Measurement-Based Feedback Quantum Control with Deep Reinforcement Learning for a Double-Well Nonlinear Potential

Sangkha Borah,[1,*] Bijita Sarma,[1] Michael Kewming,[2] Gerard J. Milburn,[2] and Jason Twamley[1]

[1]*Quantum Machines Unit, Okinawa Institute of Science and Technology Graduate University,*
Onna-son, Okinawa 904-0495, Japan

[2]*Centre for Engineered Quantum Systems, School of Mathematics and Physics, University of Queensland,*
Brisbane, Queensland, 4072 Australia

(Received 16 April 2021; revised 19 July 2021; accepted 20 September 2021; published 2 November 2021)

Closed loop quantum control uses measurement to control the dynamics of a quantum system to achieve either a desired target state or target dynamics. In the case when the quantum Hamiltonian is quadratic in x and p, there are known optimal control techniques to drive the dynamics toward particular states, e.g., the ground state. However, for nonlinear Hamiltonian such control techniques often fail. We apply deep reinforcement learning (DRL), where an artificial neural agent explores and learns to control the quantum evolution of a highly nonlinear system (double well), driving the system toward the ground state with high fidelity. We consider a DRL strategy which is particularly motivated by experiment where the quantum system is continuously but weakly measured. This measurement is then fed back to the neural agent and used for training. We show that the DRL can effectively learn counterintuitive strategies to cool the system to a nearly pure "cat" state, which has a high overlap fidelity with the true ground state.

DOI: 10.1103/PhysRevLett.127.190403

As the research on quantum communication and computation has progressed rapidly with the goal of achieving the holy grail of quantum computing, quantum state engineering has begun to take on a high profile [1–3]. Of particular importance are feedback control techniques, in which a physical system subjected to noise is continuously monitored in real time while using measurement information to impart specific driving controls to modulate the system dynamics [4]. Unlike classical systems, measurement control of a quantum mechanical system is challenging for a number of reasons. First, the act of continuously observing a quantum system introduces nonlinearity within the conditioned dynamics. Second, continuous measurement on a quantum system generally alters it, generating measurement-induced noisy dynamics, commonly known as quantum back action. Finally, applying feedback which is dependent on the noisy measurement current adds further noise into the dynamics. Consequently, a variety of feedback control schemes that work well for classical systems may not for the analogous quantum counterparts [4–6].

In recent years, machine learning (ML) has rapidly gained interest, leading to numerous technological advancements in machine vision, voice recognition, natural language processing, automatic handwriting recognition, gaming, and engineering and robotics, to name a few [7]. Various ML models broadly fall into three categories: supervised learning, unsupervised learning, and reinforcement learning (RL) [7–9]. For supervised or unsupervised methods, the ML model is provided with enough labeled or unlabeled datasets to be trained on, which it uses for discovering the predictive hidden features in the system of interest. On the other hand, RL approaches the problem differently and is not pretrained with any external data explicitly, but learns in real time based on rewards. Indeed, RL is regarded as the most effective way to benefit from the creativity of machines, where it collects experiences by performing random experiments on the system (known as the environment in RL literature), learning by trial and error. RL, specially in combination with deep neural networks, abbreviated as DRL, is poised to revolutionize the field of artificial intelligence, particularly with the emergence of autonomous systems which process, in real time, stimuli from real world environments [10].

There have already been several important applications of ML in different areas of physics, such as in statistical mechanics, many-body systems, fluid dynamics, and quantum mechanics [11–15]. Most of these applications are supervised in nature, e.g., in the quantum domain, these have been applied to solving the many-body systems [16], in the determination of high-fidelity gates and the optimization of quantum memories by dynamic decoupling [17], quantum error corrections [18–20], quantum state tomography [21–25], classification and reconstruction of optical quantum states [26]. Recently, a few applications of DRL in quantum mechanics have also appeared that include applications in quantum control [27–30], quantum state preparation and engineering [31–35], state transfer [36–38], and quantum error correction [39,40]. While the number of works utilizing DRL is increasing, a very few consider using continuous measurement outcomes explicitly toward training the DRL agent [27,32,39]. As experiments often employ such continuous quantum measurement techniques for feedback control, we will consider this type of measurement as a key ingredient of our analysis below.

