RESEARCH ARTICLE

Multiqubit and multilevel quantum reinforcement learning with quantum technologies

F. A. Cárdenas-López[1,2]*, **L. Lamata**[3], **J. C. Retamal**[1,2], **E. Solano**[3,4,5]

1 Departamento de Física, Universidad de Santiago de Chile (USACH), Santiago, Chile, **2** Center for the Development of Nanoscience and Nanotechnology, Estación Central, Santiago, Chile, **3** Department of Physical Chemistry, University of the Basque Country UPV/EHU, Bilbao, Spain, **4** IKERBASQUE, Basque Foundation for Science, Bilbao, Spain, **5** Department of Physics, Shanghai University, Shanghai, China

* francisco.cardenas@usach.cl

Abstract

We propose a protocol to perform quantum reinforcement learning with quantum technologies. At variance with recent results on quantum reinforcement learning with superconducting circuits, in our current protocol coherent feedback during the learning process is not required, enabling its implementation in a wide variety of quantum systems. We consider diverse possible scenarios for an agent, an environment, and a register that connects them, involving multiqubit and multilevel systems, as well as open-system dynamics. We finally propose possible implementations of this protocol in trapped ions and superconducting circuits. The field of quantum reinforcement learning with quantum technologies will enable enhanced quantum control, as well as more efficient machine learning calculations.

Citation: Cárdenas-López FA, Lamata L, Retamal JC, Solano E (2018) Multiqubit and multilevel quantum reinforcement learning with quantum technologies. PLoS ONE 13(7): e0200455. https://doi.org/10.1371/journal.pone.0200455

Editor: Zoltan Zimboras, Wigner Research Centre, HUNGARY

Received: April 12, 2018

Accepted: June 25, 2018

Published: July 19, 2018

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: We acknowledge support from CEDENNA basal grant No. FB0807 and Dirección de Postgrado USACH (FAC-L), FONDECYT under grant No. 1140194 (JCR), Spanish MINECO/FEDER FIS2015-69983-P and Basque Government IT986-16 (LL and ES), and Ramón y Cajal Grant RYC-2012-11391 (LL).

Competing interests: L. L. declares his affiliation with PLOS ONE as Academic Editor. This does not

Introduction

Machine Learning (ML) is a subfield of Artificial Intelligence (AI) that has attracted increasing attention in the last years. ML usually refers to a computer program which can learn from experience E with respect to some class of task T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E [1]. In other words, Machine Learning addresses the problem of how a computer algorithm can be constructed to automatically improve with experience. Several applications in this field have been implemented such as handwriting pattern recognition [2], speech recognition [3] and the development of a computer able to beat an expert Go player [4], just to name a few.

The learning process in ML can be divided in three types: supervised learning, unsupervised learning and reinforcement learning [5]. In supervised machine learning, an initial data set has the function of training the system for later prediction making or to classify data. Usually, supervised learning problems are categorized into regression (continuous output) or classification (discrete output). Unsupervised learning allows one to address problems where the training data is not necessary and only correlations between subsets in the data (clustering) are considered and analyzed. Finally, reinforcement learning [6] differs from supervised and

alter our adherence to PLOS ONE policies on sharing data and materials.

unsupervised learning in that it takes into account a scalar parameter (reward) to evaluate the input-output relation in a trial and error way. In this case, the system (so-called "agent") obtains information from its outer world ("environment") to decide which is the better way to optimize itself, for adapting to the environment.

Quantum information processing (QIP) could contribute positively in the future in the development of the machine learning field, with several quantum algorithms for machine learning with significant possible gains with respect to their classical counterparts [7–11]. More specifically, quantum algorithms have been developed and in some cases implemented for supervised and unsupervised learning problems [12–18]. However, quantum reinforcement learning has not been widely explored and just a few results have been obtained up to now [19–26]. Related topics in biomimetic quantum technologies are quantum memristors [27–30], as well as quantum Helmholtz and Boltzmann machines [31–33]. These, together with quantum reinforcement learning, may set the stage for the future development of semi-autonomous quantum devices.

The field of quantum technologies has grown extensively in the past decade. In particular, two architectures which are very promising for the implementation of a quantum computer, in terms of numbers of qubits and gate fidelities, are trapped ions [34, 35] and superconducting circuits [36–38]. Current technological progress in trapped ions has allowed us to implement quantum protocols with several ions involving high-fidelity single and two-qubit gates as well as high-fidelity readout [39, 40]. Superconducting circuits have also proven to be an excellent platform to perform quantum information processing protocols because of their individual addressing and scalability. Two-qubit quantum gates have achieved fidelities larger than 99% [41, 42] in this platform. Furthermore, technological progress in this architecture has made possible to build artificial atoms with high coherence time in coplanar [43] and 3D architecture [44], allowing for the development of feedback control with superconducting circuits [45, 46]. This feedback mechanism has inspired protocols for quantum reinforcement learning with superconducting circuits [23] where the feedback loop control allows one to reward and restart the system to obtain maximal learning fidelity.

Here, we propose a general protocol to perform quantum reinforcement learning with quantum technologies. We understand general in the sense that it goes beyond the context of qubits for embedding information in agent or environment. In this sense, and at variance with a previous result [23], we extend the realm of the quantum reinforcement learning protocol to multi-qubit, multi-level, and open quantum systems, therefore permitting a wider set of scenarios. Our protocol considers a quantum system (the agent), which interacts with an external quantum system (its environment) via an auxiliary quantum system (a register). The aim of our quantum reinforcement learning protocol is for the agent to acquire information from its environment and adapt to it, via a rewarding mechanism. In this fully quantum scenario the meaning of the learning process is the establishment of quantum correlations among the parties [21]. In our specific case, the quantum agent aims at attaining maximum quantum state overlap with the environment state, in the sense that local measurements on agent and environment will produce the same outcomes or, equivalently, that the agent and environment entangled final state is invariant under the exchange of these two subsystems. An interpretation of this outcome is that the agent can learn about the information embedded in the environment state, which has been consequently modified from a separable to an entangled state with the agent and registers. After this process we are in position of evaluating any figure of merit with the outcome measurements. Optimizing this figure of merit should be associated to a particular learning process probably requiring particular actions to be applied on the agent. Another possible result is obtained by considering projective measurements in the register systems. Only after these projective measurements agent and environment will be decoupled from them and the protocol assures that the former are in a pure correlated state, without needing to know any information about their initial states. We analyze the case where the register subspace is larger than agent and environment subspaces. The inclusion of more elements in the register subspace allows for delaying the application of the rewarding criterion to the end of the quantum protocol. This fact will enable its implementation in a wider variety of quantum platforms, besides superconducting circuits with coherent feedback. We also study quantum reinforcement learning in the case where agent, environment and register are composed of qudits. In this case, we obtain that the maximal learning fidelity is achieved in a fixed number of steps in the qudit dimension, and this number scales polynomially with the number of subsystems in the environment subspace. In addition, we analyse quantum reinforcement learning in the situation where the environment is larger than the agent. We highlight two results: the first of them is obtained when considering that the register has the same elements than the environment. In this case, two rewarding criteria are needed to obtain maximal learning fidelity and the entanglement between the agent and a specific part of the environment is a key resource. The other case is the situation where the register has more elements than the environment. In this case, only one measurement is needed to obtain maximal learning fidelity and the environment-agent entanglement is not a key resource. Based on this fact, the rewarding criterion is applied at the end of the protocol. Finally, we describe how our quantum learning protocols can be implemented in quantum platforms as trapped ions and superconducting circuits.

Quantum reinforcement learning protocol with final measurement

Here, we introduce a protocol to perform quantum reinforcement learning, which introduces significant novelties with respect to the existing literature. Unlike a previous quantum reinforcement learning result [23], the protocol described here needs one measurement at the end of the procedure and no feedback, allowing for its implementation in a variety of quantum platforms including ions and photons. The improvement relies on adding more registers than before [23] and making them interact conditionally with each other. The inclusion of ancillary systems has proven to be useful in several implementations of quantum information, because measurements on the ancillary system allow one in principle to obtain information about the main system without destroying it. Moreover, the measurement associated with the rewarding criterion is performed at the end of the protocol. This opens the possibility to implement quantum reinforcement learning protocols in architectures for which implementing coherent feedback may be a challenging problem.

The quantum reinforcement learning protocol described here works in the following way. We firstly consider an agent and environment, composed of one qubit each, and two register qubits, see Fig 1. The first step is to encode the environment information in the register states (usually this kind of operation in the context of classical reinforcement learning is called the action). Subsequently, the internal states of the registers interact conditionally with the agent (usually this kind of operation in classical reinforcement learning is called the percept). Finally, an agent-register interaction changes the agent state (partial rewarding mechanism). At this stage the rewarding criterion is satisfied, in the form of a correlated agent-environment state, in the sense that local measurements on agent and environment will produce the same outcomes. On the other hand, the agent-environment system is also entangled with the two registers, and in order to attain a correlated pure state of agent and environment, a single, final measurement may be performed on the two register states. This will produce an agent-environment state maximizing the learning fidelity defined as $\mathcal{F}_{AE} = |\langle\psi_A|\phi_E\rangle|$, where $|\psi_A\rangle$ is the agent state and $|\phi_E\rangle$ is the environment state, both after the protocol.

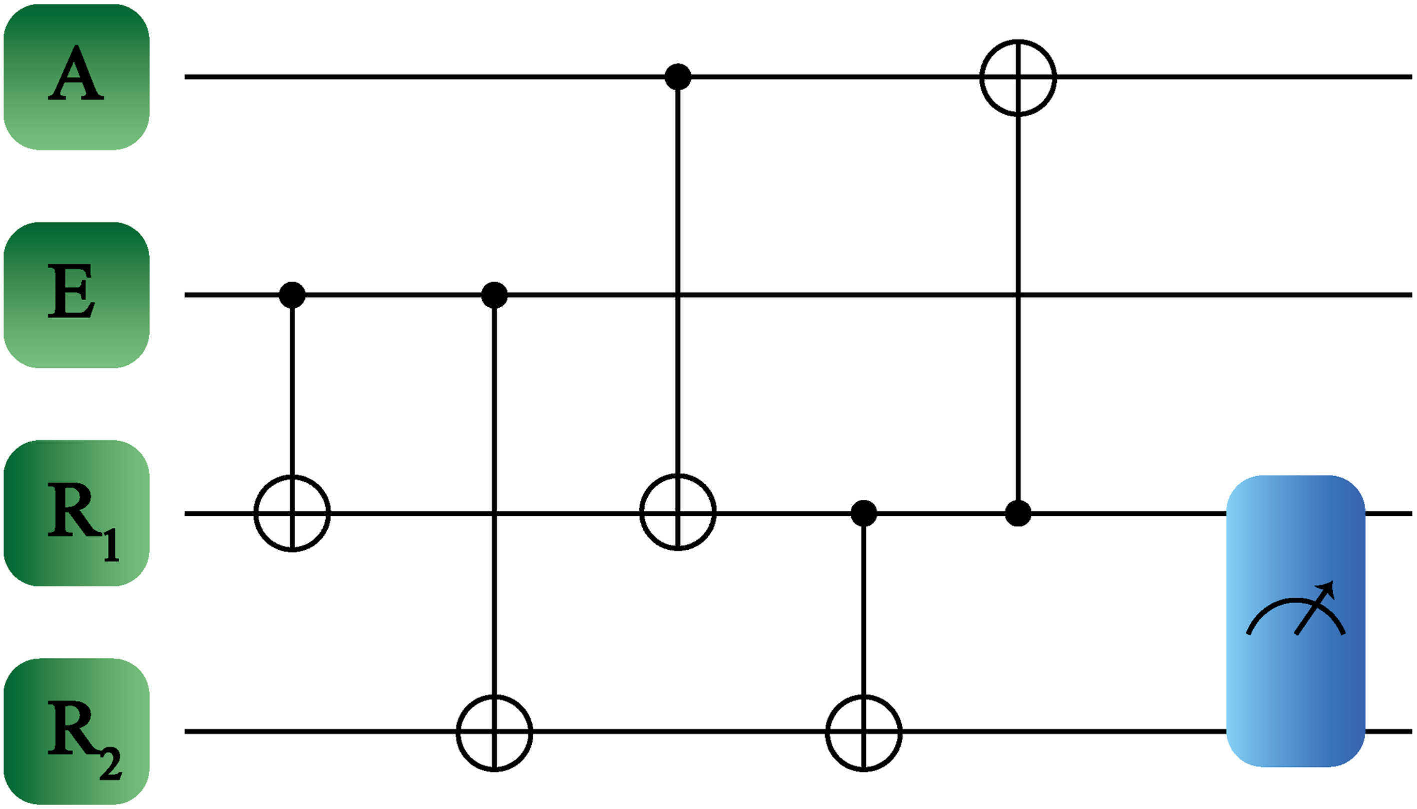

Fig 1. Proposed protocol to perform quantum reinforcement learning with final measurement. We consider a set composed of four qubits, corresponding to agent A, environment E, and registers R_1 and R_2. The considered interactions agent-register, register-register and environment-register consist of CNOT gates. The measurement in the register subspace is denoted by the rightmost box.

https://doi.org/10.1371/journal.pone.0200455.g001

To perform our quantum reinforcement learning protocol we consider that initially agent and environment are in arbitrary single-qubit pure states, whereas the register states are in their ground state, namely

$$\{|A\rangle = \alpha_A^0|0\rangle_A + \alpha_A^1|1\rangle_A, |E\rangle = \alpha_E^0|0\rangle_E + \alpha_E^1|1\rangle_E, |R\rangle = |0\rangle_1|0\rangle_2\} \tag{1}$$

$$|\Psi\rangle_0 \quad = \quad |A\rangle|E\rangle|R\rangle. \tag{2}$$

The first step in the protocol is to extract information from the environment, updating the information in the registers conditionally to the environment state. This process is done by applying a pair of CNOT gates in the environment-register subspace. Here, the first system is the control and the second the target,

$$|\Psi\rangle_1 \quad = \quad U^{\rm CNOT}_{(E,R_2)} U^{\rm CNOT}_{(E,R_1)} |\Psi\rangle_0, \tag{3}$$

$$|\Psi\rangle_1 \quad = \quad (\alpha_A^0|0\rangle_A + \alpha_A^1|1\rangle_A)(\alpha_E^0|0\rangle_E|0\rangle_1|0\rangle_2 + \alpha_E^1|1\rangle_E|1\rangle_1|1\rangle_2). \tag{4}$$

Then, the information encoded on the registers is updated conditional on the agent state. As the register subspace is larger than the agent subspace, we will choose which part of the register subspace will the agent update. Without loss of generality, let us assume that the register R_1

will be updated. The upgrade of agent subspace is performed by a CNOT gate acting in the $A-R_1$ subspace, where the agent state is the control and the register is the target,

$$\begin{aligned}
|\Psi\rangle_2 &= U^{\text{CNOT}}_{(A,R_1)}|\Psi\rangle_1, \\
|\Psi\rangle_2 &= (\alpha_A^0\alpha_E^0|0\rangle_A|0\rangle_E|0\rangle_1|0\rangle_2 + \alpha_A^0\alpha_E^1|0\rangle_A|1\rangle_E|1\rangle_1|1\rangle_2 + \alpha_A^1\alpha_E^0|1\rangle_A|0\rangle_E|1\rangle_1|0\rangle_2 \\
&\quad +\alpha_A^1\alpha_E^1|1\rangle_A|1\rangle_E|0\rangle_1|1\rangle_2).
\end{aligned} \tag{5}$$

Subsequently, the register R_2 is also updated with respect to the R_1 state. This is accomplished by applying a CNOT gate in the register subspace, where R_1 acts as control and R_2 as target,

$$\begin{aligned}
|\Psi\rangle_3 &= U^{\text{CNOT}}_{(R_1,R_2)}|\Psi\rangle_2, \\
|\Psi\rangle_3 &= (\alpha_A^0\alpha_E^0|0\rangle_A|0\rangle_E|0\rangle_1|0\rangle_2 + \alpha_A^0\alpha_E^1|0\rangle_A|1\rangle_E|1\rangle_1|0\rangle_2 + \alpha_A^1\alpha_E^0|1\rangle_A|0\rangle_E|1\rangle_1|1\rangle_2 \\
&\quad +\alpha_A^1\alpha_E^1|1\rangle_A|1\rangle_E|0\rangle_1|1\rangle_2).
\end{aligned} \tag{6}$$

Followingly, we update the agent state according to the information encoded in the register R_1. This is done by applying a CNOT gate in the R_1-A subspace, where R_1 is the control and A is the target,

$$\begin{aligned}
|\Psi\rangle_4 &= U^{\text{CNOT}}_{(R_1,A)}|\Psi\rangle_3, \\
|\Psi\rangle_4 &= (\alpha_A^0\alpha_E^0|0\rangle_A|0\rangle_E|0\rangle_1|0\rangle_2 + \alpha_A^0\alpha_E^1|1\rangle_A|1\rangle_E|1\rangle_1|0\rangle_2 + \alpha_A^1\alpha_E^0|0\rangle_A|0\rangle_E|1\rangle_1|1\rangle_2 \\
&\quad +\alpha_A^1\alpha_E^1|1\rangle_A|1\rangle_E|0\rangle_1|1\rangle_2).
\end{aligned} \tag{7}$$

We point out that, in the previous state, agent and environment are already maximally correlated, in the sense of having the same outcomes with respect to local measurements performed on either of them, or, equivalently, the state is invariant under particle exchange with respect to the agent-environment subsystem. We also remark that this state is general, valid for any initial agent and environment states. The fact that agent and environment get entangled with the two registers allows one to distinguish between identical agent-environment components that originate from different initial states, namely, to distinguish between states arising from $\alpha_A^0\alpha_E^0$ or $\alpha_A^1\alpha_E^0$, as well as from $\alpha_A^0\alpha_E^1$ or $\alpha_A^1\alpha_E^1$.

