qubits, making them well-suited for VQC-based Q-learning [22].

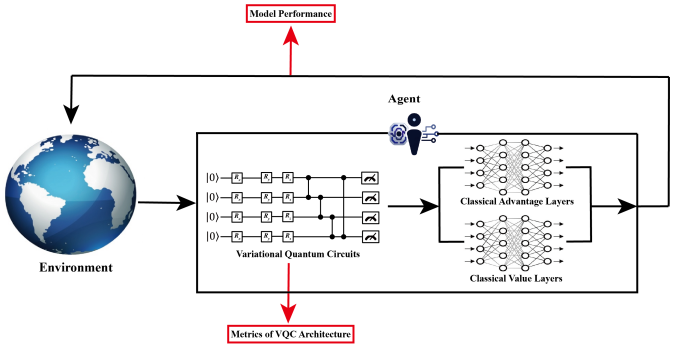

Fig. 2. Hybrid architecture of the Quantum Rainbow. The Quantum Rainbow integrates a VQC with classical Rainbow DQN layers. Initially, environmental states are processed through the VQC, passing through encoding circuits, parametrized circuits, and readout circuits. The outputs from the VQC are then input into the classical Rainbow DQN layers, which consist of value and advantage layers. These classical layers incorporate techniques such as double Q-learning, prioritized experience replay, dueling networks, multi-step learning, distributional RL, and noisy nets. This combination processes the quantum outputs to calculate the final Q-values, enabling the Quantum Rainbow agent to determine the optimal action based on these values. Throughout the training process, metrics of VQC architecture and the agent's performance are computed.

The experiments were conducted using the Pennylane and Tianshou libraries. Pennylane [53] stands out as a pivotal library for quantum machine learning research, primarily due to its comprehensive quantum simulator that enables seamless implementation of quantum circuits using conventional CPU resources to emulate quantum processing unit operations. Its compatibility with PyTorch, supporting tensors and gradient operations, offers a bridge between classical and quantum computing realms [54]. This is particularly useful for researchers to build hybrid classical-quantum models, such as our Quantum Rainbow. Tianshou is a Pytorch-based RL platform which provides efficient, adaptable, and reliable infrastructure for the implementation of cutting-edge deep RL algorithms, including Rainbow DQN [55].

Each algorithm was trained on a single NIVIDA GeForce RTX 3090 GPU for 50,000 iterations. The Quantum Rainbow has a combinatorial space of hyperparameters too large to have an exhaustive search and tuning, just like the original classical Rainbow DQN [25]. Thus, we used the set of hyperparameters that prior studies have reported as optimal [15], [22], [25]. We present the results from the optimal sets of hyperparameters that we tested from TABLE I. Note that the Quantum Rainbow algorithm demonstrated its versatility by not requiring access to quantum computing devices like NISQ devices.

A. Environments

1) Iowa Gambling Task

We employed an online data pool consisting of 617 healthy participants who performed the Iowa Gambling Task (IGT) across 10 studies [56]. Following a previous study [57], we specifically focused on subjects who completed 100 trials, resulting in a final sample size of N=504. The remaining 113 subjects had performed the Iowa Gambling Task with a different number of trials (either 95 or 150 trials). As the group of 504 subjects was considered sufficient to represent the entire data pool, we concentrated our analysis on this subset.

In the Iowa Gambling Task, participants were presented with four decks of cards, denoted as decks A, B, C, and D. Each card displayed potential gain or loss points. The net points, calculated as the difference between the gain and loss, were determined for each trial. Selecting deck A or deck B for ten choices resulted in a net loss of -250 points on average. Conversely, selecting deck C or deck D for ten choices led to a net gain of 250 points on average. Decks A and B were categorized as disadvantageous decks due to their overall loss, while decks C and D were regarded as advantageous decks owing to their overall gain.

To assess the reward for each participant, we calculated the proportion of good (deck C, D) minus bad deck (deck A, B) selections, in accordance with prior research [57]. This measure allowed us to evaluate the participants' decision-making strategies in the Iowa Gambling Task.

2) 4-Armed Bandit Task

A publicly available dataset consisting of data from 965 human participants who engaged in a drifting 4-Armed Bandit (4AB) task was utilized [58]. Each participant selected one out of four bandits during each trial, and received a continuous numerical reward, ranging from 1 to 98 points, based on the chosen bandit's current reward payout. The reward payouts were subject to drift over time, following a Gaussian walk pattern. To introduce variability in the experimental conditions, the participants were randomly assigned to one of three predetermined reward payoff schedules.

Originally designed to investigate the neural mechanisms underlying exploratory and exploitative choice [59], the task demanded rapid decision-making from the participants, as they were given only four seconds to reach a decision in each trial. In instances where the participants failed to make a timely decision, they were automatically moved to the next trial, resulting in no reward for that specific trial. Out of the 965 participants, only 127 successfully completed all 150 rounds, while the remaining participants missed at least one trial during the task. The average number of rounds completed per participant was approximately 145.