While traditionally known optimal control techniques work very well for linear, unitary, and deterministic quantum systems, there is no known generalized method for nonlinear and stochastic systems [6,27,41–47]. RL, on the contrary, is *agnostic* to the underlying physical model, but attempts to control the dynamics of the system by finding patterns from the data produced by it. In this Letter, we model the quantum evolution of a particle in a double well (DW) subject to continuous measurement at a rate Γ of the operator x^2, whose even parity avoids measurement localization of the particle's wave function to either well [48]. The DRL agent controls the quantum dynamics via a modulation of the Hamiltonian $H'(t) = H + F(t)$, with $F(t) = \mathcal{A}(t)(xp + px)$, where x and p are (dimensionless) canonical operators—a squeezing operator, whose strength $\mathcal{A}(t)$ is modulated by the DRL agent. The DRL agent is trained via the continuous measurement current, while, in real time, it acts back on the system via $F(t)$. We show that the DRL agent can be trained to cool the particle close to the ground state. Interestingly, the cooling efficiency depends on the choice of Γ, for which there is an optimal value of Γ to achieve the best cooling, which we identify numerically.

RL translates a problem at hand into a gamelike situation in which an artificial agent (also called the controller), finds a solution to the problem based on a trial-and-error approach [8,9], with no hints or suggestions on how to solve the problem itself. For this purpose, the agent is given a policy (in the case of DRL, it is the neural network itself), which is optimized based on some scalar values (reward) it receives from the environment (that includes the physics of the problem and the reward estimation function based on the observable) for each decision (action) made by it. By harnessing the power of search, coupled with many trials, the RL will gain experience from thousands of instances executed sequentially or in parallel in a sufficiently powerful computing infrastructure. After sufficient training, the agent can become skilled enough to have sophisticated tactics and superhuman abilities, as was phenomenally demonstrated by Google's AlphaGO [49,50]. To give a perspective on the applicability of RL in physics and the kinds of tasks it can solve, we provide a short demonstration to a problem in elementary mechanics, which we include as a media file in Supplemental Material [51] or the GitHub link [56].

It is possible to implement the DRL agent according to two distinct frameworks, a policy-based or value-based framework [9]. In policy-based frameworks, the policy parameters—the weights and biases of the neural network—are optimized directly based on the rewards it receives and informs its future actions on the environment [see Fig. 1(a)]. Value-based methods on the other hand, optimize the expected future return of a given value function, and deduce the policy from it [8]. It is possible to achieve the best of both these worlds by combining these two approaches in a meaningful way, known as actor-critic methods [9]. Here the actor is the policy which is being optimized, and the critic is the value function which is being learned. The actor network can be modeled using various policy-based approaches such as vanilla policy gradient [9], trust region policy optimization [57], or the more recent proximal policy optimization (PPO) [58]. In our work, we used PPO in combination with advantage actor-critic (A2C) [59] as a DRL agent. In the PPO scheme, it optimizes a clipped surrogate objective-loss function given by