Finally, by performing a projective measurement on the register subspace, the rewarding criteron is satisfied. It is easy to show that, independently of the measurement outcome, the learning fidelity $\mathcal{F}_{AE} = |\langle\psi_A|\phi_E\rangle|$ is maximal, given that agent and environment states end up being in the same state, either $|0\rangle$ or $|1\rangle$. In this case only one iteration of the protocol is sufficient in order that the agent adapts to the environment. Moreover, throughout the protocol, measurements on agent and/or environment are not required, which may allow its implementation in a variety of quantum platforms as trapped ions, superconducting circuits, and quantum photonics.

In our protocol, we do not need coherent feedback given that the registers entangle with agent and environment and as a result produce the desired agent-environment state that is invariant under permutation. It is true that the entanglement with the registers produces a mixed state in case the register states are discarded, but this is not a drawback in our protocol. Indeed, what our protocol does is, for arbitrary initial agent and environment states, which need not be known, to give a constructive way to produce a final agent-environment state perfectly correlated, in the sense of invariant under permutations in agent-environment subspace. This state is in general entangled, namely, quantum, and we do not need to perform any measurement on agent and environment during the protocol, namely, it can equally well work with photons, ions, and superconducting circuits, among others. After the production of the agent-environment-register entangled state, the registers are entangled with agent and environment, but this does not prevent us from measuring the registers at a certain desired time, and decoupling agent and environment from them. This way, we will not have measured agent and environment at any time of the protocol, and we can assure that they are perfectly correlated irrespective of their initial states, and without having any prior information about them. This may be useful, e.g., for distributing private keys in quantum cryptography for arbitrary, unknown, initial states, without the need to initialize agent and register in reference states.

Quantum reinforcement learning for multiqubit systems with final measurement

In the previous section, we have showed that by considering more than just one register the rewarding criterion in the quantum reinforcement learning algorithm can be done at the end of our protocol. The same results can be obtained when we consider more complex configurations. Indeed, by assuming that agent and register are composed of two qubits each, and four qubits act as registers, we show that the rewarding criterion can also be applied at the end of the quantum protocol. Let us illustrate this fact with an analysis for multiqubit agent, environment, and register states,

$$|A\rangle = \alpha_A^{00}|00\rangle_A + \alpha_A^{01}|01\rangle_A + \alpha_A^{10}|10\rangle_A + \alpha_A^{11}|11\rangle_A, \tag{8}$$

$$|E\rangle = \alpha_E^{00}|00\rangle_E + \alpha_E^{01}|01\rangle_E + \alpha_E^{10}|10\rangle_E + \alpha_E^{11}|11\rangle_E, \tag{9}$$

$$|R\rangle = |0\rangle_1|0\rangle_2|0\rangle_3|0\rangle_4, \tag{10}$$

$$|\Psi\rangle_0 = |A\rangle|E\rangle|R\rangle. \tag{11}$$

Following the same procedure described previously, the protocol consists mainly in three types of interaction, as shown in Fig 2. Firstly, we update the registers conditionally to the environment states. More specifically, we consider an interaction between the environment qubits E_1 and E_2 with the registers R_1 and R_2, respectively. In this description, the environment acts as control and the registers act as targets in the CNOT gates,

$$\begin{aligned}
|\Psi\rangle_1 &= U^{\text{CNOT}}_{(E_1,R_1)}U^{\text{CNOT}}_{(E_2,R_2)}, |\Psi\rangle_0, \\
|\Psi\rangle_1 &= |A\rangle(\alpha_E^{00}|00\rangle_E|0\rangle_1|0\rangle_2|0\rangle_3|0\rangle_4 + \alpha_E^{01}|01\rangle_E|0\rangle_1|1\rangle_2|0\rangle_3|0\rangle_4 \\
&\quad +\alpha_E^{10}|10\rangle_E|1\rangle_1|0\rangle_2|0\rangle_3|0\rangle_4 + \alpha_E^{11}|11\rangle_E|1\rangle_1|1\rangle_2|0\rangle_3|0\rangle_4).
\end{aligned} \tag{12}$$

Thereafter, we update similarly the remaining registers, that is, we apply a CNOT gate between the environment qubits E_1 and E_2 and the register qubits R_3 and R_4, respectively, obtaining

$$\begin{aligned}
|\Psi\rangle_2 &= U^{\text{CNOT}}_{(E_1,R_3)}U^{\text{CNOT}}_{(E_2,R_4)}|\Psi\rangle_1, \\
|\Psi\rangle_2 &= |A\rangle(\alpha_E^{00}|00\rangle_E|0\rangle_1|0\rangle_2|0\rangle_3|0\rangle_4 + \alpha_E^{01}|01\rangle_E|0\rangle_1|1\rangle_2|0\rangle_3|1\rangle_4 \\
&\quad +\alpha_E^{10}|10\rangle_E|1\rangle_1|0\rangle_2|1\rangle_3|0\rangle_4 + \alpha_E^{11}|11\rangle_E|1\rangle_1|1\rangle_2|1\rangle_3|1\rangle_4).
\end{aligned} \tag{13}$$

Next step consists in updating a part of the register subspace conditionally to the agent state.

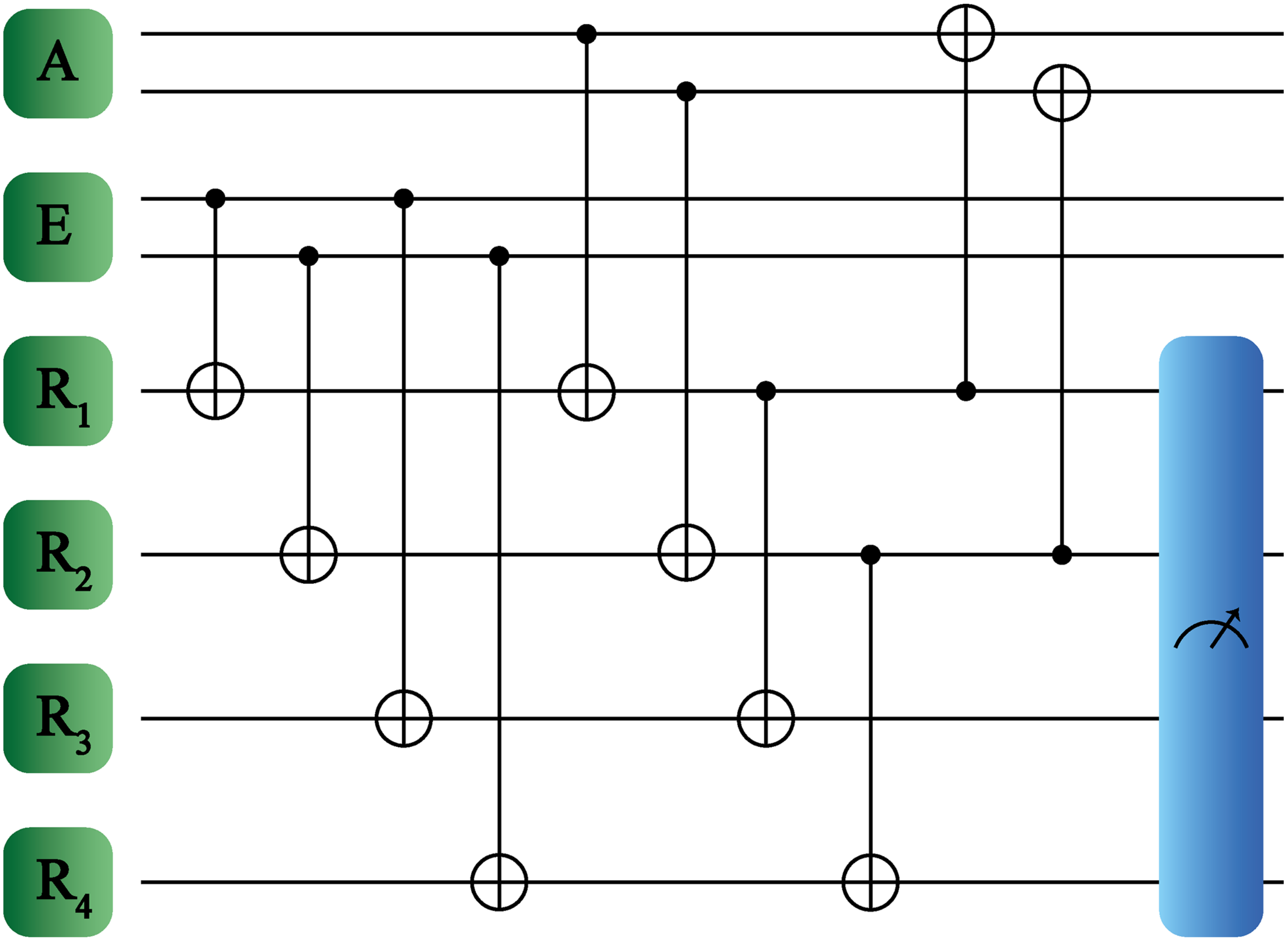

Fig 2. Schematic representation of quantum reinforcement learning protocol for multiqubit systems. Agent, environment and registers are denoted as A, E and R_1, R_2, R_3 and R_4, respectively. The measurement in the register subspace is denoted by the rightmost box.

https://doi.org/10.1371/journal.pone.0200455.g002

Thus, the registers R_1 and R_2 will be updated via A_1 and A_2, respectively,

$$\begin{array}{rcl}
|\Psi\rangle_3 & = & U^{\mathrm{CNOT}}_{(A_1,R_1)} U^{\mathrm{CNOT}}_{(A_2,R_2)} |\Psi\rangle_2, \\
|\Psi\rangle_3 & = & \alpha_A^{00}\alpha_E^{00}|00\rangle_A|00\rangle_E|0\rangle_1|0\rangle_2|0\rangle_3|0\rangle_4 + \alpha_A^{00}\alpha_E^{01}|00\rangle_A|01\rangle_E|0\rangle_1|1\rangle_2|0\rangle_3|1\rangle_4 \\
 & & +\alpha_A^{00}\alpha_E^{10}|00\rangle_A|10\rangle_E|1\rangle_1|0\rangle_2|1\rangle_3|0\rangle_4 + \alpha_A^{00}\alpha_E^{11}|00\rangle_A|11\rangle_E|1\rangle_1|1\rangle_2|1\rangle_3|1\rangle_4 \\
 & & + \alpha_A^{01}\alpha_E^{00}|01\rangle_A|00\rangle_E|0\rangle_1|1\rangle_2|0\rangle_3|0\rangle_4 + \alpha_A^{01}\alpha_E^{01}|01\rangle_A|01\rangle_E|0\rangle_1|0\rangle_2|0\rangle_3|1\rangle_4 \\
 & & +\alpha_A^{01}\alpha_E^{10}|01\rangle_A|10\rangle_E|1\rangle_1|1\rangle_2|1\rangle_3|0\rangle_4 + \alpha_A^{01}\alpha_E^{11}|01\rangle_A|11\rangle_E|1\rangle_1|0\rangle_2|1\rangle_3|1\rangle_4 \\
 & & +\alpha_A^{10}\alpha_E^{00}|10\rangle_A|00\rangle_E|1\rangle_1|0\rangle_2|0\rangle_3|0\rangle_4 + \alpha_A^{10}\alpha_E^{01}|10\rangle_A|01\rangle_E|1\rangle_1|1\rangle_2|0\rangle_3|1\rangle_4 \\
 & & +\alpha_A^{10}\alpha_E^{10}|10\rangle_A|10\rangle_E|0\rangle_1|0\rangle_2|1\rangle_3|0\rangle_4 + \alpha_A^{10}\alpha_E^{11}|10\rangle_A|11\rangle_E|0\rangle_1|1\rangle_2|1\rangle_3|1\rangle_4 \\
 & & +\alpha_A^{11}\alpha_E^{00}|11\rangle_A|00\rangle_E|1\rangle_1|1\rangle_2|0\rangle_3|0\rangle_4 + \alpha_A^{11}\alpha_E^{01}|11\rangle_A|01\rangle_E|1\rangle_1|0\rangle_2|0\rangle_3|1\rangle_4 \\
 & & +\alpha_A^{11}\alpha_E^{10}|11\rangle_A|10\rangle_E|0\rangle_1|1\rangle_2|1\rangle_3|0\rangle_4 + \alpha_A^{11}\alpha_E^{11}|11\rangle_A|11\rangle_E|0\rangle_1|0\rangle_2|1\rangle_3|1\rangle_4.
\end{array} \tag{14}$$

Afterwards, to obtain orthogonal outcomes in the register subspace we perform a pair of CNOT gates in this subspace. The interaction will be between the registers that interact with a common environment, namely, register R_1 interacts with R_3 because both have interacted with E_1. Similarly for R_2 and R_4, which have interacted with E_2. In this case, $R_1(R_2)$ is the control and $R_3(R_4)$ is the target.