B. Causal Discovery

To systematically investigate the associations among quantum entanglement, expressibility, effective dimension, and the performance of Quantum Rainbow algorithm, we first collected quantum circuit metrics and corresponding model performance data at every 100 iterations throughout the training process. To rigorously analyze the temporal relationships and causal influences among the quantum circuit parameters and model performance, we utilized three advanced time series causal discovery algorithms: Peter-Clark Momentary Conditional Independence (PCMCI), PCMCI+, and the Vector Autoregression Linear Non-Gaussian Acyclic Model (VAR-LiNGAM).

TABLE I
HYPERPARAMETERS USED FOR TRAINING QUANTUM RAINBOW

Hyperparameters	Description	Values
ϵ_{train}	Epsilon value for ϵ-greedy policy during training	0.05
ϵ_{test}	Epsilon value for ϵ-greedy policy during test	0.1
buffer-size	Buffer size for prioritized experience replay	20000
lr	Learning rate	0.001
γ	Discount rate	0.99
q	Number of quantiles for the value distribution	51
v_{min}	Value of the smallest quantile in the support set	-10
v_{max}	Value of the largest quantile in the support set	10
n_{step}	Number of steps to look ahead	5
target-update-freq	Target network update frequency	100
epoch	Maximum epochs for training	10
step-per-epoch	Number of environment steps collected per epoch	10000
step-per-collect	Number of transitions the collector would collect before the network update	10
update-per-step	Number of times the policy network would be updated per transition after (step-per-collect) transitions are collected	0.1
α	Prioritization exponent	0.5
β	Importance sample soft coefficient	0.4
batch-size	Batch size of sample data	64
n	Number of neurons in the noisy nets	128
l_V	Number of hidden value layers of the Rainbow part	1
l_A	Number of hidden advantage layers of the Rainbow part	1
l_Q	Number of quantum layers in the parametrized circuits	3
n_Q	Number of qubits	4
optimizer	Optimizer for updating network weights	Adam

1) PCMCI and PCMCI+

PCMCI [60] and its advanced variant PCMCI+ [61] represent pivotal developments in causal discovery for time series analysis, addressing the inherent complexities of temporal data. PCMCI utilizes conditional independence tests within a time series framework to elucidate causal structures, effectively handling autocorrelation and latent confounders. It leverages a two-phase process, first identifying conditional independencies through momentary information criteria, and then ascertaining causal directions using a causal discovery algorithm. PCMCI+ extends this methodology by improving the distinction between direct and indirect causation, thereby enhancing the fidelity of causal inference, especially in systems with intricate temporal dynamics. This innovation in causal analysis is instrumental in unraveling the nuanced interplay of variables over time, offering a robust tool for researchers across various disciplines to dissect and understand the causal mechanisms underpinning complex dynamical systems.

2) VARLiNGAM

The VARLiNGAM algorithm [62] represents a significant advancement in causal discovery from time series data, leveraging the strengths of vector autoregression (VAR) and Linear Non-Gaussian Acyclic Model (LiNGAM) methodologies [63]. This algorithm uniquely discerns causal relationships amidst the temporal interdependencies intrinsic to time series data. It operates on the premise that the data's temporal dynamics can be captured through VAR models, which express each variable as a linear function of its historical values. Concurrently, the LiNGAM component of VARLiNGAM capitalizes on the non-Gaussian distribution of the data to unravel the acyclic causal structure, thereby facilitating a more nuanced understanding of causality beyond mere correlation. This synthesis not only enhances the accuracy of causal inference in multivariate time series but also broadens the applicability of causal analysis in domains where temporal relationships are pivotal, offering a robust framework for disentangling the complex interplay of cause and effect over time.

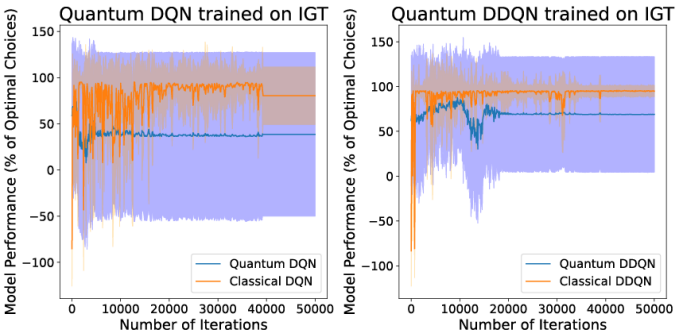

Fig. 3. Exploratory analysis with Quantum DQN and Quantum DDQN. The shaded areas indicate 95% confidence intervals. When trained on Iowa Gambling Task (IGT) environment, Quantum DQN and Quantum DDQN models underperformed compared to their classical counterparts and failed to converge to optimal solutions.

V. RESULTS

To validate the effectiveness of deep RL techniques such as prioritized experience replay, multi-step learning, and distributional RL within quantum contexts, we conducted an exploratory analysis using the Iowa Gambling Task. Our tests with Quantum DQN, based solely on the VQC structure [22], and Quantum DDQN, which integrates double Q-learning with VQC [21], [24], showed that both models struggled to converge to optimal solutions after 50,000 iterations, performing significantly worse than their classical counterparts (Fig. 3).