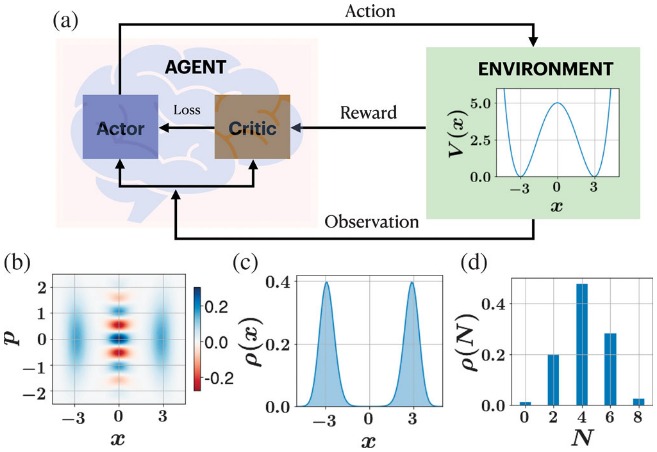

FIG. 1. (a) The working of a DRL actor-critic model. The agent consists of two networks, actor and critic, where the actor decides the action to be applied on the environment based on the suggestion made by the critic network that computes the value function based on the reward and state obtained from the DRL environment, that includes the physics of the problem—the quantum dynamics and state of particle moving in a DW and modulation function to alter quantum dynamics and the reward estimation function based on the observables (measurement results). (b) the Wigner function distribution for the ground state of the DW, and the corresponding probability distribution on position (c) and Fock-number basis (d). The ground state of the DW is an even parity state.

$$\mathcal{L}(\theta) = \hat{\mathbb{E}}_t\big(\min\{r_t(\theta)\hat{A}_t, \text{clip}[r_t(\theta), 1-\epsilon, 1+\epsilon]\}\big), \qquad (1)$$

where $r_t(\theta) = [\pi_\theta(a_t|s_t)/\pi_{\theta_{old}(a_t|s_t)}]$ is the probability ratio between current and old stochastic policies, so $r(\theta_{\text{old}}) = 1$. Furthermore, $\hat{\mathbb{E}}_t$ is the empirical average over a finite batch of samples and $\hat{A}_t$ is an estimator of the advantage function at time step t, obtained from the critic, and calculated as the difference between the Q value for action a_t at the state s_t and its average value, V: $\hat{A}_t(s_t|a_t) = Q(s_t|a_t) - V(s_t)$. Clipping the ratio within a bound specified by $\epsilon = 0.2$

ensures that the policy is not updated too much. The A2C framework allows synchronous training of multiple parallel worker environments simultaneously, which enables faster training. A more detailed theory can be found in the Supplemental Material [51]. A depiction of the DRL model employed in this study is shown in Fig. 1(a).

In this Letter, we will work with dimensionless position and momentum, denoted by (x, p). The relationship to the physical position and momentum variables is $x = Q/Q_0$, $p = P/P_0$ where Q_0, P_0 are suitable scales for position and momentum. As the canonical commutation relations are $[\hat{Q}, \hat{P}] = i\hbar$, thus $[\hat{x}, \hat{p}] = i\bar{k}$, where the dimensionless Planck's constant is defined by $\bar{k} = \hbar/(Q_0 P_0)$. The DW potential we consider is formed along the x axis by the Hamiltonian of a particle

$$H = \frac{p^2}{2} + \frac{h}{b^4}[(x-a)^2 - b^2]^2, \tag{2}$$

where b gives the location of the well's minima, h is the height of the barrier between the wells, and a is the offset along x. The ground state of this potential is a "cat" state thanks to the even parity symmetry of H in both x and p. The ground state can be depicted by the Wigner function $\mathcal{W}(x, p)$, which is shown in Fig. 1(b). The probability distribution along the x axis, i.e., $\rho(x) = \int \mathcal{W}(x, p)dp$ is shown in Fig. 1(c). Furthermore, the ground state has even parity symmetry while the first excited state has odd parity [60]. Hereafter, we will set the parameters $a = 0$, $b = 3.0$, and $h = 5$, which sets the potential to be symmetric around the origin at $x = 0$. It is worthy to note that the DW potential can now be engineered in laboratories such as in superconducting circuits, Bose-Einstein condensates, and magneto-optical setups [61].