$$\begin{aligned}
|\Psi\rangle_4 &= U^{\text{CNOT}}_{(R_1,R_3)} U^{\text{CNOT}}_{(R_2,R_4)} |\Psi\rangle_3, \\
|\Psi\rangle_4 &= \alpha_A^{00}\alpha_E^{00}|00\rangle_A|00\rangle_E|0\rangle_1|0\rangle_2|0\rangle_3|0\rangle_4 + \alpha_A^{00}\alpha_E^{01}|00\rangle_A|01\rangle_E|0\rangle_1|1\rangle_2|0\rangle_3|0\rangle_4 \\
&+\alpha_A^{00}\alpha_E^{10}|00\rangle_A|10\rangle_E|1\rangle_1|0\rangle_2|0\rangle_3|0\rangle_4 + \alpha_A^{00}\alpha_E^{11}|00\rangle_A|11\rangle_E|1\rangle_1|1\rangle_2|0\rangle_3|0\rangle_4 \\
&+\alpha_A^{01}\alpha_E^{00}|01\rangle_A|00\rangle_E|0\rangle_1|1\rangle_2|0\rangle_3|1\rangle_4 + \alpha_A^{01}\alpha_E^{01}|01\rangle_A|01\rangle_E|0\rangle_1|0\rangle_2|0\rangle_3|1\rangle_4 \\
&+\alpha_A^{01}\alpha_E^{10}|01\rangle_A|10\rangle_E|1\rangle_1|1\rangle_2|0\rangle_3|1\rangle_4 + \alpha_A^{01}\alpha_E^{11}|01\rangle_A|11\rangle_E|1\rangle_1|0\rangle_2|0\rangle_3|1\rangle_4 \\
&+\alpha_A^{10}\alpha_E^{00}|10\rangle_A|00\rangle_E|1\rangle_1|0\rangle_2|1\rangle_3|0\rangle_4 + \alpha_A^{10}\alpha_E^{01}|10\rangle_A|01\rangle_E|1\rangle_1|1\rangle_2|1\rangle_3|0\rangle_4 \\
&+\alpha_A^{10}\alpha_E^{10}|10\rangle_A|10\rangle_E|0\rangle_1|0\rangle_2|1\rangle_3|0\rangle_4 + \alpha_A^{10}\alpha_E^{11}|10\rangle_A|11\rangle_E|0\rangle_1|1\rangle_2|1\rangle_3|0\rangle_4 \\
&+\alpha_A^{11}\alpha_E^{00}|11\rangle_A|00\rangle_E|1\rangle_1|1\rangle_2|1\rangle_3|1\rangle_4 + \alpha_A^{11}\alpha_E^{01}|11\rangle_A|01\rangle_E|1\rangle_1|0\rangle_2|1\rangle_3|1\rangle_4 \\
&+\alpha_A^{11}\alpha_E^{10}|11\rangle_A|10\rangle_E|0\rangle_1|1\rangle_2|1\rangle_3|1\rangle_4 + \alpha_A^{11}\alpha_E^{11}|11\rangle_A|11\rangle_E|0\rangle_1|0\rangle_2|1\rangle_3|1\rangle_4.
\end{aligned} \tag{15}$$

Finally, we update the agent considering the states of the register in order that the rewarding criterion is satisfied. This is done by applying two CNOT gates in the agent-register subspace, where A_1 is controlled by R_1 and A_2 is controlled by R_2,

$$\begin{aligned}
|\Psi\rangle_5 &= U^{\text{CNOT}}_{(R_1,A_1)} U^{\text{CNOT}}_{(R_2,A_2)} |\Psi\rangle_4, \\
|\Psi\rangle_5 &= \alpha_A^{00}\alpha_E^{00}|00\rangle_A|00\rangle_E|0\rangle_1|0\rangle_2|0\rangle_3|0\rangle_4 + \alpha_A^{00}\alpha_E^{01}|01\rangle_A|01\rangle_E|0\rangle_1|1\rangle_2|0\rangle_3|0\rangle_4 \\
&+\alpha_A^{00}\alpha_E^{10}|10\rangle_A|10\rangle_E|1\rangle_1|0\rangle_2|0\rangle_3|0\rangle_4 + \alpha_A^{00}\alpha_E^{11}|11\rangle_A|11\rangle_E|1\rangle_1|1\rangle_2|0\rangle_3|0\rangle_4 \\
&+\alpha_A^{01}\alpha_E^{00}|00\rangle_A|00\rangle_E|0\rangle_1|1\rangle_2|0\rangle_3|1\rangle_4 + \alpha_A^{01}\alpha_E^{01}|01\rangle_A|01\rangle_E|0\rangle_1|0\rangle_2|0\rangle_3|1\rangle_4 \\
&+\alpha_A^{01}\alpha_E^{10}|10\rangle_A|10\rangle_E|1\rangle_1|1\rangle_2|0\rangle_3|1\rangle_4 + \alpha_A^{01}\alpha_E^{11}|11\rangle_A|11\rangle_E|1\rangle_1|0\rangle_2|0\rangle_3|1\rangle_4 \\
&+\alpha_A^{10}\alpha_E^{00}|00\rangle_A|00\rangle_E|1\rangle_1|0\rangle_2|1\rangle_3|0\rangle_4 + \alpha_A^{10}\alpha_E^{01}|01\rangle_A|01\rangle_E|1\rangle_1|1\rangle_2|1\rangle_3|0\rangle_4 \\
&+\alpha_A^{10}\alpha_E^{10}|10\rangle_A|10\rangle_E|0\rangle_1|0\rangle_2|1\rangle_3|0\rangle_4 + \alpha_A^{10}\alpha_E^{11}|11\rangle_A|11\rangle_E|0\rangle_1|1\rangle_2|1\rangle_3|0\rangle_4 \\
&+\alpha_A^{11}\alpha_E^{00}|00\rangle_A|00\rangle_E|1\rangle_1|1\rangle_2|1\rangle_3|1\rangle_4 + \alpha_A^{11}\alpha_E^{01}|01\rangle_A|01\rangle_E|1\rangle_1|0\rangle_2|1\rangle_3|1\rangle_4 \\
&+\alpha_A^{11}\alpha_E^{10}|10\rangle_A|10\rangle_E|0\rangle_1|1\rangle_2|1\rangle_3|1\rangle_4 + \alpha_A^{11}\alpha_E^{11}|11\rangle_A|11\rangle_E|0\rangle_1|0\rangle_2|1\rangle_3|1\rangle_4.
\end{aligned} \tag{16}$$

From the latter Eq (16), it is straightforward to see that independently of the measurement outcomes the learning fidelity is maximal. Moreover, as in the previous case, one iteration of the quantum reinforcement protocol is needed to obtain maximal learning fidelity,
$\mathcal{F}_{AE} = |\langle\psi_A|\phi_E\rangle|$.

Quantum reinforcement learning for qudit systems

So far, we have studied quantum reinforcement learning processes only for two-level systems or in pairs of them. However, there are several quantum systems which cannot be described in terms of a two-level system. For instance, quantum harmonic oscillators, electronic energy levels in an ion, and superconducting artificial atoms such as transmons [47], where for some regimes of Josephson energy they must be considered as a three-level system. In this context, it is interesting to extend the quantum reinforcement learning protocol developed here for cases where multilevel systems compound the agent, environment, and register.

To perform the previous task, we first need to define a set of logic operations that we will perform on our system. In the qubit case, the main logical operation applied is the CNOT gate, which considers a conditional interaction between two qubits, where one acts as a control while the other acts as a target. The control qubit remains unchanged whereas the target qubit output is modified by the addition modulo 2. Then, it is wise to assume that the set of logic operations between multilevel systems could be defined in terms of an addition modulo $\mathcal{D}$, where $\mathcal{D}$ stands for the dimension of one subsystem (agent, environment or register subspaces), according to

$$U|i\rangle_1|j\rangle_2 = |i\rangle_1|i \oplus j\rangle_2. \tag{17}$$

Here, $i \oplus j$ stands for the addition modulo $\mathcal{D}$. This gate is usually known as *XOR* gate [48]. For two-dimensional systems, this gate corresponds to the CNOT gate. Nevertheless, for higher dimensional systems this definition presents several disadvantages. For instance, the *XOR* gate defined as in Eq (17) is unitary but not Hermitian for $\mathcal{D} > 2$. Moreover, this logical operation is no longer its own inverse. To avoid these problems, in the literature [48] the generalized XOR gate (GXOR) has been defined as

$$\mathrm{GXOR}_{1,2}|i\rangle_1|j\rangle_2 = |i\rangle_1|i \ominus j\rangle_2, \tag{18}$$

where the operation $\ominus$ denotes the difference $i - j$ *modulo* $\mathcal{D}$. The GXOR gate of Eq (18) does not present the disadvantages pointed out in the definition of Eq (17). That is, the GXOR gate is Hermitian, unitary and $i \ominus j = 0$ only when $i = j$.

Considering our proposed protocol for single-qubit cases, we show that when we take into account multilevel systems, the number of interactions to obtain maximal learning fidelity is fixed and depends only on the number of agent subsystems in the protocol. Let us illustrate this with an example of multilevel agent-environment-register state,

$$|\Psi_0\rangle = \sum_{n=0}^{N-1}\sum_{m=0}^{N-1}\alpha_A^n\alpha_E^m|n\rangle_A|m\rangle_E|0\rangle_1|0\rangle_2. \tag{19}$$

The first step in our protocol is identical to the equivalent one in the single-qubit case. We update the register conditionally on the environment state, that is, we transfer information of the environment and encode it in the register system. This is done by applying a pair of GXOR gates acting in the environment-register subsystem. In this case, the environment interacts with both registers R_1 and R_2. The environment acts as control and both registers are targets,

$$\begin{aligned} |\Psi_1\rangle &= U^{\mathrm{GXOR}}_{(E,R_1)}|\Psi_0\rangle, \\ |\Psi_1\rangle &= \sum_{n=0}^{N-1}\sum_{m=0}^{N-1}\alpha_A^n\alpha_E^m|n\rangle_A|m\rangle_E|m\rangle_1|0\rangle_2. \end{aligned} \tag{20}$$

$$\begin{aligned} |\Psi_2\rangle &= U^{\mathrm{GXOR}}_{(E,R_2)}|\Psi_1\rangle, \\ |\Psi_2\rangle &= \sum_{n=0}^{N-1}\sum_{m=0}^{N-1}\alpha_A^n\alpha_E^m|n\rangle_A|m\rangle_E|m\rangle_1|m\rangle_2. \end{aligned} \tag{21}$$

Once the information has been transferred to the register, we update the register R_1 based on the agent state. That is, we perform a GXOR gate in the subspace composed of agent and register. Here, the agent act as a control and the register R_1 is the target,

$$\begin{aligned} |\Psi_3\rangle &= U^{\mathrm{GXOR}}_{(A,R_1)}|\Psi_2\rangle, \\ |\Psi_3\rangle &= \sum_{n=0}^{N-1}\sum_{m=0}^{N-1}\alpha_A^n\alpha_E^m|n\rangle_A|m\rangle_E|n \ominus m\rangle_1|m\rangle_2. \end{aligned} \tag{22}$$

Orthogonal outcome measurements in the register subspace are provided by interactions between the registers in this subspace. Thus, we apply a GXOR gate in the register subspace, where R_1 is the control and R_2 is the target,

$$\begin{aligned} |\Psi_4\rangle &= U^{\mathrm{GXOR}}_{(R_1,R_2)}|\Psi_3\rangle, \\ |\Psi_4\rangle &= \sum_{n=0}^{N-1}\sum_{m=0}^{N-1}\alpha_A^n\alpha_E^m|n\rangle_A|m\rangle_E|n\ominus m\rangle_1|(n\ominus m)\ominus m\rangle_2. \end{aligned} \tag{23}$$

Subsequently, the agent state is updated conditionally to the information encoded in the state of the register R_1. The GXOR gate is applied in the register-agent subspace. In this case, R_1 is the control and the agent is the target,

$$\begin{aligned} |\Psi_5\rangle &= U^{\mathrm{GXOR}}_{(R_1,A)}|\Psi_4\rangle, \\ |\Psi_5\rangle &= \sum_{n=0}^{N-1}\sum_{m=0}^{N-1}\alpha_A^n\alpha_E^m|0\ominus m\rangle_A|m\rangle_E|n\ominus m\rangle_1|n\ominus 2m\rangle_2. \end{aligned} \tag{24}$$

For the case where the multi-level system contains $\mathcal{D}=2$, we recover the result discussed previously because of $0\ominus m = m$ for that dimension. On the other hand, we are interested in systems with more energy levels, such that we need to adapt the protocol to obtain maximal learning fidelity for a fixed number of steps. In this case, we will update the agent subsystem by an iterative interaction with registers R_1 and R_2 as shown in Fig 3. Here, the agent always acts as target, while the registers are the controls. Therefore, we apply a GXOR gate between the register R_2 and the agent,

$$\begin{aligned} |\Psi_6\rangle &= U^{\mathrm{GXOR}}_{(R_2,A)}|\Psi_5\rangle, \\ |\Psi_6\rangle &= \sum_{n=0}^{N-1}\sum_{m=0}^{N-1}\alpha_A^n\alpha_E^m|n\ominus m\rangle_A|m\rangle_E|n\ominus m\rangle_1|n\ominus 2m\rangle_2. \end{aligned} \tag{25}$$

Now, by applying a GXOR gate between the register R_1 and the agent we obtain,

$$\begin{aligned} |\Psi_7\rangle &= U^{\mathrm{GXOR}}_{(R_1,A)}|\Psi_6\rangle, \\ |\Psi_7\rangle &= \sum_{n=0}^{N-1}\sum_{m=0}^{N-1}\alpha_A^n\alpha_E^m|0\rangle_A|m\rangle_E|n\ominus m\rangle_1|n\ominus 2m\rangle_2. \end{aligned} \tag{26}$$

We perform subsequently a GXOR gate in the subspace composed of R_2 and agent A,

$$\begin{aligned} |\Psi_8\rangle &= U^{\mathrm{GXOR}}_{(R_2,A)}|\Psi_7\rangle, \\ |\Psi_8\rangle &= \sum_{n=0}^{N-1}\sum_{m=0}^{N-1}\alpha_A^n\alpha_E^m|n\ominus 2m\rangle_A|m\rangle_E|n\ominus m\rangle_1|n\ominus 2m\rangle_2. \end{aligned} \tag{27}$$

Finally, applying a GXOR gate on the register-agent subspace we obtain the desired result. By considering a fixed number of interactions between the set of agent, environment and register, the learning fidelity becomes maximal independently of the outcome measurement on the register subspace, which can again be carried out at the end of the protocol,

$$\begin{aligned} |\Psi_9\rangle &= U^{\mathrm{GXOR}}_{(R_1,A)}|\Psi_8\rangle, \\ |\Psi_9\rangle &= \sum_{n=0}^{N-1}\sum_{m=0}^{N-1}\alpha_A^n\alpha_E^m|m\rangle_A|m\rangle_E|n\ominus m\rangle_1|n\ominus 2m\rangle_2. \end{aligned} \tag{28}$$

Thus, in a machine learning protocol where the learning units are composed by multilevel

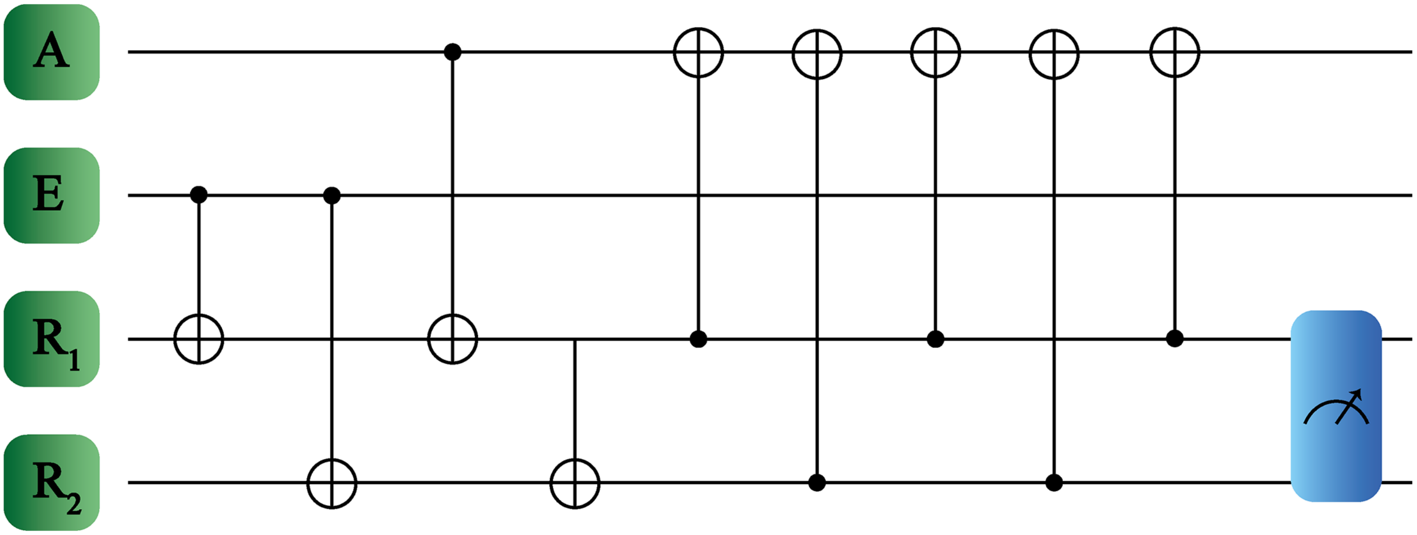

Fig 3. Quantum reinforcement learning protocol for qudits. The systems involved are denoted as agent A, environment E and registers R_1, R_2. In this case, the logical quantum gates which are applied in the learning protocol correspond to GXOR gates. The measurement process in the register subspace is denoted with the rightmost box.

https://doi.org/10.1371/journal.pone.0200455.g003

systems (see Fig 3), the number of logical operations required to obtain maximal learning fidelity does not depend on the system dimension.