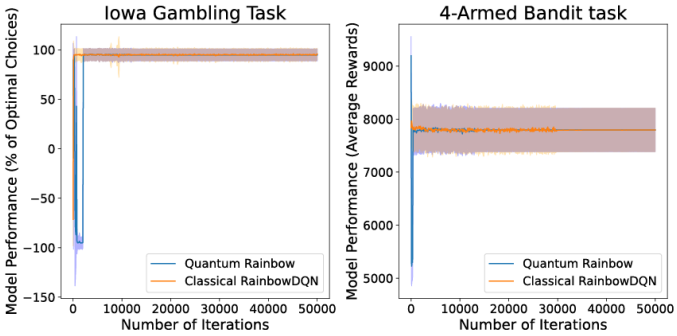

Fig. 4. Model Performance of the Quantum Rainbow trained on Iowa Gambling Task and 4-Armed Bandit tasks.The shaded areas indicate 95% confidence intervals. The Quantum Rainbow, integrating brain-inspired deep RL techniques into VQCs, matched the classical Rainbow DQN in achieving optimal solutions.

Despite previous demonstrations of these quantum algorithms' efficacy in computer-simulated environments [17], [21], [22], our findings highlight their limitations in tasks that model human behavior. In response, we explored the integration of deep RL techniques that align closely with human neurocognitive functions into VQCs. Our implementation of the Quantum Rainbow algorithm, which combines these techniques, demonstrated comparable performance with the classical Rainbow DQN, successfully converging to optimal solutions in both the Iowa Gambling and 4-Armed Bandit tasks—more complex decision-making tasks, modeled after human cognitive processes, than computer-simulated RL environments (Fig. 4).

These findings underscore the potential of incorporating deep RL techniques, particularly those that mirror human neurocognitive functions, into VQCs to enhance the adaptability and effectiveness of RL algorithms in complex, real-world settings.

The Quantum Rainbow model exhibits a marked reduction in parameter count compared to its classical counterpart, the Rainbow DQN. The Quantum Rainbow consists of three quantum layers—encoding circuits, parametrized circuits, and readout circuits—and two classical layers—value and advantage layers. In the parametrized circuits, each layer employs R_y and R_z rotations for each qubit, resulting in a total of $2n_Q l_Q$ quantum parameters, where n_Q represents the number of qubits and l_Q the number of quantum layers. Additionally, trainable weights are introduced at the input and output stages, thereby incorporating classical parameters into the quantum layers. Specifically, the encoding circuits include parameters proportional to the environmental state space, and the readout circuits include parameters corresponding to the number of possible actions for initial Q-value estimation. Assuming the number of qubits equals the state space, the classical parameters within the quantum layers amount to $s + a = n_Q + a$, where s and a denote the state and action space, respectively.

The classical layers of the model consist of:

- Value Layer: $(a+1)n + n(n+1)l_V + (n+1)q$
- Advantage Layer: $(a+1)n + n(n+1)l_A + (n+1)aq$

Here, n represents the number of neurons, l_V and l_A the number of hidden layers in the value and advantage layers, respectively, a the action space, and q the number of quantiles in the Rainbow DQN. The cumulative classical parameter count in these layers is $2n(a+1)+n(n+1)(l_V+l_A)+(n+1)(a+1)q$.

The total number of parameters in the Quantum Rainbow can be summarized as follows:

$$2n_Q l_Q + s + a + 2n(a+1) + n(n+1)(l_V + l_A) + (n+1)(a+1)q$$

In contrast, the classical Rainbow DQN consists of three layers: feature, value, and advantage layers. The parameters are distributed as follows:

- Feature Layer: $sn + n(n+1)l_F$
- Value Layer: $n(n+1)l_V + (n+1)q$
- Advantage Layer: $n(n+1)l_A + (n+1)aq$,

with l_F representing the number of hidden layers in the feature layers.

Consequently, the total number of parameters for the classical Rainbow DQN is:

$$sn + 2n(n+1) + n(n+1)(l_F + l_V + l_A) + (n+1)(a+1)q. \quad (18)$$

Despite using the same hyperparameters as detailed in TABLE I, the Quantum Rainbow utilizes only 67,231 parameters (24 quantum parameters, 67,207 classical parameters), while the classical Rainbow DQN uses 148,991 parameters. This demonstrates that the Quantum Rainbow achieves comparable performance on tasks such as Iowa Gambling Task, and the 4-Armed Bandit task, with fewer than half the parameters required by the classical model.

Further analysis was conducted to elucidate the associations between the metrics of VQC architecture and the performance of the Quantum Rainbow agent. To ensure the accuracy of our causal inferences, we employed the Akaike Information Criterion (AIC) to evaluate the fit of each causal discovery algorithm across different RL environments. The AIC evaluations, as detailed in TABLE II, indicated that the VARLiNGAM consistently provided the best model fit.

Based on the optimal model fit provided by VARLiNGAM, we further analyzed the causal relationships between the quantum circuit metrics and the Quantum Rainbow's performance. To maintain conciseness and clarity, we present a single causal diagram that effectively summarizes the relationships derived from both Iowa Gambling Task and 4-Armed Bandit task environments. This diagram reflects the robustness of VARLiNGAM in capturing the essential dynamics between the VQC's architecture and the Quantum Rainbow's performance,

TABLE II
MODEL FIT OF CAUSAL DISCOVERY ANALYSES

Algorithm	IGT	4AB
PCMCI	2.440	247.705
PCMCI+	1.628	41.198
VARLiNGAM	-10.353	0.326

thereby highlighting the algorithm's potential in optimizing quantum-enhanced RL applications.