To provide data to the agent, we consider that the quantum system is subject to a continuous measurement process and these measurement results are provided to the agent in real time. This continuous measurement also induces back action on the quantum system and noise on both the conditioned quantum dynamics and also on the observed measurement data. We can describe the dynamical evolution of the conditioned density operator $\rho_c(t)$, conditioned on a stochastic measurement record to time t via a quantum stochastic master equation [4,62] given by

$$d\rho_c(t) = -i[H, \rho_c]dt + \mathcal{D}[A]\rho_c dt + \mathcal{H}[A]\rho_c dW(t), \tag{3}$$

where A is a Hermitian observable operator under continuous measurement (known as the measurement operator), and $\mathcal{H}[A]$ and $\mathcal{D}[A]$ are superoperators given by

$$\mathcal{H}[A]\rho_c(t) = [\{A, \rho_c(t)\} - \mathrm{tr}(\{A, \rho_c\})]\rho(t), \tag{4}$$

$$\mathcal{D}[A]\rho_c(t) = \frac{1}{2}[2A\rho_c(t)A^+ - \{\rho_c(t), A^+A\}], \tag{5}$$

where $\{\cdot,\cdot\}$, denotes the anticommutator. Furthermore, $dW(t)$ in Eq. (3) represents a Wiener increment. It has mean zero and variance equal to dt. The measurement result current $I(t)$ are described by a classical stochastic process that satisfies an Ito stochastic differential equation (see Supplemental Material [51])

$$I(t)dt = \gamma g\left(\langle A(t)\rangle_c dt + \frac{1}{\sqrt{4\Gamma}}dW(t)\right), \tag{6}$$

where g is a gain with inverse units to that of the measurement operator A, meaning $I(t)$ has the units of frequency. For the context of the present work, we have $A = \sqrt{\Gamma}x^2$ and where Γ is the measurement rate and quantifies the quality of the measurement (see Supplemental Material [51]). As we have fixed units, so that x, p are dimensionless, we can set $\gamma g = 1$.

To cool the system to the ground state (which is a cat state) via continuous measurement, it is important to choose the stochastic operator A in Eq. (3) as $\sqrt{\Gamma}x^2$ instead of $\sqrt{\Gamma}x$, as the latter would collapse the state to either of the two minima of the DW [48]. At each interaction, the agent adds a squeezing term $F(t) \equiv \mathcal{A}(t)(xp + px)$ to the Hamiltonian [Eq. (2)], attempting to adjust the values of $A(t)$ in the continuous range $\mathcal{A}(t) \in [-5, 5]$, to maximize the reward. The choice of such feedback is motivated by the physics of the problem, which we explain in detail in the supporting information, backed by an analysis using Bayesian control driven by the conditional mean of the measurement record following the method of Stockton *et al.* [47]. It is possible to implement xp-type Hamiltonian terms via motion in a magnetic field [63]. In each episode of the training process, the DRL interacts with the environment 1000 times, in intervals of $\delta t = 0.01$, and each time applies an action to the environment. Further detail of the implementation and other technicalities of the DRL can be found in Supplemental Material [51]. The amplitude of the measurement noise depends on two parameters—(a) the measurement strength Γ and (b) the measurement time δt. Since the Wiener noise in Eq. (6) is a Gaussian with variance δt, the noise term in the measurement current $\mathcal{I}(t)$ scales at least as $1/\sqrt{4\Gamma\delta t}$. Because of this, one might expect the DRL to learn more efficiently for larger values of Γ, however, this is not the case. This makes it critical to choose an optimal value for Γ along with the measurement time δt. For a choice of $\delta t = 0.01$, we observe that optimal learning occurs near $\Gamma \sim 0.1$, and worsens for other values of Γ. Similar effects can be observed in Markovian measurement-based direct feedback, which we discuss in Supplemental Material [51]. Larger Γ values result in an increase in noise, while smaller Γ values return a very low signal-to-noise ratio in the continuous measurement process. The agent tends to learn most efficiently when the dynamics fluctuate in a limited domain around the DW minima. The effectiveness of the agent learning is shown in

