Featured in Physics

Machine Learning for Long-Distance Quantum Communication

Julius Wallnöfer,[1,2,*] Alexey A. Melnikov,[2,3,4,5] Wolfgang Dür,[2] and Hans J. Briegel[2,6]

[1] *Department of Physics, Freie Universität Berlin, Arnimallee 14, 14195 Berlin, Germany*

[2] *Institute for Theoretical Physics, University of Innsbruck, Technikerstraße 21a, 6020 Innsbruck, Austria*

[3] *Department of Physics, University of Basel, Klingelbergstrasse 82, 4056 Basel, Switzerland*

[4] *Valiev Institute of Physics and Technology, Russian Academy of Sciences, Nakhimovskii prospekt 36/1, 117218 Moscow, Russia*

[5] *Terra Quantum AG, St. Gallerstr. 16a, 9400 Rorschach, Switzerland*

[6] *Department of Philosophy, University of Konstanz, Fach 17, 78457 Konstanz, Germany*

(Received 29 May 2020; accepted 20 July 2020; published 3 September 2020)

Machine learning can help us in solving problems in the context of big-data analysis and classification, as well as in playing complex games such as Go. But can it also be used to find novel protocols and algorithms for applications such as large-scale quantum communication? Here we show that machine learning can be used to identify central quantum protocols, including teleportation, entanglement purification, and the quantum repeater. These schemes are of importance in long-distance quantum communication, and their discovery has shaped the field of quantum information processing. However, the usefulness of learning agents goes beyond the mere reproduction of known protocols; the same approach allows one to find improved solutions to long-distance communication problems, in particular when dealing with asymmetric situations where the channel noise and segment distance are nonuniform. Our findings are based on the use of projective simulation, a model of a learning agent that combines reinforcement learning and decision making in a physically motivated framework. The learning agent is provided with a universal gate set, and the desired task is specified via a reward scheme. From a technical perspective, the learning agent has to deal with stochastic environments and reactions. We utilize an idea reminiscent of hierarchical skill acquisition, where solutions to subproblems are learned and reused in the overall scheme. This is of particular importance in the development of long-distance communication schemes, and opens the way to using machine learning in the design and implementation of quantum networks.

DOI: 10.1103/PRXQuantum.1.010301

I. INTRODUCTION

Humans have invented technologies with transforming impacts on society. One such example is the internet, which significantly influences our everyday life. The quantum internet [1,2] could become the next generation of such a world-spanning network, and promises applications that go beyond its classical counterpart. These include, e.g., distributed quantum computation, secure communication, and distributed quantum sensing. Quantum technologies are now on the brink of being commercially used, and the quantum internet is conceived as one of the key applications in this context. Such quantum technologies are based on the invention of a number of central protocols and schemes, for instance quantum cryptography [3–7] and teleportation [8]. Additional schemes that solve fundamental problems such as the accumulation of channel noise and decoherence have been discovered and have also shaped future research. These include, e.g., entanglement purification [9–11] and the quantum repeater [12], which allow the possibility of scalable long-distance quantum communication. These schemes are considered key results whose discoveries represent breakthroughs in the field of quantum information processing. But to what extent are human minds required to find such schemes?

Here we show that many of these central quantum protocols can in fact be found using machine learning by phrasing the problem in the framework of reinforcement learning (RL) [13–15], the framework at the forefront of modern artificial intelligence [16–18]. By using projective simulation (PS) [19], a physically motivated framework for RL, we show that teleportation, entanglement swapping, and entanglement purification can be found by a

*julius.wallnoefer@fu-berlin.de

PS agent. We equip the agent with a universal gate set, and specify the desired task via a reward scheme. With certain specifications of the structure of the action and percept spaces, RL then leads to the rediscovery of the desired protocols. Based on these elementary schemes, we then show that such an artificial agent can also learn more complex tasks and discover long-distance communication protocols, the so-called quantum repeaters [12]. The usage of elementary protocols learned previously is of central importance in this case. We also equip the agent with the possibility to call subagents, thereby allowing the design of a hierarchical scheme [20,21] that offers flexibility to deal with various environmental situations. The proper combination of optimized block actions discovered by the subagents is the central element in this learning stage, which allows the agent to find a scalable, efficient scheme for long-distance communication. We are aware that we make use of existing knowledge in the specific design of the challenges. Rediscovering existing protocols under such guidance is naturally very different from the original achievement (by humans) of conceiving of and proposing them in the first place, an essential part of which includes the identification of relevant concepts and resources. However, the agent not only rediscovers known protocols and schemes, but also can go beyond known solutions. In particular, we find that in asymmetric situations, where the channel noise and decoherence are nonuniform, the schemes found by the agent outperform human-designed schemes that are based on known solutions for symmetric cases.