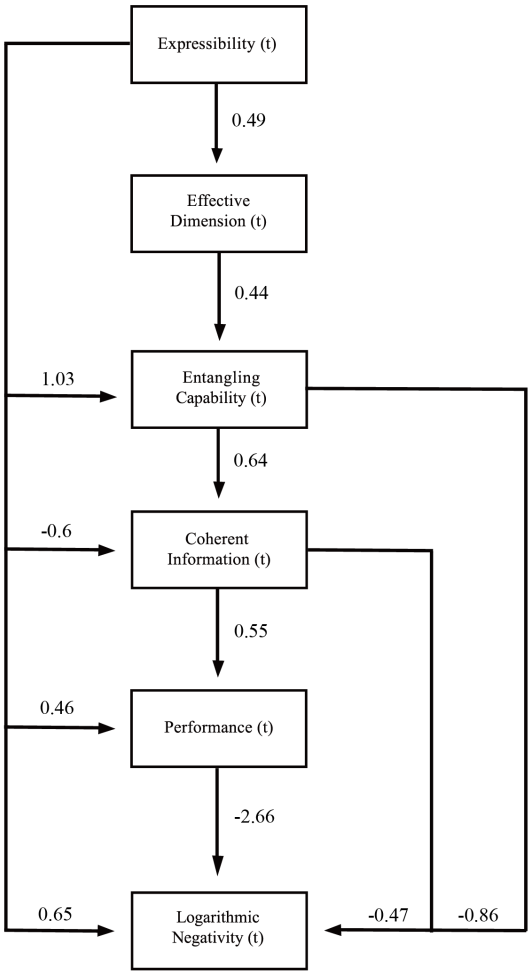

Fig. 5. Causal Diagram Illustrating the Relationship between VQC Architecture Metrics and Quantum Rainbow Model Performance

The results presented in Fig. 5 reveals the intricate causal relationships between key quantum circuit metrics and their collective impact on the performance of the Quantum Rainbow model. Notably, coherent information directly contributes to model performance with a substantial positive standardized coefficient ($\beta = 0.55$), indicating that the preservation of quantum information is a crucial factor for the model's success. Expressibility also shows a positive direct causal impact ($\beta = 0.46$), suggesting that the model's ability to represent a diverse set of quantum states is beneficial for its performance.

Furthermore, the analysis indicates that effective dimension and entangling capability positively influence model performance, albeit indirectly. These findings support the notion that the ability of the Quantum Rainbow to utilize entanglement and maintain a high level of system expressibility is advantageous to its computational efficacy.

Conversely, an inverse relationship between model performance and logarithmic negativity is observed, with a robust negative standardized coefficient ($\beta = -2.66$). This indicates that improved model performance correlates with a decreased reliance on bipartite entanglement, as measured by logarithmic negativity, at that specific point in time.

These insights highlight the complexity of the quantum circuit metrics' interrelations and their temporal evolution in affecting the Quantum Rainbow's computational power and efficiency. These results enrich our comprehension of the dynamic influences that quantum properties exert on model performance and offer strategic direction for refining quantum circuits. This advancement is crucial for optimizing the performance of quantum machine learning models, thus propelling the development of more sophisticated quantum computing applications.

VI. Conclusion

In this paper, we examined the performance of the Quantum Rainbow, a novel hybrid quantum-classical algorithm, across a range of learning environments. Our analysis revealed that Quantum Rainbow excels particularly in human behavioral task environments. This enhanced performance in complex cognitive tasks emphasizes the utility and adaptability of integrating deep RL techniques with VQCs, especially those that emulate human neurocognitive functions, including prioritized experience replay, multi-step learning, and distributional RL.

The Quantum Rainbow model has demonstrated remarkable efficiency by achieving comparable performance to its classical counterpart, the Rainbow DQN, with only 45.13% of the parameters. This reduction in parameter count not only simplifies the model but significantly enhances memory efficiency. This aspect of quantum computing, often referred to as quantum supremacy in memory consumption, was previously observed in a study by Chen et al. [17], where a VQC-DQN displayed reduced parameter complexity compared to traditional classical DQNs in cognitive-radio environments. Our findings extend these advantages to more diverse applications, including behavioral tasks like the Iowa Gambling Task and the 4-Armed Bandit task, suggesting that the Quantum Rainbow could efficiently handle more complex, high-dimensional state spaces that were not analyzed in this study.

Additionally, the streamlined parameterization of the Quantum Rainbow mitigates the risk of overfitting, enhancing the model's generalizability—a critical challenge in large deep learning models, which are prone to overfitting, especially as the sample size increases. Contrarily, prior research [64], [65] indicates that quantum algorithms can maintain robust out-of-distribution performance even with limited parameters and sample sizes. Our results corroborate these findings, showcasing that the Quantum Rainbow not only matches but potentially exceeds the performance of classical DQN algorithms under constrained conditions.