Example

Here, we exemplify how our reinforcement learning protocol works in qudit systems. We consider, without loss of generality, the case for dimension $\mathcal{D} = 4$. In this case, the agent-environment-register state has the following form,

$$|A\rangle = \alpha_A^0|0\rangle_A + \alpha_A^1|1\rangle_A + \alpha_A^2|2\rangle_A + \alpha_A^3|3\rangle_A, \tag{29}$$

$$|E\rangle = \alpha_E^0|0\rangle_E + \alpha_E^1|1\rangle_E + \alpha_E^2|2\rangle_E + \alpha_E^3|3\rangle_E \tag{30}$$

$$|R\rangle = |0\rangle_1|0\rangle_2 \tag{31}$$

$$|\Psi\rangle_0 = |A\rangle|E\rangle|R\rangle. \tag{32}$$

As mentioned previously, the considered quantum gate is a GXOR gate with subtraction modulo 4. The first step is to update the register according to the environment information,

$$\begin{aligned} |\Psi\rangle_1 &= U^{\mathrm{GXOR}}_{(E,R_1)}|\Psi\rangle_0, \\ |\Psi\rangle_1 &= (\alpha_A^0|0\rangle_A + \alpha_A^1|1\rangle_A + \alpha_A^2|2\rangle_A + \alpha_A^3|3\rangle_A) \\ &\quad (\alpha_E^0|0\rangle_E|0\rangle_1|0\rangle_2 + \alpha_E^1|1\rangle_E|1\rangle_1|0\rangle_2 + \alpha_E^2|2\rangle_E|2\rangle_1|0\rangle_2 + \alpha_E^3|3\rangle_E|3\rangle_1|0\rangle_2), \end{aligned} \tag{33}$$

$$\begin{aligned} |\Psi\rangle_2 &= U^{\mathrm{GXOR}}_{(E,R_2)}|\Psi\rangle_1, \\ |\Psi\rangle_2 &= (\alpha_A^0|0\rangle_A + \alpha_A^1|1\rangle_A + \alpha_A^2|2\rangle_A + \alpha_A^3|3\rangle_A) \\ &\quad (\alpha_E^0|0\rangle_E|0\rangle_1|0\rangle_2 + \alpha_E^1|1\rangle_E|1\rangle_1|1\rangle_2 + \alpha_E^2|2\rangle_E|2\rangle_1|2\rangle_2 + \alpha_E^3|3\rangle_E|3\rangle_1|3\rangle_2). \end{aligned} \tag{34}$$

Subsequently, the register is updated conditional to the agent state,

$$\begin{aligned}
|\Psi\rangle_3 &= U^{\mathrm{GXOR}}_{(A,R_1)}|\Psi\rangle_2,\\
|\Psi\rangle_3 &= \alpha_A^0\alpha_E^0|0\rangle_A|0\rangle_E|0\rangle_1|0\rangle_2+\alpha_A^0\alpha_E^1|0\rangle_A|1\rangle_E|3\rangle_1|1\rangle_2+\alpha_A^0\alpha_E^2|0\rangle_A|2\rangle_E|2\rangle_1|2\rangle_2\\
&+\alpha_A^0\alpha_E^3|0\rangle_A|3\rangle_E|1\rangle_1|3\rangle_2+\alpha_A^1\alpha_E^0|1\rangle_A|0\rangle_E|1\rangle_1|0\rangle_2+\alpha_A^1\alpha_E^1|1\rangle_A|1\rangle_E|0\rangle_1|1\rangle_2\\
&+\alpha_A^1\alpha_E^2|1\rangle_A|2\rangle_E|3\rangle_1|2\rangle_2+\alpha_A^1\alpha_E^3|1\rangle_A|3\rangle_E|2\rangle_1|3\rangle_2+\alpha_A^2\alpha_E^0|2\rangle_A|0\rangle_E|2\rangle_1|0\rangle_2\\
&+\alpha_A^2\alpha_E^1|2\rangle_A|1\rangle_E|1\rangle_1|1\rangle_2+\alpha_A^2\alpha_E^2|2\rangle_A|2\rangle_E|0\rangle_1|2\rangle_2+\alpha_A^2\alpha_E^3|2\rangle_A|3\rangle_E|3\rangle_1|3\rangle_2\\
&+\alpha_A^3\alpha_E^0|3\rangle_A|0\rangle_E|3\rangle_1|0\rangle_2+\alpha_A^3\alpha_E^1|3\rangle_A|1\rangle_E|2\rangle_1|1\rangle_2+\alpha_A^3\alpha_E^2|3\rangle_A|2\rangle_E|1\rangle_1|2\rangle_2\\
&+\alpha_A^3\alpha_E^3|3\rangle_A|3\rangle_E|0\rangle_1|3\rangle_2.
\end{aligned}\tag{35}$$

Then, to obtain orthogonal outcome measurements in the register basis, we perform an interaction in the register subspace,

$$\begin{aligned}
|\Psi\rangle_4 &= U^{\mathrm{GXOR}}_{(R_1,R_2)}|\Psi\rangle_3,\\
|\Psi\rangle_4 &= \alpha_A^0\alpha_E^0|0\rangle_A|0\rangle_E|0\rangle_1|0\rangle_2+\alpha_A^0\alpha_E^1|0\rangle_A|1\rangle_E|3\rangle_1|2\rangle_2+\alpha_A^0\alpha_E^2|0\rangle_A|2\rangle_E|2\rangle_1|0\rangle_2\\
&+\alpha_A^0\alpha_E^3|0\rangle_A|3\rangle_E|1\rangle_1|2\rangle_2+\alpha_A^1\alpha_E^0|1\rangle_A|0\rangle_E|1\rangle_1|1\rangle_2+\alpha_A^1\alpha_E^1|1\rangle_A|1\rangle_E|0\rangle_1|3\rangle_2\\
&+\alpha_A^1\alpha_E^2|1\rangle_A|2\rangle_E|3\rangle_1|1\rangle_2+\alpha_A^1\alpha_E^3|1\rangle_A|3\rangle_E|2\rangle_1|3\rangle_2+\alpha_A^2\alpha_E^0|2\rangle_A|0\rangle_E|2\rangle_1|2\rangle_2\\
&+\alpha_A^2\alpha_E^1|2\rangle_A|1\rangle_E|1\rangle_1|0\rangle_2+\alpha_A^2\alpha_E^2|2\rangle_A|2\rangle_E|0\rangle_1|2\rangle_2+\alpha_A^2\alpha_E^3|2\rangle_A|3\rangle_E|3\rangle_1|0\rangle_2\\
&+\alpha_A^3\alpha_E^0|3\rangle_A|0\rangle_E|3\rangle_1|3\rangle_2+\alpha_A^3\alpha_E^1|3\rangle_A|1\rangle_E|2\rangle_1|1\rangle_2+\alpha_A^3\alpha_E^2|3\rangle_A|2\rangle_E|1\rangle_1|3\rangle_2\\
&+\alpha_A^3\alpha_E^3|3\rangle_A|3\rangle_E|0\rangle_1|1\rangle_2.
\end{aligned}\tag{36}$$

Now, we need to apply iterative interactions in the register-agent subspace to update the agent in each step until we get maximal learning fidelity with respect to the environment. We start by performing a GXOR gate between the register R_1 and the agent,

$$\begin{aligned}
|\Psi\rangle_5 &= U^{\mathrm{GXOR}}_{(R_1,A)}|\Psi\rangle_4,\\
|\Psi\rangle_5 &= \alpha_A^0\alpha_E^0|0\rangle_A|0\rangle_E|0\rangle_1|0\rangle_2+\alpha_A^0\alpha_E^1|3\rangle_A|1\rangle_E|3\rangle_1|2\rangle_2+\alpha_A^0\alpha_E^2|2\rangle_A|2\rangle_E|2\rangle_1|0\rangle_2\\
&+\alpha_A^0\alpha_E^3|1\rangle_A|3\rangle_E|1\rangle_1|2\rangle_2+\alpha_A^1\alpha_E^0|0\rangle_A|0\rangle_E|1\rangle_1|1\rangle_2+\alpha_A^1\alpha_E^1|3\rangle_A|1\rangle_E|0\rangle_1|3\rangle_2\\
&+\alpha_A^1\alpha_E^2|2\rangle_A|2\rangle_E|3\rangle_1|1\rangle_2+\alpha_A^1\alpha_E^3|1\rangle_A|3\rangle_E|2\rangle_1|3\rangle_2+\alpha_A^2\alpha_E^0|0\rangle_A|0\rangle_E|2\rangle_1|2\rangle_2\\
&+\alpha_A^2\alpha_E^1|3\rangle_A|1\rangle_E|1\rangle_1|0\rangle_2+\alpha_A^2\alpha_E^2|2\rangle_A|2\rangle_E|0\rangle_1|2\rangle_2+\alpha_A^2\alpha_E^3|1\rangle_A|3\rangle_E|3\rangle_1|0\rangle_2\\
&+\alpha_A^3\alpha_E^0|0\rangle_A|0\rangle_E|3\rangle_1|3\rangle_2+\alpha_A^3\alpha_E^1|3\rangle_A|1\rangle_E|2\rangle_1|1\rangle_2+\alpha_A^3\alpha_E^2|2\rangle_A|2\rangle_E|1\rangle_1|3\rangle_2\\
&+\alpha_A^3\alpha_E^3|1\rangle_A|3\rangle_E|0\rangle_1|1\rangle_2.
\end{aligned}\tag{37}$$

Hereafter, we apply the GXOR gate in the R_2-agent subspace,

$$\begin{aligned}
|\Psi\rangle_6 &= U^{\mathrm{GXOR}}_{(R_2,A)}|\Psi\rangle_5,\\
|\Psi\rangle_6 &= \alpha_A^0\alpha_E^0|0\rangle_A|0\rangle_E|0\rangle_1|0\rangle_2+\alpha_A^0\alpha_E^1|3\rangle_A|1\rangle_E|3\rangle_1|2\rangle_2+\alpha_A^0\alpha_E^2|2\rangle_A|2\rangle_E|2\rangle_1|0\rangle_2\\
&+\alpha_A^0\alpha_E^3|1\rangle_A|3\rangle_E|1\rangle_1|2\rangle_2+\alpha_A^1\alpha_E^0|1\rangle_A|0\rangle_E|1\rangle_1|1\rangle_2+\alpha_A^1\alpha_E^1|0\rangle_A|1\rangle_E|0\rangle_1|3\rangle_2\\
&+\alpha_A^1\alpha_E^2|3\rangle_A|2\rangle_E|3\rangle_1|1\rangle_2+\alpha_A^1\alpha_E^3|2\rangle_A|3\rangle_E|2\rangle_1|3\rangle_2+\alpha_A^2\alpha_E^0|2\rangle_A|0\rangle_E|2\rangle_1|2\rangle_2\\
&+\alpha_A^2\alpha_E^1|1\rangle_A|1\rangle_E|1\rangle_1|0\rangle_2+\alpha_A^2\alpha_E^2|0\rangle_A|2\rangle_E|0\rangle_1|2\rangle_2+\alpha_A^2\alpha_E^3|3\rangle_A|3\rangle_E|3\rangle_1|0\rangle_2\\
&+\alpha_A^3\alpha_E^0|3\rangle_A|0\rangle_E|3\rangle_1|3\rangle_2+\alpha_A^3\alpha_E^1|2\rangle_A|1\rangle_E|2\rangle_1|1\rangle_2+\alpha_A^3\alpha_E^2|1\rangle_A|2\rangle_E|1\rangle_1|3\rangle_2\\
&+\alpha_A^3\alpha_E^3|0\rangle_A|3\rangle_E|0\rangle_1|1\rangle_2.
\end{aligned}\tag{38}$$

Afterwards, we perform a GXOR gate between R_1 and A,

$$\begin{aligned}
|\Psi\rangle_7 &= U^{\mathrm{GXOR}}_{(R_1,A)}|\Psi\rangle_6, \\
|\Psi\rangle_7 &= \alpha_A^0\alpha_E^0|0\rangle_A|0\rangle_E|0\rangle_1|0\rangle_2 + \alpha_A^0\alpha_E^1|0\rangle_A|1\rangle_E|3\rangle_1|2\rangle_2 + \alpha_A^0\alpha_E^2|0\rangle_A|2\rangle_E|2\rangle_1|0\rangle_2 \\
&\quad +\alpha_A^0\alpha_E^3|0\rangle_A|3\rangle_E|1\rangle_1|2\rangle_2 + \alpha_A^1\alpha_E^0|0\rangle_A|0\rangle_E|1\rangle_1|1\rangle_2 + \alpha_A^1\alpha_E^1|0\rangle_A|1\rangle_E|0\rangle_1|3\rangle_2 \\
&\quad +\alpha_A^1\alpha_E^2|0\rangle_A|2\rangle_E|3\rangle_1|1\rangle_2 + \alpha_A^1\alpha_E^3|0\rangle_A|3\rangle_E|2\rangle_1|3\rangle_2 + \alpha_A^2\alpha_E^0|0\rangle_A|0\rangle_E|2\rangle_1|2\rangle_2 \\
&\quad +\alpha_A^2\alpha_E^1|0\rangle_A|1\rangle_E|1\rangle_1|0\rangle_2 + \alpha_A^2\alpha_E^2|0\rangle_A|2\rangle_E|0\rangle_1|2\rangle_2 + \alpha_A^2\alpha_E^3|0\rangle_A|3\rangle_E|3\rangle_1|0\rangle_2 \\
&\quad +\alpha_A^3\alpha_E^0|0\rangle_A|0\rangle_E|3\rangle_1|3\rangle_2 + \alpha_A^3\alpha_E^1|0\rangle_A|1\rangle_E|2\rangle_1|1\rangle_2 + \alpha_A^3\alpha_E^2|0\rangle_A|2\rangle_E|1\rangle_1|3\rangle_2 \\
&\quad +\alpha_A^3\alpha_E^3|0\rangle_A|3\rangle_E|0\rangle_1|1\rangle_2.
\end{aligned} \tag{39}$$