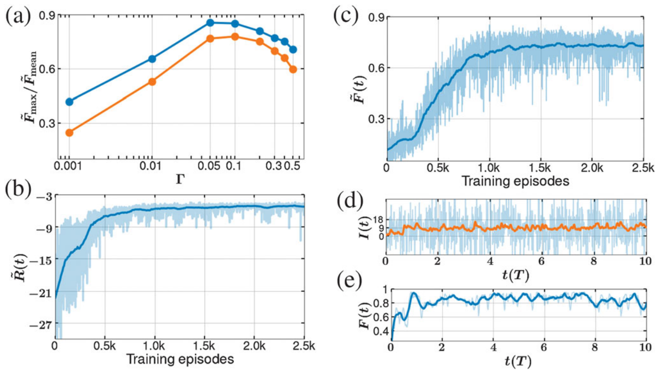

FIG. 2. (a) Effectiveness of the agent's learning process as a function of measurement strength Γ, indicating the existence of a "sweet spot" around $\Gamma \sim 0.1$. We plot the maximum (shown in blue lines) and mean (shown in brown lines) episodic $\tilde{F}$ of ten successive deterministic episodes where $\tilde{F}$ represents the mean fidelity over an episode using trained agents, (b) the episodic time evolution of the mean reward $\tilde{R}(t)$ during the training of the agent when $\Gamma = 0.1$ [light (dark) blue includes (average) noise], (c) episodic mean fidelity $\tilde{F}$ of the instantaneous $\rho(t)$ with the ground state of the DW, (d) measurement current $I(t)$ illustrating that a trained agent is able to keep the wave function near the well minima (the conditional average $\langle x_c^2 \rangle$ is shown in brown), and (e) the corresponding variation of fidelity for a trained episode, for a deterministic episode of the trained agent. Similar performance can be obtained using fidelity as the reward function (see Supplemental Material [51]).

Fig. 2(a), in terms of mean and maximum fidelities from ten successive deterministic episodes of the agents trained on the measurement current $I(t)$.

An important ingredient in the DRL is a suitable choice for the reward function. Many research studies have previously used fidelity or energy as the reward function. However, such a function is not practically available in experiments. Instead, we propose a measurement-based reward function $R(t) = -|I(t)/(\gamma g) - 3^2|$, where $I(t)$ is the measurement current [Eq. (6)]. This function obtains its maximum reward of 0, when $\langle x_c^2 \rangle = 3^2$, at the well minima positions. The learning process of the agent is shown in Fig. 2(b), along with the fidelity of the instantaneous state of the system with respect to the DW ground state in Fig. 2(c). Although such a fidelity is not possible to evaluate experimentally in real time, we present this as a check of the learning process. In a given episode the trained agent is able to adapt the feedback in such a way that the particle oscillates near the minima of the DW, as shown in Fig. 2(d) (for $\gamma g = 1$). The corresponding variation of fidelity for the episode is shown in Fig. 2(e). It is worthy to note that with a different choice of the reward function it might be possible to obtain better and more stable learning, see Supplemental Material [51] for details.

At the beginning of each episode, the environment must be reset to an initial state [initial density matrix $\rho(0)$], which is needed to start the stochastic master equation solver. We have found that the choice of $\rho(0)$ plays a crucial role in determining the total reward that can typically be achieved by the agent. When $\rho(0)$ is a thermal or coherent state, the DRL converges to an average fidelity of about 60%. However, if we use a small cat state or the ground state of the DW itself, the agent is able to achieve a mean trained fidelity of over 80% with noisy measurement data. The parity of the initial state is crucial, as the stochastic process of continuous measurement and feedback we have chosen is parity conserving. The target ground state of the DW, however, has even parity, and so choosing an initial state with a component of odd parity will lower the ultimate fidelity achievable. We achieve similar high performance if we start with a thermal state projected on to even parity, as done for Fig. 2. The explicit comparison of the performance of the trained agent with an untrained one is demonstrated in the Supplemental Material media [51] or the GitHub link [64].

We benchmark our results against the state-based Bayesian feedback protocol (where the feedback is based on an estimate of the state) as proposed by Doherty *et al.* [45,47]. In the context of the present work, the protocol can be simplified to provide feedback of the form $\mathcal{F}(t) = -(\langle x_c^2 \rangle(t) - 3^2) \times (xp + px)$, where $\langle x_c^2 \rangle(t)$ denotes the conditional mean of the observable x^2. We find that this Bayesian control achieves a mean fidelity of ~85%. However, $\langle x_c^2 \rangle(t)$ is not a quantity directly accessible in real experiments. When the Bayesian feedback is instead driven by the noisy measurement current $I(t)$ (which is available in experiments), we find that Bayesian feedback demonstrates almost no control over the dynamics. Numerical simulations with 1000 copies of the system (ensemble), evolving under a given feedback based on the mean of the measurement currents during each time step, yields an average fidelity of ~42%. This is considerably worse than the performance of the DRL agent. A more detailed discussion can be found in the Supplemental Material [51].