Further, our study delved into the causal relationships between quantum circuit architecture and the performance of the Quantum Rainbow, utilizing causal discovery algorithms. We determined that quantum entanglement metrics—specifically logarithmic negativity, coherent information, and entangling

capability—play significant roles in influencing model performance, albeit in varying manners.

The results of our time-series causal discovery analysis have elucidated pivotal aspects of VQC performance within the Quantum Rainbow model. Coherent information, serving as a measure of quantum information preservation, was found to have a direct positive effect on the model's performance. This suggests that the ability of the VQC to maintain the integrity of quantum information through a noisy quantum channel is paramount to its overall effectiveness.

The entangling capability of the VQC, indicative of the circuit's ability to generate quantum states that exhibit quantum correlations, was observed to have an indirect positive influence on model performance. This underscores the significance of entanglement as an essential resource for quantum computation, contributing to the model's computational power [46].

The negative correlation observed between model performance and logarithmic negativity presents an intriguing facet of the Quantum Rainbow model's operational dynamics. While logarithmic negativity is traditionally utilized to quantify the degree of entanglement in bipartite quantum states, it may not accurately capture multipartite entanglement characteristics [66]. This discrepancy, in the context of high coherent information and high entangling capabilities, hints at the VQC's generation of a more nuanced form of entanglement, potentially of the multipartite variety. Multipartite entanglement, which is known to facilitate complex quantum state manipulations, stands as a cornerstone for the enhanced computational capabilities of quantum algorithms [67], [68].

The significance of multipartite entanglement extends beyond the bipartite scenarios, particularly in quantum technologies, where it serves as a vital resource for executing advanced computational tasks. This has been supported by literature that underscores its indispensable role in the development of quantum computational technologies (e.g., Grover's algorithm) [66]–[70]. The findings of our study, aligned with this body of research, suggest that the computational efficacy of the Quantum Rainbow model may be bolstered by leveraging multipartite entanglement. This insight not only reaffirms the necessity of a multipartite perspective in the assessment of quantum models but also sets the stage for future explorations into optimizing VQC designs to exploit this complex entanglement form, paving the way towards realizing the full promise of quantum computing.

Moreover, expressibility, which reflects the VQC's ability to generate a diverse set of quantum states, directly correlates with enhanced model performance, affirming the importance of versatile state preparation in quantum machine learning. Likewise, the effective dimension, indicative of the circuit's capacity to exploit a higher-dimensional Hilbert space, indirectly fosters superior performance, underscoring the value of leveraging the full potential of quantum systems.

Our findings demonstrate that the Quantum Rainbow model thrives on maintaining coherent quantum states and exploiting the complex entanglement structures that emerge within the VQC. The inverse relationship between performance and logarithmic negativity prompts a re-evaluation of entanglement measures in the context of quantum machine learning, suggesting that alternative forms of entanglement may be leveraged to enhance computational abilities. These findings pave the way for further research into the optimization of VQCs for quantum machine learning, potentially leading to more efficient and powerful quantum algorithms. Ultimately, this study not only confirms the practical viability of the Quantum Rainbow in advanced cognitive tasks but also paves the way for future investigations into optimizing quantum circuit design for enhanced computational performance.

ACKNOWLEDGMENT

The authors extend their gratitude to the members of the Connectome Lab for their invaluable support and critical contributions to the development of the Quantum Rainbow algorithm. Special thanks are also due to Seoyeon Park for her exceptional work on the sketches of Fig. 5. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1C1C1006503, RS-2023-00266787, RS-2023-00265406, RS-2024-00421268), by Creative-Pioneering Researchers Program through Seoul National University (No. 200-20230058), by Semi-Supervised Learning Research Grant by SAMSUNG(No.A0426-20220118), by Identify the network of brain preparation steps for concentration Research Grant by LooxidLabs (No.339-20230001), by Institute of Information and communications Technology Planning and Evaluation (IITP) grant funded by the Korea government (MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)] and by the National Supercomputing Center with supercomputing resources including technical support (KSC-2023-CRE-0568) and by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2021S1A3A2A02090597).

REFERENCES

[1] E. O. Neftci and B. B. Averbeck, "Reinforcement learning in artificial and biological systems," *Nature Machine Intelligence*, vol. 1, no. 3, pp. 133–143, 2019.

[2] Y. Niv, "Reinforcement learning in the brain," *Journal of Mathematical Psychology*, vol. 53, no. 3, pp. 139–154, 2009.

[3] C. Foo, A. Lozada, J. Aljadeff, Y. Li, J. W. Wang, P. A. Slesinger, and D. Kleinfeld, "Reinforcement learning links spontaneous cortical dopamine impulses to reward," *Current Biology*, 2021.

[4] P. W. Glimcher, "Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis," *Proceedings of the National Academy of Sciences*, vol. 108, pp. 15647–15654, 2011.

[5] M. Watabe-Uchida, N. Eshel, and N. Uchida, "Neural circuitry of reward prediction error," *Annual Review of Neuroscience*, vol. 40, no. 1, pp. 373–394, 2017.