Subsequently, an interaction in the R_2-agent subspace is performed,

$$\begin{aligned}
|\Psi\rangle_8 &= U^{\mathrm{GXOR}}_{(R_2,A)}|\Psi\rangle_7, \\
|\Psi\rangle_8 &= \alpha_A^0\alpha_E^0|0\rangle_A|0\rangle_E|0\rangle_1|0\rangle_2 + \alpha_A^0\alpha_E^1|2\rangle_A|1\rangle_E|3\rangle_1|2\rangle_2 + \alpha_A^0\alpha_E^2|0\rangle_A|2\rangle_E|2\rangle_1|0\rangle_2 \\
&\quad +\alpha_A^0\alpha_E^3|2\rangle_A|3\rangle_E|1\rangle_1|2\rangle_2 + \alpha_A^1\alpha_E^0|1\rangle_A|0\rangle_E|1\rangle_1|1\rangle_2 + \alpha_A^1\alpha_E^1|3\rangle_A|1\rangle_E|0\rangle_1|3\rangle_2 \\
&\quad +\alpha_A^1\alpha_E^2|1\rangle_A|2\rangle_E|3\rangle_1|1\rangle_2 + \alpha_A^1\alpha_E^3|3\rangle_A|3\rangle_E|2\rangle_1|3\rangle_2 + \alpha_A^2\alpha_E^0|2\rangle_A|0\rangle_E|2\rangle_1|2\rangle_2 \\
&\quad +\alpha_A^2\alpha_E^1|0\rangle_A|1\rangle_E|1\rangle_1|0\rangle_2 + \alpha_A^2\alpha_E^2|2\rangle_A|2\rangle_E|0\rangle_1|2\rangle_2 + \alpha_A^2\alpha_E^3|0\rangle_A|3\rangle_E|3\rangle_1|0\rangle_2 \\
&\quad +\alpha_A^3\alpha_E^0|3\rangle_A|0\rangle_E|3\rangle_1|3\rangle_2 + \alpha_A^3\alpha_E^1|1\rangle_A|1\rangle_E|2\rangle_1|1\rangle_2 + \alpha_A^3\alpha_E^2|3\rangle_A|2\rangle_E|1\rangle_1|3\rangle_2 \\
&\quad +\alpha_A^3\alpha_E^3|1\rangle_A|3\rangle_E|0\rangle_1|1\rangle_2.
\end{aligned} \tag{40}$$

Finally, we apply a GXOR gate between R_1 and the agent,

$$\begin{aligned}
|\Psi\rangle_9 &= U^{\mathrm{GXOR}}_{(R_1,A)}|\Psi\rangle_8, \\
|\Psi\rangle_9 &= \alpha_A^0\alpha_E^0|0\rangle_A|0\rangle_E|0\rangle_1|0\rangle_2 + \alpha_A^0\alpha_E^1|1\rangle_A|1\rangle_E|3\rangle_1|2\rangle_2 + \alpha_A^0\alpha_E^2|2\rangle_A|2\rangle_E|2\rangle_1|0\rangle_2 \\
&\quad +\alpha_A^0\alpha_E^3|3\rangle_A|3\rangle_E|1\rangle_1|2\rangle_2 + \alpha_A^1\alpha_E^0|0\rangle_A|0\rangle_E|1\rangle_1|1\rangle_2 + \alpha_A^1\alpha_E^1|1\rangle_A|1\rangle_E|0\rangle_1|3\rangle_2 \\
&\quad +\alpha_A^1\alpha_E^2|2\rangle_A|2\rangle_E|3\rangle_1|1\rangle_2 + \alpha_A^1\alpha_E^3|3\rangle_A|3\rangle_E|2\rangle_1|3\rangle_2 + \alpha_A^2\alpha_E^0|0\rangle_A|0\rangle_E|2\rangle_1|2\rangle_2 \\
&\quad +\alpha_A^2\alpha_E^1|1\rangle_A|1\rangle_E|1\rangle_1|0\rangle_2 + \alpha_A^2\alpha_E^2|2\rangle_A|2\rangle_E|0\rangle_1|2\rangle_2 + \alpha_A^2\alpha_E^3|3\rangle_A|3\rangle_E|3\rangle_1|0\rangle_2 \\
&\quad +\alpha_A^3\alpha_E^0|0\rangle_A|0\rangle_E|3\rangle_1|3\rangle_2 + \alpha_A^3\alpha_E^1|1\rangle_A|1\rangle_E|2\rangle_1|1\rangle_2 + \alpha_A^3\alpha_E^2|2\rangle_A|2\rangle_E|1\rangle_1|3\rangle_2 \\
&\quad +\alpha_A^3\alpha_E^3|3\rangle_A|3\rangle_E|0\rangle_1|1\rangle_2.
\end{aligned} \tag{41}$$

As we can see, based in the quantum protocol described previously (see Fig 3), we have shown that for a fixed number of interactions, we obtain maximal learning fidelity even though the system has an arbitrary dimension.

Quantum reinforcement learning in multiqudit systems

In the previous section, we proved that for an agent and environment composed of a multilevel system each, the quantum reinforcement learning protocol entails maximal learning fidelity for a fixed number of steps, irrespective of the dimension. Here, using this result, we also prove that for more than one multilevel system in agent, environment, and register subspaces, the number of steps is also fixed and scales with the number of individual subsystems that compose both agent and environment subsystems. To be more specific, in the single-multilevel case the needed total steps are nine. For two multilevel systems, we show that the number of required steps are eighteen, and in general, $9n$, with n being the number of multilevel subsystems. The possible initial states of our protocol consist in arbitrary superpositions for both agent and environment states and the register states are in their ground state,

$$|\Psi_0\rangle = \sum_{n,m=0}^{N-1}\sum_{p,q=0}^{N-1}\alpha_A^{nm}\alpha_E^{pq}|n\rangle_A|m\rangle_A|p\rangle_E|q\rangle_E|0\rangle_1|0\rangle_2|0\rangle_3|0\rangle_4. \tag{42}$$

The first step in the protocol consists in encoding the environment information in the register states. This is done by applying a pair of GXOR gates. The gates are applied in the environment-register subspace, while the interaction in this case is the same as the one described previously. Namely, E_1 controls R_1 and E_2 controls R_2.

$$\begin{aligned} |\Psi_1\rangle &= U^{\mathrm{GXOR}}_{(E_2,R_2)}U^{\mathrm{GXOR}}_{(E_1,R_1)}|\Psi_0\rangle, \\ |\Psi_1\rangle &= \sum_{n,m=0}^{N-1}\sum_{p,q=0}^{N-1}\alpha_A^{nm}\alpha_E^{pq}|n\rangle_A|m\rangle_A|p\rangle_E|q\rangle_E|p\rangle_1|q\rangle_2|0\rangle_3|0\rangle_4. \end{aligned} \tag{43}$$

Similarly, in the second step we encode the environment information in the other two registers (R_3 and R_4) through GXOR gates. Here, the control system is the environment while the targets are the registers.

$$\begin{aligned} |\Psi_2\rangle &= U^{\mathrm{GXOR}}_{(E_2,R_4)}U^{\mathrm{GXOR}}_{(E_1,R_3)}|\Psi_1\rangle, \\ |\Psi_2\rangle &= \sum_{n,m=0}^{N-1}\sum_{p,q=0}^{N-1}\alpha_A^{nm}\alpha_E^{pq}|n\rangle_A|m\rangle_A|p\rangle_E|q\rangle_E|p\rangle_1|q\rangle_2|p\rangle_3|q\rangle_4. \end{aligned} \tag{44}$$

Subsequently, a part of the register subspace is updated conditional on the agent information. Therefore, we apply a pair of GXOR gates on the agent-register subspace. In this case, agents A_1 and A_2 are controls and registers R_1 and R_2 targets.

$$\begin{aligned} |\Psi_3\rangle &= U^{\mathrm{GXOR}}_{(A_2,R_2)}U^{\mathrm{GXOR}}_{(A_1,R_1)}|\Psi_2\rangle, \\ |\Psi_3\rangle &= \sum_{n,m=0}^{N-1}\sum_{p,q=0}^{N-1}\alpha_A^{nm}\alpha_E^{pq}|n\rangle_A|m\rangle_A|p\rangle_E|q\rangle_E|n\ominus p\rangle_1|m\ominus q\rangle_2|p\rangle_3|q\rangle_4. \end{aligned} \tag{45}$$

Now, we update the register subspace considering interactions between register components which have been acted upon with the same part of the environment. Namely, the register R_3 will be updated with the control of R_1 (Similarly with R_4 being controlled with R_2).

$$\begin{aligned} |\Psi_4\rangle &= U^{\mathrm{GXOR}}_{(R_2,R_4)}U^{\mathrm{GXOR}}_{(R_1,R_3)}|\Psi_3\rangle, \\ |\Psi_4\rangle &= \sum_{n,m=0}^{N-1}\sum_{p,q=0}^{N-1}\alpha_A^{nm}\alpha_E^{pq}|n\rangle_A|m\rangle_A|p\rangle_E|q\rangle_E|n\ominus p\rangle_1|m\ominus q\rangle_2|n\ominus 2p\rangle_3|m\ominus 2q\rangle_4. \end{aligned} \tag{46}$$

Subsequently, we need to apply successive interactions between agent states and register states to obtain maximal learning fidelity. We show that applying the same interactions as for the single multilevel case for the triplet formed by agent A_1 with the environment parts R_1 and R_3 (similarly A_2 with R_2 and R_4), the maximal learning fidelity is reached. It is straightforward to show that

$$\begin{aligned} |\Psi_9\rangle &= U^{\mathrm{GXOR}}_{(R_2,A_2)} U^{\mathrm{GXOR}}_{(R_1,A_1)} U^{\mathrm{GXOR}}_{(R_4,A_2)} U^{\mathrm{GXOR}}_{(R_3,A_1)} \times \\ &\quad U^{\mathrm{GXOR}}_{(R_2,A_2)} U^{\mathrm{GXOR}}_{(R_1,A_1)} U^{\mathrm{GXOR}}_{(R_4,A_2)} U^{\mathrm{GXOR}}_{(R_3,A_1)} U^{\mathrm{GXOR}}_{(R_2,A_2)} \times \\ &\quad U^{\mathrm{GXOR}}_{(R_1,A_1)} |\Psi_4\rangle, \\ |\Psi_9\rangle &= \sum_{n,m=0}^{N-1} \sum_{p,q=0}^{N-1} \alpha_A^{nm} \alpha_E^{pq} |p\rangle_A |q\rangle_A |p\rangle_E |q\rangle_E |n \ominus p\rangle_1 |m \ominus q\rangle_2 |n \ominus 2p\rangle_3 |m \ominus 2q\rangle_4. \end{aligned} \tag{47}$$

Summarizing, for the case studied in this section, we demonstrate that the number of operations required to obtain maximal learning fidelity does not depend on the learning unit dimension and it is equal to eighteen operations, which correspond to the double of the required steps in the single multiqubit case. It is straightforward to realize that the number of needed operations to achieve maximal learning fidelity in a machine learning protocol composed by n subsystems for agent and environment is equal to $9n$. Namely, the number of operations scales polynomially, indeed linearly, with the number of subsystems.

Quantum reinforcement learning in larger environments

Up to now, the quantum reinforcement learning protocol described here always considers that the agent and the environment have the same number of subsystems, as well as the same dimension. In these cases, we have shown that by adding more system registers the quantum protocol improves in the sense that only one iteration and one measurement is enough to obtain maximal learning fidelity. Nevertheless, in more realistic scenarios, the agent must adapt to larger or more complex surroundings. Here, we discuss the situation where the environment has more subsystems than the agent, and therefore a larger dimension. As the environment has more information than the agent, it is expect that not all available surrounding information will be transferred to the agent. Indeed, we prove that by depending on the register-environment interaction, the agent can encode the information from one specific part of the environment. In this case, unlike the protocol previously discussed, we achieve maximal learning fidelity after applying one measurement and a rewarding iteration (feedback).

The proposed quantum protocol is shown in Fig 4. Here, one two-level system forms the agent, while register and environment are constituted each by two qubits. Each environment qubit interacts with one qubit from the register, such that this interaction updates the registers conditionally to the environment information. Then, one part of the register subspace is also upgraded conditionally to the agent state. Subsequently, we perform a measurement on the register subspace, such that depending on the measurement outcomes we apply a conditional operation in the agent-register subspace until the agent adapts to a specific part of the environment. To illustrate this, let us introduce a possible agent-register-subspace state which has the following form,

$$|A\rangle = \alpha_A^0 |0\rangle_A + \alpha_A^1 |1\rangle_A \tag{48}$$

$$|E\rangle = \alpha_E^{00} |00\rangle_E + \alpha_E^{01} |01\rangle_E + \alpha_E^{10} |10\rangle_E + \alpha_E^{11} |11\rangle_E \tag{49}$$

$$|R\rangle = |0\rangle_1 |0\rangle_2, \tag{50}$$

$$|\Psi\rangle_0 = |A\rangle |E\rangle |R\rangle. \tag{51}$$

The first step is to transfer quantum information from the environment onto the registers.

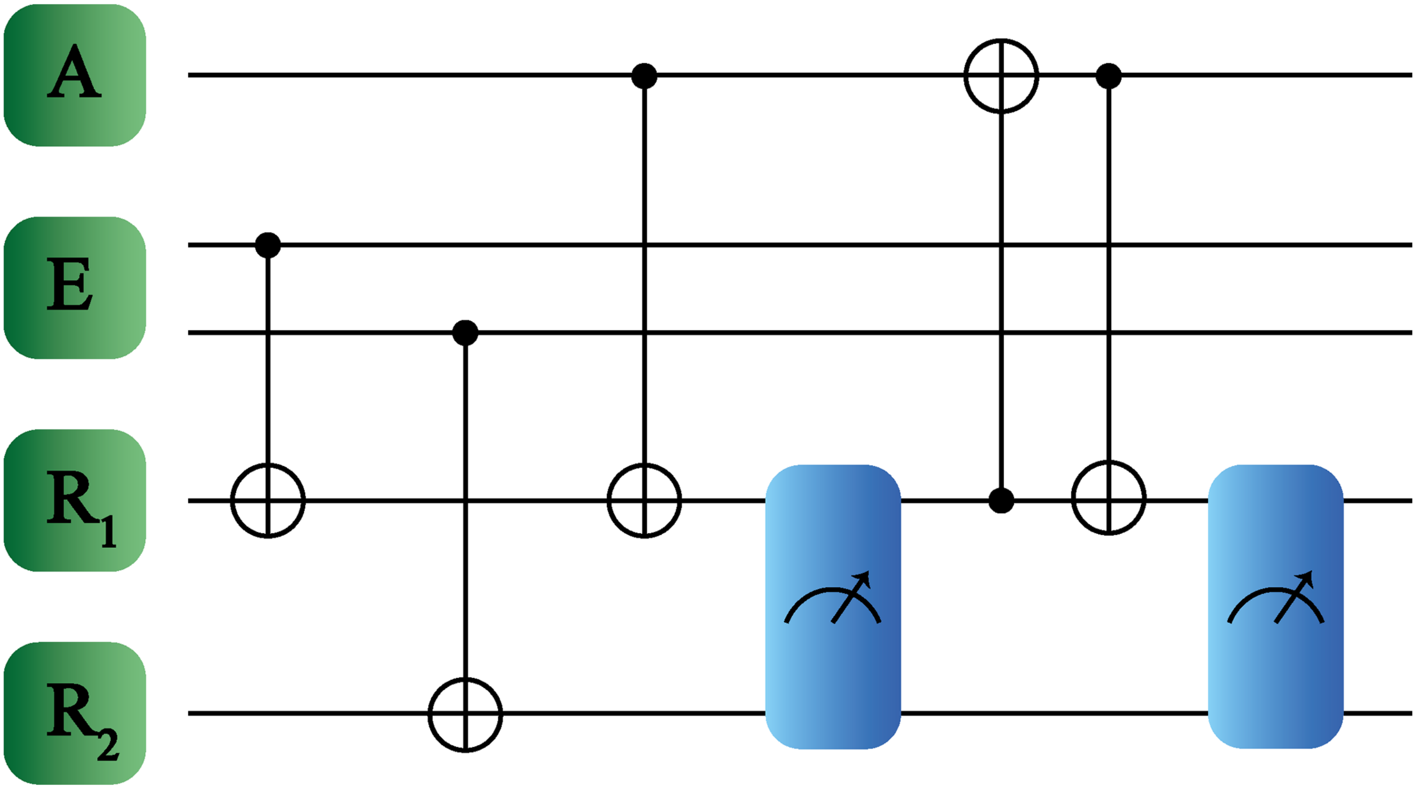

Fig 4. Quantum reinforcement learning for larger environment systems. The systems involved are denoted as agent A, environment E and registers R_1, R_2, where E contains now two qubits while A just one. The logical gates applied between the different subsystems are CNOT gates. In this case, to obtain maximal learning fidelity, it is required to perform two separate measurements denoted by the blue boxes.