We have found that the DRL shows a robust control when the measurement efficiency $\eta > 50\%$, as shown in Fig. S5(a) of the Supplemental Material [51], implying that the DRL agent is able to find patterns in the underlying dynamics even when the stochasticity is significantly increased. Similarly, the DRL shows no significant drop of fidelity under additional dephasing of the form $\sim\sqrt{\gamma}a^\dagger a$, but this fidelity does drop for damping $\sim\sqrt{\gamma}a$, where γ is the decoherence rate (see Supplemental Material [51]).

On the computational side, the challenge for DRL control is the significant computational expense, e.g., 3–4 days of simulation time, even with fast computational CPUs, the bottleneck being the slow stochastic solver routines. We expect improved performance if the agent is made to learn in a dynamic combination of supervised (under a supervised setting an ML agent can learn more effectively from fewer data points, but is not reward based), and reinforcement learning (which is reward based and hence useful for feedback control), as done for image recognition in earlier studies [65–67]. It is possible to reshape the data to images of actions and measurement records which would enable the usage of convolution neural networks (CNN) in GPUs/TPUs with multicore support and utilize different image compression techniques, e.g., deep compressed sensing technique, proposed recently by researchers from DeepMind [68]. From the physics side, use of proper filters (as normally done in experiments) to filter the noisy signals prior to inputting into the DRL is expected to be crucial. In addition, the use of better reward estimation, such as combining constraints on current, fidelity (using tomography), and energy, is expected to be useful for further improvement. An even further innovation would be to use RL in combination with various optimal and Bayesian control protocols, as recently explored in applications outside quantum mechanics [20,69,70].

In conclusion, we have demonstrated the usefulness of deep reinforcement learning to tailor the nontrivial feedback parameters in a nonlinear system to engineer evolution toward the ground state. We found that the artificial agent can discover novel strategies solely based on measurement records to engineer high-fidelity "cat states" for the quantum double well.

The supporting media files for this Letter are openly available from the GitHub repository [56,64].

The authors thank the supercomputing facilities provided by the Okinawa Institute of Science and Technology (OIST) Graduate University and financial support. G. J. M. and M. K. acknowledge the support of the Australian Research Council Centre of Excellence for Engineered Quantum Systems CE170100009.

[*]sangkha.borah@oist.jp

[1] F. Verstraete, M. M. Wolf, and J. Ignacio Cirac, Nat. Phys. **5**, 633 (2009).

[2] M. Motta, C. Sun, A. T. K. Tan, M. J. O'Rourke, E. Ye, A. J. Minnich, F. G. S. L. Brandão, and G. K.-L. Chan, Nat. Phys. **16**, 205 (2020).

[3] P. J. Love, Nat. Phys. **16**, 130 (2020).

[4] H. M. Wiseman and G. J. Milburn, *Quantum Measurement and Control* (Cambridge University Press, Cambridge, 2009).

[5] K. Jacobs and D. A. Steck, Contemp. Phys. **47**, 279 (2006).

[6] J. Zhang, Y.-x. Liu, R.-B. Wu, K. Jacobs, and F. Nori, Phys. Rep. **679**, 1 (2017).

[7] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep Learning* (MIT Press, Cambridge, 2016), Vol. 1.

[8] 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, Nature (London) **518**, 529 (2015).

[9] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. (The MIT Press, Cambridge, 2018).

[10] Y. Li, arXiv:1908.06973.

[11] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, Rev. Mod. Phys. **91**, 045002 (2019).