[6] M. M. Botvinick, Y. Niv, and A. G. Barto, "Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective," *Cognition*, vol. 113, no. 3, pp. 262–280, 2009.

[7] K. Juechems and C. Summerfield, "Where does value come from?," *Trends in Cognitive Sciences*, vol. 23, no. 10, pp. 836–850, 2019.

[8] A. Zador, S. Escola, B. Richards, B. Ölveczky, Y. Bengio, K. Boahen, M. Botvinick, D. Chklovskii, A. Churchland, C. Clopath, J. DiCarlo, S. Ganguli, J. Hawkins, K. Körding, A. Koulakov, Y. LeCun, T. Lillicrap, A. Marblestone, B. Olshausen, A. Pouget, C. Savin, T. Sejnowski, E. Simoncelli, S. Solla, D. Sussillo, A. S. Tolias, and D. Tsao, "Catalyzing next-generation artificial intelligence through neuroai," *Nature Communications*, vol. 14, no. 1, p. 1597, 2023.

[9] M. Botvinick, J. X. Wang, W. Dabney, K. J. Miller, and Z. Kurth-Nelson, "Deep reinforcement learning and its neuroscientific implications," *Neuron*, vol. 107, no. 4, pp. 603–616, 2020.

[10] J. Gläscher, N. Daw, P. Dayan, and J. P. O'Doherty, "States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning," *Neuron*, vol. 66, no. 4, pp. 585–595, 2010.

[11] W. Schultz, P. Dayan, and P. R. Montague, "A neural substrate of prediction and reward," *Science*, vol. 275, no. 5306, pp. 1593–1599, 1997.

[12] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[13] M. G. Bellemare, W. Dabney, and R. Munos, "A distributional perspective on reinforcement learning," in *Proceedings of the 34th International Conference on Machine Learning* (D. Precup and Y. W. Teh, eds.), vol. 70 of *Proceedings of Machine Learning Research*, pp. 449–458, PMLR, 06–11 Aug 2017.

[14] W. Dabney, Z. Kurth-Nelson, N. Uchida, C. K. Starkweather, D. Hassabis, R. Munos, and M. Botvinick, "A distributional code for value in dopamine-based reinforcement learning," *Nature*, vol. 577, no. 7792, pp. 671–675, 2020.

[15] J. S. O. Ceron and P. S. Castro, "Revisiting rainbow: Promoting more insightful and inclusive deep reinforcement learning research," in *Proceedings of the 38th International Conference on Machine Learning* (M. Meila and T. Zhang, eds.), vol. 139 of *Proceedings of Machine Learning Research*, pp. 1373–1383, PMLR, 18–24 Jul 2021.

[16] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever, "Deep double descent: where bigger models and more data hurt*," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2021, p. 124003, dec 2021.

[17] S. Y. C. Chen, C. H. H. Yang, J. Qi, P. Y. Chen, X. Ma, and H. S. Goan, "Variational quantum circuits for deep reinforcement learning," *IEEE Access*, vol. 8, pp. 141007–141024, 2020.

[18] S. Jerbi, C. Gyurik, S. Marshall, H. Briegel, and V. Dunjko, "Parametrized quantum policies for reinforcement learning," in *Advances in Neural Information Processing Systems* (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), vol. 34, pp. 28362–28375, Curran Associates, Inc., 2021.

[19] T.-A. Drăgan, M. Monnet, C. B. Mendl, and J. M. Lorenz, "Quantum Reinforcement Learning for Solving a Stochastic Frozen Lake Environment and the Impact of Quantum Architecture Choices," *arXiv e-prints*, p. arXiv:2212.07932, Dec. 2022.

[20] D. Dong, C. Chen, H. Li, and T. J. Tarn, "Quantum reinforcement learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 38, no. 5, pp. 1207–1220, 2008.

[21] O. Lockwood and M. Si, "Reinforcement learning with quantum variational circuit," *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 16, pp. 245–251, Oct. 2020.

[22] A. Skolik, S. Jerbi, and V. Dunjko, "Quantum agents in the gym: a variational quantum algorithm for deep q-learning," *Quantum*, vol. 6, p. 720, 2022.

[23] M. Schuld and N. Killoran, "Is quantum advantage the right goal for quantum machine learning?," *PRX Quantum*, vol. 3, p. 030101, Jul 2022.

[24] O. Lockwood and M. Si, "Playing atari with hybrid quantum-classical reinforcement learning," in *NeurIPS 2020 Workshop on Pre-registration in Machine Learning* (L. Bertinetto, J. F. Henriques, S. Albanie, M. Paganini, and G. Varol, eds.), vol. 148 of *Proceedings of Machine Learning Research*, pp. 285–301, PMLR, 11 Dec 2021.

[25] M. Hessel, J. Modayil, H. van Hasselt, T. Schaul, G. Ostrovski, W. Dabney, D. Horgan, B. Piot, M. Azar, and D. Silver, "Rainbow: Combining improvements in deep reinforcement learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[26] R. E. Bellman, *Dynamic Programming*. Princeton, NJ: Princeton University Press, 1957.

[27] H. van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, 2016.