https://doi.org/10.1371/journal.pone.0200455.g004

This is done by applying a pair of CNOT gates in the environment-register subspaces,

$$\begin{aligned} |\Psi\rangle_1 &= U^{\text{CNOT}}_{(E,R_2)} U^{\text{CNOT}}_{(E,R_1)} |\Psi\rangle_0, \\ |\Psi\rangle_1 &= (\alpha_A^0 |0\rangle_A + \alpha_A^1 |1\rangle_A) \\ &\quad (\alpha_E^{00}|00\rangle_E|0\rangle_1|0\rangle_2 + \alpha_E^{01}|01\rangle_E|0\rangle_1|1\rangle_2 + \alpha_E^{10}|10\rangle_E|1\rangle_1|0\rangle_2 + \alpha_E^{11}|11\rangle_E|1\rangle_1|1\rangle_2). \end{aligned} \tag{52}$$

Subsequently, the register R_1 is updated conditionally to the agent information. Therefore, a CNOT gate is applied in the agent-register subspace, where the agent qubit is the control and the register R_1 is the target,

$$\begin{aligned} |\Psi\rangle_2 &= U^{\text{CNOT}}_{(A,R_1)} |\Psi\rangle_1, \\ |\Psi\rangle_2 &= \alpha_A^0\alpha_E^{00}|0\rangle_A|00\rangle_E|0\rangle_1|0\rangle_2 + \alpha_A^0\alpha_E^{01}|0\rangle_A|01\rangle_E|0\rangle_1|1\rangle_2 \\ &\quad + \alpha_A^0\alpha_E^{10}|0\rangle_A|10\rangle_E|1\rangle_1|0\rangle_2 + \alpha_A^0\alpha_E^{11}|0\rangle_A|11\rangle_E|1\rangle_1|1\rangle_2 \\ &\quad + \alpha_A^1\alpha_E^{00}|1\rangle_A|00\rangle_E|1\rangle_1|0\rangle_2 + \alpha_A^1\alpha_E^{01}|1\rangle_A|01\rangle_E|1\rangle_1|1\rangle_2 \\ &\quad + \alpha_A^1\alpha_E^{10}|1\rangle_A|10\rangle_E|0\rangle_1|0\rangle_2 + \alpha_A^1\alpha_E^{11}|1\rangle_A|11\rangle_E|0\rangle_1|1\rangle_2. \end{aligned} \tag{53}$$

Afterwards, we perform a measurement on the register subspace. In this case, the wave function is projected into the four possible measurement outcomes,

$$\begin{aligned}
\mathrm{M}_1 &= (\alpha_A^0\alpha_E^{00}|0\rangle_A|00\rangle_E + \alpha_A^1\alpha_E^{10}|1\rangle_A|10\rangle_E)|0\rangle_1|0\rangle_2\\
&= (\alpha_A^0\alpha_E^{00}|0\rangle_A|0\rangle_{E_1} + \alpha_A^1\alpha_E^{10}|1\rangle_A|1\rangle_{E_1})|0\rangle_{E_2}|0\rangle_1|0\rangle_2,\\
\mathrm{M}_2 &= (\alpha_A^0\alpha_E^{01}|0\rangle_A|01\rangle_E + \alpha_A^1\alpha_E^{11}|1\rangle_A|11\rangle_E)|0\rangle_1|1\rangle_2\\
&= (\alpha_A^0\alpha_E^{01}|0\rangle_A|0\rangle_{E_1} + \alpha_A^1\alpha_E^{11}|1\rangle_A|1\rangle_{E_1})|1\rangle_{E_2}|0\rangle_1|1\rangle_2,\\
\mathrm{M}_3 &= (\alpha_A^1\alpha_E^{00}|1\rangle_A|00\rangle_E + \alpha_A^0\alpha_E^{10}|0\rangle_A|10\rangle_E)|1\rangle_1|0\rangle_2\\
&= (\alpha_A^1\alpha_E^{00}|1\rangle_A|0\rangle_{E_1} + \alpha_A^0\alpha_E^{10}|0\rangle_A|1\rangle_{E_1})|0\rangle_{E_2}|1\rangle_1|0\rangle_2,\\
\mathrm{M}_4 &= (\alpha_A^0\alpha_E^{11}|0\rangle_A|11\rangle_E + \alpha_A^1\alpha_E^{01}|1\rangle_A|01\rangle_E)|1\rangle_1|1\rangle_2\\
&= (\alpha_A^0\alpha_E^{11}|0\rangle_A|1\rangle_{E_1} + \alpha_A^1\alpha_E^{01}|1\rangle_A|0\rangle_{E_1})|1\rangle_{E_2}|1\rangle_1|1\rangle_2.
\end{aligned} \tag{54}$$

As we can see, the projective measurement on the register subspace produces that agent and one part of the environment subspace (E_1) is in an entangled state. At this stage, we can apply the rewarding criterion which consists in performing a CNOT gate operation in the register-agent subspace. The register qubit R_1 is the control and the agent is the target,

$$\begin{aligned}
\mathrm{M}_{1a} = \mathrm{U}^{\mathrm{CNOT}}_{(\mathrm{R}_1,\mathrm{A})}\mathrm{M}_1 = (\alpha_\mathrm{A}^0\alpha_\mathrm{E}^{00}|0\rangle_\mathrm{A}|0\rangle_{\mathrm{E}_1} + \alpha_\mathrm{A}^1\alpha_\mathrm{E}^{10}|1\rangle_\mathrm{A}|1\rangle_{\mathrm{E}_1})|0\rangle_{\mathrm{E}_2}|0\rangle_1|0\rangle_2,\\
\mathrm{M}_{2a} = \mathrm{U}^{\mathrm{CNOT}}_{(\mathrm{R}_1,\mathrm{A})}\mathrm{M}_2 = (\alpha_\mathrm{A}^0\alpha_\mathrm{E}^{01}|0\rangle_\mathrm{A}|0\rangle_{\mathrm{E}_1} + \alpha_\mathrm{A}^1\alpha_\mathrm{E}^{11}|1\rangle_\mathrm{A}|1\rangle_{\mathrm{E}_1})|1\rangle_{\mathrm{E}_2}|0\rangle_1|1\rangle_2,\\
\mathrm{M}_{3a} = \mathrm{U}^{\mathrm{CNOT}}_{(\mathrm{R}_1,\mathrm{A})}\mathrm{M}_3 = (\alpha_\mathrm{A}^1\alpha_\mathrm{E}^{00}|0\rangle_\mathrm{A}|0\rangle_{\mathrm{E}_1} + \alpha_\mathrm{A}^0\alpha_\mathrm{E}^{10}|1\rangle_\mathrm{A}|1\rangle_{\mathrm{E}_1})|0\rangle_{\mathrm{E}_2}|1\rangle_1|0\rangle_2,\\
\mathrm{M}_{4a} = \mathrm{U}^{\mathrm{CNOT}}_{(\mathrm{R}_1,\mathrm{A})}\mathrm{M}_4 = (\alpha_\mathrm{A}^0\alpha_\mathrm{E}^{11}|1\rangle_\mathrm{A}|1\rangle_{\mathrm{E}_1} + \alpha_\mathrm{A}^1\alpha_\mathrm{E}^{01}|0\rangle_\mathrm{A}|0\rangle_{\mathrm{E}_1})|1\rangle_{\mathrm{E}_2}|1\rangle_1|1\rangle_2.
\end{aligned} \tag{55}$$

Finally, we perform a CNOT gate in the agent-register subspace to obtain orthogonal measurement outcomes. The qubit agent is the control and the qubit register R_1 is the target, according to

$$\begin{aligned}
\mathrm{M}_{1b} = \mathrm{U}^{\mathrm{CNOT}}_{(\mathrm{A},\mathrm{R}_1)}\mathrm{M}_{1a} = \alpha_\mathrm{A}^0\alpha_\mathrm{E}^{00}|0\rangle_\mathrm{A}|00\rangle_\mathrm{E}|0\rangle_1|0\rangle_2 + \alpha_\mathrm{A}^1\alpha_\mathrm{E}^{10}|1\rangle_\mathrm{A}|10\rangle_\mathrm{E}|1\rangle_1|0\rangle_2,\\
\mathrm{M}_{2b} = \mathrm{U}^{\mathrm{CNOT}}_{(\mathrm{A},\mathrm{R}_1)}\mathrm{M}_{2a} = \alpha_\mathrm{A}^0\alpha_\mathrm{E}^{01}|0\rangle_\mathrm{A}|01\rangle_\mathrm{E}|0\rangle_1|1\rangle_2 + \alpha_\mathrm{A}^1\alpha_\mathrm{E}^{11}|1\rangle_\mathrm{A}|11\rangle_\mathrm{E}|1\rangle_1|1\rangle_2,\\
\mathrm{M}_{3b} = \mathrm{U}^{\mathrm{CNOT}}_{(\mathrm{A},\mathrm{R}_1)}\mathrm{M}_{3a} = \alpha_\mathrm{A}^1\alpha_\mathrm{E}^{00}|0\rangle_\mathrm{A}|00\rangle_\mathrm{E}|1\rangle_1|0\rangle_2 + \alpha_\mathrm{A}^0\alpha_\mathrm{E}^{10}|1\rangle_\mathrm{A}|10\rangle_\mathrm{E}|0\rangle_1|0\rangle_2,\\
\mathrm{M}_{4b} = \mathrm{U}^{\mathrm{CNOT}}_{(\mathrm{A},\mathrm{R}_1)}\mathrm{M}_{4a} = \alpha_\mathrm{A}^1\alpha_\mathrm{E}^{01}|0\rangle_\mathrm{A}|01\rangle_\mathrm{E}|1\rangle_1|1\rangle_2 + \alpha_\mathrm{A}^0\alpha_\mathrm{E}^{11}|1\rangle_\mathrm{A}|11\rangle_\mathrm{E}|0\rangle_1|1\rangle_2.
\end{aligned} \tag{56}$$

In this quantum reinforcement learning protocol, we perform interactions between the environment and the register subspaces. Nevertheless, the agent is updated only regarding the information encoded in register R_1. Thus, the maximal learning fidelity is achieved with respect to the first qubit of the environment.

Let us now consider another configuration similar to the one studied previously in this article, where the register is formed by a larger number of subsystems than the environment. Here, additionally, the environment we consider is larger than the agent. We prove that, for this system configuration, maximal learning fidelity between the agent and one part of the environment is achieved in one rewarding process. For this configuration, the maximal fidelity does not depend on the entanglement present in the agent-environment subspace. The general agent-register-environment state is

$$|A\rangle = \alpha_A^0|0\rangle_A + \alpha_A^1|1\rangle_A, \tag{57}$$

$$|E\rangle = (\alpha_E^0|0\rangle_{E_1} + \alpha_E^1|1\rangle_{E_1})|0\rangle_{E_2} + (\beta_E^0|0\rangle_{E_1} + \beta_E^1|1\rangle_{E_1})|1\rangle_{E_2}, \tag{58}$$

$$|R\rangle = |0\rangle_1|0\rangle_2|0\rangle_3|0\rangle_4, \tag{59}$$

$$|\Psi\rangle_0 = |A\rangle|E\rangle|R\rangle. \tag{60}$$

The quantum protocol consists in updating the registers $R_{1,2}$ conditionally to the environment state $E_{1,2}$,

$$\begin{aligned}
|\Psi\rangle_1 &= U^{\text{CNOT}}_{(E_2,R_2)}U^{\text{CNOT}}_{(E_1,R_1)}|\Psi\rangle_0,\\
|\Psi\rangle_1 &= (\alpha_A^0|0\rangle_A + \alpha_A^1|1\rangle_A)(\alpha_E^0|0\rangle_{E_1}|0\rangle_{E_2}|0\rangle_1|0\rangle_2|0\rangle_3|0\rangle_4 + \alpha_E^1|1\rangle_{E_1}|0\rangle_{E_2}|1\rangle_1|0\rangle_2|0\rangle_3|0\rangle_4\\
&\quad +\beta_E^0|0\rangle_{E_1}|1\rangle_{E_2}|0\rangle_1|1\rangle_2|0\rangle_3|0\rangle_4 + \beta_E^1|1\rangle_{E_1}|1\rangle_{E_2}|1\rangle_1|1\rangle_2|0\rangle_3|0\rangle_4).
\end{aligned} \tag{61}$$

After this, we also update the information of the registers $R_{3,4}$ conditionally to the environment state $E_{1,2}$,

$$\begin{aligned}
|\Psi\rangle_2 &= U^{\text{CNOT}}_{(E_2,R_4)}U^{\text{CNOT}}_{(E_1,R_3)}|\Psi\rangle_1,\\
|\Psi\rangle_2 &= (\alpha_A^0|0\rangle_A + \alpha_A^1|1\rangle_A)(\alpha_E^0|0\rangle_{E_1}|0\rangle_{E_2}|0\rangle_1|0\rangle_2|0\rangle_3|0\rangle_4 + \alpha_E^1|1\rangle_{E_1}|0\rangle_{E_2}|1\rangle_1|0\rangle_2|1\rangle_3|0\rangle_4\\
&\quad +\beta_E^0|0\rangle_{E_1}|1\rangle_{E_2}|0\rangle_1|1\rangle_2|0\rangle_3|1\rangle_4 + \beta_E^1|1\rangle_{E_1}|1\rangle_{E_2}|1\rangle_1|1\rangle_2|1\rangle_3|1\rangle_4).
\end{aligned} \tag{62}$$

Now, the register R_1 is updated conditionally to the agent state,

$$\begin{aligned}
|\Psi\rangle_3 &= U^{\text{CNOT}}_{(A,R_1)}|\Psi\rangle_2,\\
|\Psi\rangle_3 &= \alpha_A^0\alpha_E^0|0\rangle_A|0\rangle_{E_1}|0\rangle_{E_2}|0\rangle_1|0\rangle_2|0\rangle_3|0\rangle_4 + \alpha_A^0\alpha_E^1|0\rangle_A|1\rangle_{E_1}|0\rangle_{E_2}|1\rangle_1|0\rangle_2|1\rangle_3|0\rangle_4\\
&\quad +\alpha_A^0\beta_E^0|0\rangle_A|0\rangle_{E_1}|1\rangle_{E_2}|0\rangle_1|1\rangle_2|0\rangle_3|1\rangle_4 + \alpha_A^0\beta_E^1|0\rangle_A|1\rangle_{E_1}|1\rangle_{E_2}|1\rangle_1|1\rangle_2|1\rangle_3|1\rangle_4\\
&\quad +\alpha_A^1\alpha_E^0|1\rangle_A|0\rangle_{E_1}|0\rangle_{E_2}|1\rangle_1|0\rangle_2|0\rangle_3|0\rangle_4 + \alpha_A^1\alpha_E^1|1\rangle_A|1\rangle_{E_1}|0\rangle_{E_2}|0\rangle_1|0\rangle_2|1\rangle_3|0\rangle_4\\
&\quad +\alpha_A^1\beta_E^0|1\rangle_A|0\rangle_{E_1}|1\rangle_{E_2}|1\rangle_1|1\rangle_2|0\rangle_3|1\rangle_4 + \alpha_A^1\beta_E^1|1\rangle_A|1\rangle_{E_1}|1\rangle_{E_2}|0\rangle_1|1\rangle_2|1\rangle_3|1\rangle_4.
\end{aligned} \tag{63}$$

Then, the next step would consist in updating a part of the register subspace from the information encoded in the other part. However, this step is not necessary because the number of terms in Eq (63) is smaller than all the possible measurement outcomes in the register subspace. Thus, the register is always projected onto orthogonal measurement outcomes. On the other hand, we update the agent state from the information encoding in the register R_1. Therefore, we perform a CNOT gate in the register-agent subspace, where the register R_1 is the control and the agent is the target,

$$\begin{aligned}
|\Psi\rangle_4 &= U^{\text{CNOT}}_{(R_1,A)}|\Psi\rangle_3,\\
|\Psi\rangle_4 &= \alpha_A^0\alpha_E^0|0\rangle_A|0\rangle_{E_1}|0\rangle_{E_2}|0\rangle_1|0\rangle_2|0\rangle_3|0\rangle_4 + \alpha_A^0\alpha_E^1|1\rangle_A|1\rangle_{E_1}|0\rangle_{E_2}|1\rangle_1|0\rangle_2|1\rangle_3|0\rangle_4\\
&\quad +\alpha_A^0\beta_E^0|0\rangle_A|0\rangle_{E_1}|1\rangle_{E_2}|0\rangle_1|1\rangle_2|0\rangle_3|1\rangle_4 + \alpha_A^0\beta_E^1|1\rangle_A|1\rangle_{E_1}|1\rangle_{E_2}|1\rangle_1|1\rangle_2|1\rangle_3|1\rangle_4\\
&\quad +\alpha_A^1\alpha_E^0|0\rangle_A|0\rangle_{E_1}|0\rangle_{E_2}|1\rangle_1|0\rangle_2|0\rangle_3|0\rangle_4 + \alpha_A^1\alpha_E^1|1\rangle_A|1\rangle_{E_1}|0\rangle_{E_2}|0\rangle_1|0\rangle_2|1\rangle_3|0\rangle_4\\
&\quad +\alpha_A^1\beta_E^0|0\rangle_A|0\rangle_{E_1}|1\rangle_{E_2}|1\rangle_1|1\rangle_2|0\rangle_3|1\rangle_4 + \alpha_A^1\beta_E^1|1\rangle_A|1\rangle_{E_1}|1\rangle_{E_2}|0\rangle_1|1\rangle_2|1\rangle_3|1\rangle_4.
\end{aligned} \tag{64}$$

By measuring the register subspace, we obtain that agent and environment qubit E_1 achieve maximal fidelity.