[12] S. L. Brunton, B. R. Noack, and P. Koumoutsakos, Annu. Rev. Fluid Mech. **52**, 477 (2020).

[13] J. Carrasquilla and R. G. Melko, Nat. Phys. **13**, 431 (2017).

[14] V. Dunjko and H. J. Briegel, Rep. Prog. Phys. **81**, 074001 (2018).

[15] P. Mehta, M. Bukov, C.-H. Wang, A. G. R. Day, C. Richardson, C. K. Fisher, and D. J. Schwab, Phys. Rep. **810**, 1 (2019).

[16] G. Carleo and M. Troyer, Science **355**, 602 (2017).

[17] M. August and X. Ni, Phys. Rev. A **95**, 012335 (2017).

[18] P. Baireuther, T. E. O'Brien, B. Tarasinski, and C. W. J. Beenakker, Quantum **2**, 48 (2018).

[19] G. Torlai and R. G. Melko, Phys. Rev. Lett. **119**, 030501 (2017).

[20] S. Krastanov and L. Jiang, Sci. Rep. **7**, 11003 (2017).

[21] G. Torlai, G. Mazzola, J. Carrasquilla, M. Troyer, R. Melko, and G. Carleo, Nat. Phys. **14**, 447 (2018).

[22] M. Neugebauer, L. Fischer, A. Jäger, S. Czischek, S. Jochim, M. Weidemüller, and M. Gärttner, Phys. Rev. A **102**, 042604 (2020).

[23] S. Lohani, B. T. Kirby, M. Brodsky, O. Danaci, and R. T. Glasser, Mach. Learn. **1**, 035007 (2020).

[24] S. Ahmed, C. S. Muñoz, F. Nori, and A. F. Kockum, arXiv:2008.03240.

[25] A. M. Palmieri, E. Kovlakov, F. Bianchi, D. Yudin, S. Straupe, J. D. Biamonte, and S. Kulik, npj Quantum Inf. **6**, 20 (2020).

[26] S. Ahmed, C. S. Muñoz, F. Nori, and A. F. Kockum, arXiv:2012.02185.

[27] Z. T. Wang, Y. Ashida, and M. Ueda, Phys. Rev. Lett. **125**, 100401 (2020).

[28] M. Y. Niu, S. Boixo, V. N. Smelyanskiy, and H. Neven, npj Quantum Inf. **5**, 33 (2019).

[29] X.-M. Zhang, Z. Wei, R. Asad, X.-C. Yang, and X. Wang, npj Quantum Inf. **5**, 85 (2019).

[30] H. Xu, L. Wang, H. Yuan, and X. Wang, Phys. Rev. A **103**, 042615 (2021).

[31] X.-M. Zhang, Z. Wei, R. Asad, X.-C. Yang, and X. Wang, npj Quantum Inf. **5**, 85 (2019).

[32] J. Mackeprang, D. B. R. Dasari, and J. Wrachtrup, Quantum Mach. Intell. **2**, 5 (2020).

[33] T. Haug, W.-K. Mok, J.-B. You, W. Zhang, C. E. Png, and L.-C. Kwek, Mach. Learn. **2**, 01LT02 (2020).

[34] S.-F. Guo, F. Chen, Q. Liu, M. Xue, J.-J. Chen, J.-H. Cao, T.-W. Mao, M. K. Tey, and L. You, Phys. Rev. Lett. **126**, 060401 (2021).

[35] M. Bilkis, M. Rosati, R. M. Yepes, and J. Calsamiglia, Phys. Rev. Research **2**, 033295 (2020).

[36] R. Porotti, D. Tamascelli, M. Restelli, and E. Prati, Commun. Phys. **2**, 1 (2019).

[37] Y. Ding, Y. Ban, J. D. Martín-Guerrero, E. Solano, J. Casanova, and X. Chen, Phys. Rev. A **103**, L040401 (2021).

[38] I. Paparelle, L. Moro, and E. Prati, Phys. Lett. A **384**, 126266 (2020).

[39] T. Fösel, P. Tighineanu, T. Weiss, and F. Marquardt, Phys. Rev. X **8**, 031084 (2018).