[28] H. van Hasselt, "Double q-learning," *Advances in Neural Information Processing Systems 23 (NIPS 2010)*, vol. 23, p. 2613–2621, 2010.

[29] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," in *International Conference on Learning Representations*, 2016.

[30] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, and N. Freitas, "Dueling network architectures for deep reinforcement learning," in *Proceedings of The 33rd International Conference on Machine Learning* (M. F. Balcan and K. Q. Weinberger, eds.), vol. 48 of *Proceedings of Machine Learning Research*, (New York, New York, USA), pp. 1995–2003, PMLR, 20–22 Jun 2016.

[31] R. S. Sutton, "Learning to predict by the methods of temporal differences," *Machine Learning*, vol. 3, no. 1, pp. 9–44, 1988.

[32] M. Fortunato, M. G. Azar, B. Piot, J. Menick, I. Osband, A. Graves, V. Mnih, R. Munos, D. Hassabis, O. Pietquin, C. Blundell, and S. Legg, "Noisy networks for exploration," 2019.

[33] M. Benedetti, E. Lloyd, S. Sack, and M. Fiorentini, "Parameterized quantum circuits as machine learning models," *Quantum Science and Technology*, vol. 4, no. 4, p. 043001, 2019.

[34] Z. He, L. Li, S. Zheng, Y. Li, and H. Situ, "Variational quantum compiling with double q-learning," *New Journal of Physics*, vol. 23, no. 3, p. 033002, 2021.

[35] G. Vidal and R. F. Werner, "Computable measure of entanglement," *Physical Review A*, vol. 65, p. 032314, Feb 2002.

[36] M. B. Plenio, "Logarithmic negativity: A full entanglement monotone that is not convex," *Physical Review Letters*, vol. 95, p. 090503, Aug 2005.

[37] K. Wang, Z. Song, X. Zhao, Z. Wang, and X. Wang, "Detecting and quantifying entanglement on near-term quantum devices," *npj Quantum Information*, vol. 8, no. 1, p. 52, 2022.

[38] B. Schumacher and M. A. Nielsen, "Quantum data processing and error correction," *Physical Review A*, vol. 54, no. 4, pp. 2629–2635, 1996.

[39] S. Lloyd, "Capacity of the noisy quantum channel," *Physical Review A*, vol. 55, no. 3, pp. 1613–1622, 1997.

[40] M. F. Cornelio, M. C. de Oliveira, and F. F. Fanchini, "Entanglement irreversibility from quantum discord and quantum deficit," *Physical Review Letters*, vol. 107, no. 2, p. 020502, 2011.

[41] C. H. Bennett, D. P. DiVincenzo, J. A. Smolin, and W. K. Wootters, "Mixed-state entanglement and quantum error correction," *Physical Review A*, vol. 54, no. 5, pp. 3824–3851, 1996.

[42] X. Lin, Z. Chen, and Z. Wei, "Quantifying quantum entanglement via a hybrid quantum-classical machine learning framework," *Physical Review A*, vol. 107, no. 6, p. 062409, 2023.

[43] M. Schuld, A. Bocharov, K. M. Svore, and N. Wiebe, "Circuit-centric quantum classifiers," *Physical Review A*, vol. 101, no. 3, p. 032308, 2020.

[44] A. Kandala, A. Mezzacapo, K. Temme, M. Takita, M. Brink, J. M. Chow, and J. M. Gambetta, "Hardware-efficient variational quantum

eigensolver for small molecules and quantum magnets," *Nature*, vol. 549, no. 7671, pp. 242–246, 2017.

[45] D. A. Meyer and N. R. Wallach, "Global entanglement in multiparticle systems," *Journal of Mathematical Physics*, vol. 43, no. 9, pp. 4273–4278, 2002.

[46] S. Sim, P. D. Johnson, and A. Aspuru-Guzik, "Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms," *Advanced Quantum Technologies*, vol. 2, no. 12, p. 1900070, 2019.

[47] U. Azad and A. Sinha, "qleet: visualizing loss landscapes, expressibility, entangling power and training trajectories for parameterized quantum circuits," *Quantum Information Processing*, vol. 22, no. 6, p. 256, 2023.

[48] T. Hubregtsen, J. Pichlmeier, P. Stecher, and K. Bertels, "Evaluation of parameterized quantum circuits: on the relation between classification accuracy, expressibility, and entangling capability," *Quantum Machine Intelligence*, vol. 3, no. 1, p. 9, 2021.

[49] K. Życzkowski and H.-J. Sommers, "Average fidelity between random quantum states," *Physical Review A*, vol. 71, p. 032313, Mar 2005.

[50] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[51] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[52] A. Abbas, D. Sutter, C. Zoufal, A. Lucchi, A. Figalli, and S. Woerner, "The power of quantum neural networks," *Nature Computational Science*, vol. 1, no. 6, pp. 403–409, 2021.