Quantum reinforcement learning for mixed states

Let us consider now the situation where the environment evolves under a noisy mechanism (for qubit states, noisy mechanisms can be depolarizing noise as well as amplitude damping). In this case, the density matrix describing the environment state reads

$$\rho = \begin{pmatrix} \rho_{00} & \rho_{01} \\ \rho_{01}^{*} & \rho_{11} \end{pmatrix}. \tag{65}$$

We focus now our attention in the application of the quantum reinforcement learning protocol in this type of state. We will show that, by adding more registers, two main results will be obtained. Firstly, even though the environment is in a mixed state, the learning fidelity will be maximal for any measurement outcome in the register basis. Additionally, the measurement outcomes provide relevant information about the coherences of the mixed state. To apply the quantum protocol, we express the mixed state in term of its (non-unique) purification, such as

$$|\Psi_{E+e}\rangle = \left[\sqrt{\rho_{00}}|0\rangle_E + \frac{\rho_{10}}{\sqrt{\rho_{00}}}|1\rangle_E\right]|e_1\rangle + \left[\sqrt{\rho_{11} - \frac{|\rho_{10}|^2}{\rho_{00}}}\right]|1\rangle_E|e_2\rangle, \tag{66}$$

$$|\psi_e\rangle = \frac{\rho_{10}}{\sqrt{\rho_{00}}}|e_1\rangle + \left[\sqrt{\rho_{11} - \frac{|\rho_{10}|^2}{\rho_{00}}}\right]|e_2\rangle \rightarrow |\Psi_{E+e}\rangle = \sqrt{\rho_{00}}|0\rangle_E|e_1\rangle + \sqrt{\rho_{11}}|1\rangle_E|\bar{\psi_e}\rangle. \tag{67}$$

Here, $|\bar{\psi_e}\rangle$ is a normalized vector in the purification Hilbert space. As we can see, the coefficient of the quantum state written in its extended Hilbert space (environment + purification) depends only on the diagonal terms of the mixed state. Moreover, to obtain additional information about the mixed state, we need to perform unitary transformations on it in such a way that the information related to the coherences is in the diagonal of the state after the transformation. To be more specific, we need to perform unitary transformations such that the mixed state can be written as follows,

$$\bar{\rho} \rightarrow U_y \rho U_y^{\dagger} = \frac{1}{2}\begin{pmatrix} 1 + (\rho_{01} + \rho_{01}^{*}) & \rho_{11} - \rho_{00} + (\rho_{01} - \rho_{01}^{*}) \\ \rho_{11} - \rho_{00} - (\rho_{01} - \rho_{01}^{*}) & 1 - (\rho_{01} + \rho_{01}^{*}) \end{pmatrix}, \tag{68}$$

$$\tilde{\rho} \rightarrow U_x \rho U_x^{\dagger} = \frac{1}{2}\begin{pmatrix} 1 - i(\rho_{01} - \rho_{01}^{*}) & \rho_{01} + \rho_{01}^{*} + i(\rho_{11} - \rho_{00}) \\ \rho_{01} + \rho_{01}^{*} - i(\rho_{11} - \rho_{00}) & 1 + i(\rho_{01} - \rho_{01}^{*}) \end{pmatrix}. \tag{69}$$

To carry out this task, we need to add three more registers, where each of them has the function to encode information of diagonal, real, and imaginary part of the coherence terms, respectively. A possible state for the space composed of agent, mixed environment and register is given by

$$|A\rangle = \alpha_A^0|0\rangle_A + \alpha_A^1|1\rangle_A, \tag{70}$$

$$|\Psi_{E+e}\rangle = \sqrt{\rho_{00}}|0\rangle_E|e_1\rangle + \sqrt{\rho_{11}}|1\rangle_E|\overline{\psi}_e\rangle \tag{71}$$

$$|R\rangle = |0\rangle_1|0\rangle_2\frac{1}{\sqrt{3}}(|1\rangle_3|0\rangle_4|0\rangle_5 + |0\rangle_3|1\rangle_4|0\rangle_5 + |0\rangle_3|0\rangle_4|1\rangle_5) \tag{72}$$

$$|\Psi\rangle_0 = |A\rangle|\Psi_{E+e}\rangle|R\rangle. \tag{73}$$

The first step is to apply a unitary transformation, which is conditional to the state of the register R_3, R_4 and R_5. In case that the register state is $|1\rangle_3|0\rangle_4|0\rangle_5$, we apply the transformation $\mathcal{U}_1 = \mathbb{I}_{R_3} \otimes \mathbb{I}_{R_4} \otimes \mathbb{I}_{R_5}$. If the register state is in the state $|0\rangle_3|1\rangle_4|0\rangle_5$, we apply the transformation $\mathcal{U}_2 = \mathbb{I}_{R_3} \otimes U_y \otimes \mathbb{I}_{R_5}$. Finally, if the register state is in the state $|0\rangle_3|0\rangle_4|1\rangle_5$ the unitary transformation is given by $\mathcal{U}_3 = \mathbb{I}_{R_3} \otimes \mathbb{I}_{R_4} \otimes U_x$. Hence, the state after this transformation is given by unitary transformation in the environment state according to

$$\begin{aligned}
|\Psi\rangle_1 &= |A\rangle|\psi_{E+e}\rangle|0\rangle_1|0\rangle_2|1\rangle_3|0\rangle_4|0\rangle_5 + |A\rangle U_y|\psi_{E+e}\rangle|0\rangle_1|0\rangle_2|0\rangle_3|1\rangle_4|0\rangle_5 \\
&\quad +|A\rangle U_x|\psi_{E+e}\rangle|0\rangle_1|0\rangle_2|0\rangle_3|0\rangle_4|1\rangle_5, \\
|\Psi_1\rangle &= \frac{1}{\sqrt{3}}(\alpha_A^0|0\rangle_A + \alpha_A^1|1\rangle_A)[(\sqrt{\rho_{00}}|0\rangle_E|e_1\rangle + \sqrt{\rho_{11}}|1\rangle_E|\overline{\psi}_e\rangle)|0\rangle_1|0\rangle_2|1\rangle_3|0\rangle_4|0\rangle_5 \\
&\quad +\left(\sqrt{\frac{1}{2} + \operatorname{Re}(\rho_{01})}|0\rangle_E|e_1\rangle + \sqrt{\frac{1}{2} - \operatorname{Re}(\rho_{01})}|1\rangle_E|\overline{\psi}_e\rangle\right)|0\rangle_1|0\rangle_2|0\rangle_3|1\rangle_4|0\rangle_5 \\
&\quad +\left(\sqrt{\frac{1}{2} + \operatorname{Im}(\rho_{01})}|0\rangle_E|e_1\rangle + \sqrt{\frac{1}{2} - \operatorname{Im}(\rho_{01})}|1\rangle_E|\overline{\psi}_e\rangle\right)|0\rangle_1|0\rangle_2|0\rangle_3|0\rangle_4|1\rangle_5].
\end{aligned} \tag{74}$$

Afterwards, we apply the quantum protocol as we did in the first section. Namely, we first update the register conditionally to the information of the environment. Then, we update the register R_1 conditionally to the information of the agent. Subsequently, to obtain orthogonal measurement outcomes we perform CNOT gates in the register subspace (R_1 is the control and R_2 is the agent). Finally, the agent is updated in terms of the information encoded in register R_1 (where A is the target and R_1 is the control),

$$
\begin{aligned}
|\Psi\rangle_5 \quad = \quad & \frac{1}{\sqrt{3}}\Bigg(\alpha_A^0\sqrt{\rho_{00}}|0\rangle_A|0\rangle_E|e_1\rangle|0\rangle_1|0\rangle_2|1\rangle_3|0\rangle_4|0\rangle_5 \\
& +\alpha_A^0\sqrt{\rho_{11}}|1\rangle_A|1\rangle_E|\overline{\psi}_e\rangle|1\rangle_1|0\rangle_2|1\rangle_3|0\rangle_4|0\rangle_5 \\
& +\alpha_A^1\sqrt{\rho_{00}}|0\rangle_A|0\rangle_E|e_1\rangle|1\rangle_1|1\rangle_2|1\rangle_3|0\rangle_4|0\rangle_5 \\
& +\alpha_A^1\sqrt{\rho_{11}}|1\rangle_A|1\rangle_E|\overline{\psi}_e\rangle|0\rangle_1|1\rangle_2|1\rangle_3|0\rangle_4|0\rangle_5 \\
& +\alpha_A^0\sqrt{\frac{1}{2}+\ \mathrm{Re}\,(\rho_{01})}|0\rangle_A|0\rangle_E|e_1\rangle|0\rangle_1|0\rangle_2|0\rangle_3|1\rangle_4|0\rangle_5 \\
& +\alpha_A^0\sqrt{\frac{1}{2}-\ \mathrm{Re}\,(\rho_{01})}|1\rangle_A|1\rangle_E|\overline{\psi}_e\rangle|1\rangle_1|0\rangle_2|0\rangle_3|1\rangle_4|0\rangle_5 \\
& +\alpha_A^1\sqrt{\frac{1}{2}+\ \mathrm{Re}\,(\rho_{01})}|0\rangle_A|0\rangle_E|e_1\rangle|1\rangle_1|1\rangle_2|0\rangle_3|1\rangle_4|0\rangle_5 \\
& +\alpha_A^1\sqrt{\frac{1}{2}-\ \mathrm{Re}\,(\rho_{01})}|1\rangle_A|1\rangle_E|\overline{\psi}_e\rangle|0\rangle_1|1\rangle_2|0\rangle_3|1\rangle_4|0\rangle_5 \\
& +\alpha_A^0\sqrt{\frac{1}{2}+\ \mathrm{Im}\,(\rho_{01})}|0\rangle_A|0\rangle_E|e_1\rangle|0\rangle_1|0\rangle_2|0\rangle_3|0\rangle_4|1\rangle_5 \\
& +\alpha_A^0\sqrt{\frac{1}{2}-\ \mathrm{Im}\,(\rho_{01})}|1\rangle_A|1\rangle_E|\overline{\psi}_e\rangle|1\rangle_1|0\rangle_2|0\rangle_3|0\rangle_4|1\rangle_5 \\
& +\alpha_A^1\sqrt{\frac{1}{2}+\ \mathrm{Im}\,(\rho_{01})}|0\rangle_A|0\rangle_E|e_1\rangle|1\rangle_1|1\rangle_2|0\rangle_3|0\rangle_4|1\rangle_5 \\
& +\alpha_A^1\sqrt{\frac{1}{2}-\ \mathrm{Im}\,(\rho_{01})}|1\rangle_A|1\rangle_E|\overline{\psi}_e\rangle|0\rangle_1|1\rangle_2|0\rangle_3|0\rangle_4|1\rangle_5\Bigg).
\end{aligned}
\tag{75}
$$

This quantum reinforcement learning protocol exhibits two features. First, by performing projective measurements on registers R_1, R_2 and R_3, we recover the result studied in the first section, i.e., the learning fidelity is maximal independently of the measurement outcomes in the register subspace. The second feature is that, for specific measurement outcomes in a part of the register subspace, we obtain information about the population (diagonal) and the coherences (off-diagonal) of the mixed state. This feature can be used in problems such as partial cloning in cases where the system in which we can extract information evolves under loss mechanisms.

Analysis of implementation in quantum technologies

An interesting result obtained in this manuscript is that in most of the cases, for the considered quantum reinforcement learning protocols, adding more registers improves the rewarding process. That is, via a purely unitary evolution, without coherent feedback, a maximally positively-correlated agent environment state is achieved, in the sense that the final agent contains the same quantum information as the considered final environment. This means that the agent has acquired the needed information about the environment and accordingly modified it, being this a quantum process. In our formalism, typically, one measurement at the end of the protocol is enough to obtain maximal learning fidelity in one iteration of the process. In this sense, several quantum architectures could benefit of this fact, given that coherent feedback is not needed in this case. For instance, we focus our attention in two prominent platforms, namely, trapped ions and superconducting circuits.

Trapped ions

As we have pointed out along the manuscript, the performance of our proposed quantum protocols is based on the quality of the quantum gates between different subsystems. In this case, the realization of high-fidelity quantum gates is essential to perform the quantum protocol proposed here. Technological progress in trapped ions has enabled to implement single [49] and two-qubit quantum gates [50] with a large fidelity. For the single-qubit gate, e.g., a Beryllium hyperfine transition can be driven with microwave fields or lasers, being the error associated with single-qubit gates below 10^{-4}. For two-qubit gates, the use of either microwaves or a laser beam with modulated amplitude allows for the interaction of both qubits (electronic levels of, e.g., Beryllium or Calcium ions) at the same time. Adiabatic elimination of the motion allows one to obtain maximally entangled states of both ions. The fidelity of trapped-ion two-qubit gates can reach nowadays above 99.9% [51, 52]. Trapped-ion technologies offer long coherences times, which can reach up to the range of seconds [53] for Calcium atoms. In addition, this platform enables state preparation and readout with high fidelity [39, 54, 55]. Here, the use of hyperfine states and the microwave fields improve the optical pumping fidelity and improve the relaxation time T_1 allowing to obtain fidelity readouts of 99.9999% [54].

Superconducting circuits

As in trapped ions, the technological progress in superconducting circuits has grown significantly in the latter years. For instance, artificial atoms whose coherence times are in the microsecond range have been built in coplanar [43] and 3D architectures [44]. On the other hand, integrated Josephson quantum processors allows one to implement quantum gates between two-level systems even in cases where the qubits do not have identical frequencies, as well as making them interact via a quantum bus [56]. The Xmon qubits achieve two-qubit gate fidelities above 99% [41, 42]. These technological progresses have developed feedback loop control in this platform. This feedback protocol relies on high fidelity readout, as well as on conditional control on the outcome of a quantum non-demolition measurement [45, 46]. Even though in the quantum reinforcement learning protocols in this paper coherent feedback is not required, this may be a useful ingredient in other quantum reinforcement learning proposals [23].