[40] H. P. Nautrup, N. Delfosse, V. Dunjko, H. J. Briegel, and N. Friis, Quantum **3**, 215 (2019).

[41] J. Werschnik and E. K. U. Gross, J. Phys. B **40**, R175 (2007).

[42] A. P. Peirce, M. A. Dahleh, and H. Rabitz, Phys. Rev. A **37**, 4950 (1988).

[43] P. Doria, T. Calarco, and S. Montangero, Phys. Rev. Lett. **106**, 190501 (2011).

[44] E. Zahedinejad, S. Schirmer, and B. C. Sanders, Phys. Rev. A **90**, 032310 (2014).

[45] A. C. Doherty and K. Jacobs, Phys. Rev. A **60**, 2700 (1999).

[46] A. C. Doherty, S. Habib, K. Jacobs, H. Mabuchi, and S. M. Tan, Phys. Rev. A **62**, 012105 (2000).

[47] J. K. Stockton, R. van Handel, and H. Mabuchi, Phys. Rev. A **70**, 022106 (2004).

[48] K. Jacobs, L. Tian, and J. Finn, Phys. Rev. Lett. **102**, 057208 (2009).

[49] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, Nature (London) **529**, 484 (2016).

[50] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis, Nature (London) **550**, 354 (2017).

[51] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.127.190403 for brief theory of reinforcement learning and weak continuous measurement, technical details of the implementation and additional results, and media files illustrating the main results, which includes Refs. [52–55].

[52] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, arXiv:1606.01540.

[53] A. Raffin, A. Hill, M. Ernestus, A. Gleave, A. Kanervisto, and N. Dormann, Stable baselines3, https://github.com/DLR-RM/stable-baselines3 (2019).

[54] J. R. Johansson, P. D. Nation, and F. Nori, Comput. Phys. Commun. **183**, 1760 (2012).

[55] A. Paszke *et al.*, arXiv:1912.01703.

[56] S. Borah, Quantum Machines Unit, A little video of how deep reinforcement learning works to control the motion of a ball moving on an upside-down harmonic oscillator, https://github.com/QuantumMachinesUnit/ml-iho-demo (2021).

[57] J. Schulman, S. Levine, P. Moritz, M. I. Jordan, and P. Abbeel, arXiv:1502.05477.

[58] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, arXiv:1707.06347.

[59] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. P. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, arXiv:1602.01783.

[60] V. Jelic and F. Marsiglio, Eur. J. Phys. **33**, 1651 (2012).

[61] M. Abdi, P. Degenfeld-Schonburg, M. Sameti, C. Navarrete-Benlloch, and M. J. Hartmann, Phys. Rev. Lett. **116**, 233604 (2016).

[62] G. J. Milburn, Phys. Rev. A **36**, 744 (1987).

[63] E. Romero-Sánchez, W. P. Bowen, M. R. Vanner, K. Xia, and J. Twamley, Phys. Rev. B **97**, 024109 (2018).

[64] S. Borah, Quantum Machines Unit, Supporting media files for the paper measurement based feedback quantum control with deep reinforcement learning, https://github.com/QuantumMachinesUnit/ml-dw (2021).

[65] D. Kangin and N. Pugeault, in *Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN)* (IEEE, New York, 2018), pp. 1–8.

[66] E. Senft, P. Baxter, J. Kennedy, S. Lemaignan, and T. Belpaeme, Pattern Recogn. Lett. **99**, 77 (2017).

[67] L. Wang, W. Zhang, X. He, and H. Zha, arXiv:1807.01473.

[68] Y. Wu, M. Rosca, and T. Lillicrap, arXiv:1905.06723.

[69] S. D.-C. Shashua and S. Mannor, arXiv:2002.07171.

[70] A. Carron, M. Todescato, R. Carli, L. Schenato, and G. Pillonetto, in *Proceedings of the 2016 IEEE 55th Conference on Decision and Control (CDC)* (IEEE, New York, 2016), pp. 4594–4599.