[53] V. Bergholm, J. Izaac, M. Schuld, C. Gogolin, S. Ahmed, V. Ajith, M. S. Alam, G. Alonso-Linaje, B. AkashNarayanan, A. Asadi, J. M. Arrazola, U. Azad, S. Banning, C. Blank, T. R. Bromley, B. A. Cordier, J. Ceroni, A. Delgado, O. D. Matteo, A. Dusko, T. Garg, D. Guala, A. Hayes, R. Hill, A. Ijaz, T. Isacsson, D. Ittah, S. Jahangiri, P. Jain, E. Jiang, A. Khandelwal, K. Kottmann, R. A. Lang, C. Lee, T. Loke, A. Lowe, K. McKiernan, J. J. Meyer, J. A. Montañez-Barrera, R. Moyard, Z. Niu, L. J. O'Riordan, S. Oud, A. Panigrahi, C.-Y. Park, D. Polatajko, N. Quesada, C. Roberts, N. Sá, I. Schoch, B. Shi, S. Shu, S. Sim, A. Singh, I. Strandberg, J. Soni, A. Száva, S. Thabet, R. A. Vargas-Hernández, T. Vincent, N. Vitucci, M. Weber, D. Wierichs, R. Wiersema, M. Willmann, V. Wong, S. Zhang, and N. Killoran, "Pennylane: Automatic differentiation of hybrid quantum-classical computations," 2022.

[54] Y. Kwak, W. J. Yun, S. Jung, J.-K. Kim, and J. Kim, "Introduction to quantum reinforcement learning: Theory and pennylane-based implementation," in *2021 International Conference on Information and Communication Technology Convergence (ICTC)*, pp. 416–420, 2021.

[55] J. Weng, H. Chen, D. Yan, K. You, A. Duburcq, M. Zhang, Y. Su, H. Su, and J. Zhu, "Tianshou: A highly modularized deep reinforcement learning library," *Journal of Machine Learning Research*, vol. 23, no. 267, pp. 1–6, 2022.

[56] H. Steingroever, D. J. Fridberg, A. Horstmann, K. L. Kjome, V. Kumari, S. D. Lane, T. V. Maia, J. L. McClelland, T. Pachur, P. Premkumar, J. C. Stout, R. Wetzels, S. Wood, D. A. Worthy, and E.-J. Wagenmakers, "Data from 617 healthy participants performing the iowa gambling task: A "many labs" collaboration," *Journal of Open Psychology Data*, 2015.

[57] J.-A. Li, D. Dong, Z. Wei, Y. Liu, Y. Pan, F. Nori, and X. Zhang, "Quantum reinforcement learning during human decision-making," *Nature Human Behaviour*, vol. 4, no. 3, pp. 294–307, 2020.

[58] B. Bahrami and J. Navajas, "4 arm bandit task dataset," 2022.

[59] N. D. Daw, J. P. O'Doherty, P. Dayan, B. Seymour, and R. J. Dolan, "Cortical substrates for exploratory decisions in humans," *Nature*, vol. 441, no. 7095, pp. 876–879, 2006.

[60] J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic, "Detecting and quantifying causal associations in large nonlinear time series datasets," *Science Advances*, vol. 5, no. 11, p. eaau4996, 2019.

[61] J. Runge, "Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets," in *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)* (J. Peters and D. Sontag, eds.), vol. 124 of *Proceedings of Machine Learning Research*, pp. 1388–1397, PMLR, 03–06 Aug 2020.

[62] A. Hyvärinen, K. Zhang, S. Shimizu, and P. O. Hoyer, "Estimation of a structural vector autoregression model using non-gaussianity," *Journal of Machine Learning Research*, vol. 11, no. 56, pp. 1709–1731, 2010.

[63] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen, "A linear non-gaussian acyclic model for causal discovery," *J. Mach. Learn. Res.*, vol. 7, p. 2003–2030, dec 2006.

[64] M. C. Caro, H.-Y. Huang, M. Cerezo, K. Sharma, A. Sornborger, L. Cincio, and P. J. Coles, "Generalization in quantum machine learning from few training data," *Nature Communications*, vol. 13, no. 1, p. 4919, 2022.

[65] M. C. Caro, H.-Y. Huang, N. Ezzell, J. Gibbs, A. T. Sornborger, L. Cincio, P. J. Coles, and Z. Holmes, "Out-of-distribution generalization for learning quantum dynamics," *Nature Communications*, vol. 14, no. 1, p. 3751, 2023.

[66] G. Adesso, A. Serafini, and F. Illuminati, "Quantification and scaling of multipartite entanglement in continuous variable systems," *Physical Review Letters*, vol. 93, p. 220504, Nov 2004.

[67] R. Jozsa and N. Linden, "On the role of entanglement in quantum-computational speed-up," *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 459, no. 2036, pp. 2011–2032, 2003. doi: 10.1098/rspa.2002.1097.

[68] J. L. Beckey, N. Gigena, P. J. Coles, and M. Cerezo, "Computable and operationally meaningful multipartite entanglement measures," *Physical Review Letters*, vol. 127, no. 14, p. 140501, 2021. PRL.

[69] D. Bruß and C. Macchiavello, "Multipartite entanglement in quantum algorithms," *Physical Review A*, vol. 83, p. 052313, May 2011.

[70] L. Bugalho, B. C. Coutinho, F. A. Monteiro, and Y. Omar, "Distributing multipartite entanglement over noisy quantum networks," *Quantum*, vol. 7, p. 920, 2023.