Discussion

In summary, we propose a protocol to perform quantum reinforcement learning which does not require coherent feedback and, therefore, may be implemented in a variety of quantum technologies. Our learning protocol, being mostly unitary (except with the final register measurement) considers learning in a loose sense: while it does not depend on feedback, the protocol achieves its aim regardless of the initial state of agent and environment. In this aspect, it is general, and obtains a similar goal than Ref. [23] without the need of feedback, enabling its implementation in a variety of quantum platforms. We also point out that one may employ different performance measures than the one considered here, depending on the agent possible aims. Adding more registers than in previous proposals in the literature [23], the rewarding criterion can be applied at the end of the protocol, while agent and environment need not be measured directly, although only via the registers. We also obtain that when the considered systems are composed of qudits, the number of steps needed to obtain maximal learning fidelity is fixed in each qudit dimension and scales polynomially with the number of qudit subsystems. We consider as well environment states which are mixtures, while the agent can also in this case acquire the appropriate information from them. Theoretically, all the cases considered of qubit, multiqubit, qudit, and multiqudit, have many similarities. Even though the protocols are not directly transformable into one another, a d-dimensional qudit can be rewritten as a $\log_2(d)$ multiqubit system, while a multiqudit system with n qudits is equivalent to an $n \log_2(d)$ multiqubit system. Therefore, in this respect, it is intuitive that the results for all these protocols (namely, that maximal fidelity can be attained) should be related. Nevertheless, it is valuable to show that the protocol can be scaled up to multiqudit systems with many parties and high dimensions, given that this will be an ultimate goal of a scalable quantum device. Implementations of these protocols in trapped ions and superconducting circuits seem feasible with current platforms.

Author Contributions

Conceptualization: F. A. Cárdenas-López, L. Lamata, J. C. Retamal, E. Solano.

Formal analysis: F. A. Cárdenas-López, L. Lamata, J. C. Retamal, E. Solano.

Funding acquisition: L. Lamata, J. C. Retamal, E. Solano.

Investigation: F. A. Cárdenas-López, L. Lamata, J. C. Retamal, E. Solano.

Methodology: F. A. Cárdenas-López, L. Lamata, J. C. Retamal, E. Solano.

Project administration: L. Lamata, J. C. Retamal, E. Solano.

Resources: L. Lamata, J. C. Retamal, E. Solano.

Supervision: L. Lamata, J. C. Retamal, E. Solano.

Validation: F. A. Cárdenas-López, L. Lamata, J. C. Retamal, E. Solano.

Writing – original draft: F. A. Cárdenas-López, L. Lamata, J. C. Retamal, E. Solano.

Writing – review & editing: F. A. Cárdenas-López, L. Lamata, J. C. Retamal, E. Solano.

References

1. Michalski RS, Carbonell JG, Mitchell TM. Machine learning: An artificial intelligence approach. Springer Science & Business Media; 2013.
2. Plamondon R, Srihari SN. Online and off-line handwriting recognition: a comprehensive survey. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2000; 22(1):63–84. https://doi.org/10.1109/34.824821
3. Lee KF, Hon HW, Hwang MY, Mahajan S, Reddy R. The SPHINX speech recognition system. In: International Conference on Acoustics, Speech, and Signal Processing,; 1989. p. 445–448 vol.1.
4. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, et al. Mastering the game of Go with deep neural networks and tree search. Nature. 2016; 529(7587):484–489. https://doi.org/10.1038/nature16961 PMID: 26819042
5. Russell SJ, Norvig P. Artificial Intelligence: A Modern Approach (International Edition). Pearson US Imports & PHIPEs; 2002.
6. Sutton RS, Barto AG. Reinforcement learning: An introduction. vol. 1. MIT press Cambridge; 1998.
7. Wittek P. Quantum machine learning: what quantum computing means to data mining. Academic Press; 2014.
8. Schuld M, Sinayskiy I, Petruccione F. An introduction to quantum machine learning. Contemporary Physics. 2015; 56(2):172–185. https://doi.org/10.1080/00107514.2014.964942
9. Adcock J, Allen E, Day M, Frick S, Hinchliff J, Johnson M, et al. Advances in quantum machine learning. arXiv preprint arXiv:151202900. 2015;.
10. Biamonte J, Wittek P, Pancotti N, Rebentrost P, Wiebe N, Lloyd S. Quantum Machine Learning. Nature. 2017; 549, 195–202. https://doi.org/10.1038/nature23474 PMID: 28905917
11. Dunjko V, Briegel HJ. Machine learning & artificial intelligence in the quantum domain. Rep. Prog. Phys. 2018; 81:074001. https://doi.org/10.1088/1361-6633/aab406 PMID: 29504942

12. Bonner R, Freivalds R. A survey of quantum learning. Quantum Computation and Learning. 2003; p. 106.

13. Aïmeur E, Brassard G, Gambs S. Quantum speed-up for unsupervised learning. Machine Learning. 2013; 90(2):261–287. https://doi.org/10.1007/s10994-012-5316-5

14. Lloyd S, Mohseni M, Rebentrost P. Quantum algorithms for supervised and unsupervised machine learning. arXiv preprint arXiv:13070411. 2013;.

15. Rebentrost P, Mohseni M, Lloyd S. Quantum Support Vector Machine for Big Data Classification. Phys Rev Lett. 2014; 113:130503. https://doi.org/10.1103/PhysRevLett.113.130503 PMID: 25302877

16. Alvarez-Rodriguez U, Lamata L, Escandell-Montero P, Martín-Guerrero JD, Solano E. Supervised Quantum Learning without Measurements. Scientific Reports. 2017; 7(1):13645. https://doi.org/10.1038/s41598-017-13378-0 PMID: 29057923

17. Cai XD, Wu D, Su ZE, Chen MC, Wang XL, Li L, et al. Entanglement-Based Machine Learning on a Quantum Computer. Phys Rev Lett. 2015; 114:110504. https://doi.org/10.1103/PhysRevLett.114.110504 PMID: 25839250

18. Li Z, Liu X, Xu N, Du J. Experimental Realization of a Quantum Support Vector Machine. Phys Rev Lett. 2015; 114:140504. https://doi.org/10.1103/PhysRevLett.114.140504 PMID: 25910101

19. Dong D, Chen C, Li H, Tarn TJ. Quantum Reinforcement Learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 2008; 38(5):1207–1220. https://doi.org/10.1109/TSMCB.2008.925743

20. Paparo GD, Dunjko V, Makmal A, Martin-Delgado MA, Briegel HJ. Quantum Speedup for Active Learning Agents. Phys Rev X. 2014; 4:031002. doi: 10.1103/PhysRevX.4.031002

21. Dunjko V, Taylor JM, Briegel HJ. Quantum-Enhanced Machine Learning. Phys Rev Lett. 2016; 117:130501. https://doi.org/10.1103/PhysRevLett.117.130501 PMID: 27715099

22. Crawford D, Levit A, Ghadermarzy N, Oberoi JS, Ronagh P. Reinforcement Learning Using Quantum Boltzmann Machines. arXiv preprint arXiv:161205695. 2016;.

23. Lamata L. Basic protocols in quantum reinforcement learning with superconducting circuits. Scientific Reports. 2017; 7:1609. https://doi.org/10.1038/s41598-017-01711-6 PMID: 28487535

24. Friis N, Melnikov AA, Kirchmair G, and Briegel HJ. Coherent controlization using superconducting qubits. Scientific Reports. 2015; 5:18036. https://doi.org/10.1038/srep18036 PMID: 26667893

25. Dunjko V, Friis N, and H. J. Briegel Quantum-enhanced deliberation of learning agents using trapped ions New J. Phys. 2015; 17:023006.

26. T. Sriarunothai et al., Speeding-up the decision making of a learning agent using an ion trap quantum processor arXiv:1709.01366.

27. Pfeiffer P, Egusquiza IL, Di Ventra M, Sanz M, Solano E. Quantum memristors. Scientific Reports. 2016; 6:29507 EP –. https://doi.org/10.1038/srep29507 PMID: 27381511

28. Salmilehto J, Deppe F, Di Ventra M, Sanz M, Solano E. Quantum Memristors with Superconducting Circuits. Scientific Reports. 2017; 7:42044 EP –. https://doi.org/10.1038/srep42044 PMID: 28195193

29. Sanz M, Lamata L, Solano E. Invited article: Quantum memristors in quantum photonics. APL Photonics. 2018; 3:080801. https://doi.org/10.1063/1.5036596

30. Shevchenko SN, Pershin YV, Nori F. Qubit-Based Memcapacitors and Meminductors. Phys Rev Applied. 2016; 6:014006. https://doi.org/10.1103/PhysRevApplied.6.014006

31. Benedetti M, Realpe-Gómez J, Perdomo-Ortiz A. Quantum-assisted Helmholtz machines: A quantum-classical deep learning framework for industrial datasets in near-term devices. Quant. Sci. Tech. 2018; 3:034007. https://doi.org/10.1088/2058-9565/aabd98

32. Benedetti M, Realpe-Gómez J, Biswas R, Perdomo-Ortiz A. Estimation of effective temperatures in quantum annealers for sampling applications: A case study with possible applications in deep learning. Phys Rev A. 2016; 94:022308. https://doi.org/10.1103/PhysRevA.94.022308

33. Perdomo-Ortiz A, Benedetti M, Realpe-Gómez J, Biswas R. Opportunities and challenges for quantum-assisted machine learning in near-term quantum computers. arXiv preprint arXiv:170809757. 2017;.

34. Leibfried D, Blatt R, Monroe C, Wineland D. Quantum dynamics of single trapped ions. Rev Mod Phys. 2003; 75:281–324. https://doi.org/10.1103/RevModPhys.75.281

35. Haffner H, Roos CF, Blatt R. Quantum computing with trapped ions. Physics Reports. 2008; 469(4):155–203. https://doi.org/10.1016/j.physrep.2008.09.003

36. Blais A, Gambetta J, Wallraff A, Schuster DI, Girvin SM, Devoret MH, et al. Quantum-information processing with circuit quantum electrodynamics. Phys Rev A. 2007; 75:032329. https://doi.org/10.1103/PhysRevA.75.032329

37. Clarke J, Wilhelm FK. Superconducting quantum bits. Nature. 2008; 453(7198):1031–1042. https://doi.org/10.1038/nature07128 PMID: 18563154

38. Wendin G. Quantum information processing with superconducting circuits: a review. Rep Prog Phys. 2017; 80:106001. https://doi.org/10.1088/1361-6633/aa7e1a PMID: 28682303

39. Harty TP, Allcock DTC, Ballance CJ, Guidoni L, Janacek HA, Linke NM, et al. High-Fidelity Preparation, Gates, Memory, and Readout of a Trapped-Ion Quantum Bit. Phys Rev Lett. 2014; 113:220501. https://doi.org/10.1103/PhysRevLett.113.220501 PMID: 25494060

40. Ballance CJ, Harty TP, Linke NM, Sepiol MA, Lucas DM. High-Fidelity Quantum Logic Gates Using Trapped-Ion Hyperfine Qubits. Phys Rev Lett. 2016; 117:060504. https://doi.org/10.1103/PhysRevLett.117.060504 PMID: 27541450

41. Barends R, Kelly J, Megrant A, Veitia A, Sank D, Jeffrey E, et al. Superconducting quantum circuits at the surface code threshold for fault tolerance. Nature. 2014; 508(7497):500–503. https://doi.org/10.1038/nature13171 PMID: 24759412

42. Barends R, Shabani A, Lamata L, Kelly J, Mezzacapo A, Heras UL, et al. Digitized adiabatic quantum computing with a superconducting circuit. Nature. 2016; 534(7606):222–226. https://doi.org/10.1038/nature17658 PMID: 27279216

43. Barends R, Kelly J, Megrant A, Sank D, Jeffrey E, Chen Y, et al. Coherent Josephson Qubit Suitable for Scalable Quantum Integrated Circuits. Phys Rev Lett. 2013; 111:080502. https://doi.org/10.1103/PhysRevLett.111.080502 PMID: 24010421

44. Paik H, Schuster DI, Bishop LS, Kirchmair G, Catelani G, Sears AP, et al. Observation of High Coherence in Josephson Junction Qubits Measured in a Three-Dimensional Circuit QED Architecture. Phys Rev Lett. 2011; 107:240501. https://doi.org/10.1103/PhysRevLett.107.240501 PMID: 22242979

45. Ristè D, Bultink CC, Lehnert KW, DiCarlo L. Feedback Control of a Solid-State Qubit Using High-Fidelity Projective Measurement. Phys Rev Lett. 2012; 109:240502. https://doi.org/10.1103/PhysRevLett.109.240502 PMID: 23368293

46. Ristè D, DiCarlo L. Digital feedback in superconducting quantum circuits. arXiv preprint arXiv:150801385. 2015;.

47. Koch J, Yu TM, Gambetta J, Houck AA, Schuster DI, Majer J, et al. Charge-insensitive qubit design derived from the Cooper pair box. Phys Rev A. 2007; 76:042319. https://doi.org/10.1103/PhysRevA.76.042319

48. Alber G, Delgado A, Gisin N, Jex I. Generalized quantum XOR-gate for quantum teleportation and state purification in arbitrary dimensional Hilbert spaces. arXiv preprint quant-ph/0008022. 2000;.

49. Brown KR, Wilson AC, Colombe Y, Ospelkaus C, Meier AM, Knill E, et al. Single-qubit-gate error below $\mathbf{10^{-4}}$ in a trapped ion. Phys Rev A. 2011; 84:030303. https://doi.org/10.1103/PhysRevA.84.030303

50. Benhelm J, Kirchmair G, Roos CF, Blatt R. Towards fault-tolerant quantum computing with trapped ions. Nat Phys. 2008; 4(6):463–466. https://doi.org/10.1038/nphys961

51. Gaebler JP, Tan TR, Lin Y, Wan Y, Bowler R, Keith AC, et al. High-Fidelity Universal Gate Set for $^{9}Be^{+}$ Ion Qubits. Phys Rev Lett. 2016; 117:060505. https://doi.org/10.1103/PhysRevLett.117.060505 PMID: 27541451

52. Harty TP, Sepiol MA, Allcock DTC, Ballance CJ, Tarlton JE, Lucas DM. High-Fidelity Trapped-Ion Quantum Logic Using Near-Field Microwaves. Phys Rev Lett. 2016; 117:140501. https://doi.org/10.1103/PhysRevLett.117.140501 PMID: 27740823

53. Langer C, Ozeri R, Jost JD, Chiaverini J, DeMarco B, Ben-Kish A, et al. Long-Lived Qubit Memory Using Atomic Ions. Phys Rev Lett. 2005; 95:060502. https://doi.org/10.1103/PhysRevLett.95.060502 PMID: 16090932

54. Myerson AH, Szwer DJ, Webster SC, Allcock DTC, Curtis MJ, Imreh G, et al. High-Fidelity Readout of Trapped-Ion Qubits. Phys Rev Lett. 2008; 100:200502. https://doi.org/10.1103/PhysRevLett.100.200502 PMID: 18518518

55. Noek R, Vrijsen G, Gaultney D, Mount E, Kim T, Maunz P, et al. High speed, high fidelity detection of an atomic hyperfine qubit. Opt Lett. 2013; 38(22):4735–4738. https://doi.org/10.1364/OL.38.004735 PMID: 24322119

56. Blais A, Huang RS, Wallraff A, Girvin SM, Schoelkopf RJ. Cavity quantum electrodynamics for superconducting electrical circuits: An architecture for quantum computation. Phys Rev A. 2004; 69:062320. https://doi.org/10.1103/PhysRevA.69.062320