From a technical perspective, the agent is situated in a stochastic environment [13,14,22], as measurements with random outcomes are central elements of some of the schemes considered. This requires the agent to learn proper reactions to all measurement outcomes, e.g., the required correction operations in a teleportation protocol depending on the outcomes of (Bell) measurements. Additional elements include abort operations, as not all measurement outcomes lead to a situation where the resulting state can be used further. This happens, for instance, in entanglement purification, where the process needs to be restarted in some cases as the resulting state is no longer entangled. The overall scheme is thus probabilistic. These are challenges that have not been treated in projective simulation before, but the PS agent can in fact deal with such challenges. Another interesting element is the usage of block actions that have been learned previously. This is a mechanism similar to hierarchical skill learning in robotics [20,21] and to clip composition in PS [19,23,24], where previously learned tasks are used to solve more complex challenges and problems. Here we use this concept for long-distance communication schemes. The initial situation is a quantum channel that is subdivided by multiple repeater stations that share entangled pairs with their neighboring stations. Previously learned protocols, namely entanglement swapping and entanglement purification, are used as new primitives. In addition, the agent is allowed to employ subagents that operate in the same way but deal with a problem on a smaller scale, i.e., they find optimized block actions for shorter distances that the main agent can employ on a larger scale. This allows the agent to deal with large systems and rediscover the quantum repeater, with its favorable scaling. The ability to delegate is of special importance in asymmetric situations, as such block actions need to be learned separately for different initial states of the environment—in our case, the fidelity of the elementary pairs might vary drastically either because they correspond to segments with different channel noise or because they are of different length. In this case, the agent outperforms human-designed protocols that are tailored to symmetric situations.

The paper is organized as follows. In Sec. II, we provide background information on reinforcement learning and projective simulation, and discuss our approach to applying these techniques to problems in quantum communication. In Sec. III, we show that the PS agent can find solutions to elementary quantum protocols, thereby rediscovering teleportation, entanglement swapping, entanglement purification, and the elementary repeater cycle. In Sec. IV, we present results for a scaling repeater in a symmetric and an asymmetric setting, and summarize and conclude the paper in Sec. V.

II. PROJECTIVE SIMULATION FOR QUANTUM COMMUNICATION TASKS

In this paper, the process of designing quantum communication protocols is viewed as a RL problem. RL, and more generally machine learning (ML), is becoming increasingly more useful in the automation of problem-solving in quantum information science [25–27]. First, ML has been shown to be capable of designing new quantum experiments [24,28–30] and new quantum algorithms [31,32]. Next, by building a bridge between knowledge about quantum algorithms and actual near-term experimental capabilities, ML can be used to identify problems in which a quantum advantage over a classical approach can be obtained [33–35]. Then, ML can be used to realize such algorithms and protocols in quantum devices, by autonomously learning how to control [36–38], error-correct [39–42], and measure [43] quantum devices. Finally, given experimental data, ML can reconstruct quantum states of physical systems [44–46], learn a compact representation of these states, and characterize them [47–49].

Here we propose learning quantum communication protocols by a trial-and-error process. This process is visualized in Fig. 1 as an interaction between a RL agent and its environment: by trial and error, the agent manipulates

quantum states and hence constructs communication protocols. In each interaction step, the RL agent perceives the current state of the protocol (the environment) and chooses one of the available operations (actions). This action modifies the previous version of the protocol, and the interaction step ends. In addition to the state of the protocol, the agent gets feedback in each interaction step. This feedback is specified by a reward function, which depends on the specific quantum communication task (a)–(d) in Fig. 1. A reward is interpreted by the RL agent, and its memory is updated.

The RL approach described here is used for two reasons. First, there is a similarity between a target quantum communication protocol and a typical RL target. A target quantum communication protocol is a sequence of elementary operations leading to a desired quantum state, whereas a target of a RL agent is a sequence of actions that maximizes the achievable reward. In both cases the solution is therefore a sequence, which makes it natural to assign each elementary quantum operation a corresponding action, and to assign each desired state a reward. Second, the way the targets described are achieved is similar in RL and quantum communication protocols. In both cases, an initial search (exploration) over a large number of operation (or action) sequences is needed. This search space can be viewed as a network, where the states of a quantum communication environment are vertices and the basic quantum operations are edges. The structure of a complex network formed in this way is similar to that observed in quantum experiments [24], which makes the search problem equivalent to navigation in mazes—a reference problem in RL [14,50–52].

It should also be said that the role of the RL agent goes beyond mere parameter estimation for the following reasons. First, using simple search methods (e.g., a brute-force or guided search) would fail for the problem sizes considered: e.g., in the teleportation task discussed in Sec. III A, the number of possible states of the communication environment is at least $7^{14} > 0.6 \times 10^{12}$ [53]. Second, the RL agent learns in the space of its memory parameters, but this is not the case with optimization techniques (e.g., genetic algorithms, simulated annealing, or gradient descent algorithms) that would search directly in the parameter space of communication protocols. Optimizing directly in the space of protocols, which consist of both actions and stochastic responses of the environment, can be efficient only if the space is sufficiently small [14]. Additional complication is introduced by the fact that reward signals are often sparse in quantum communication tasks, and hence the reward gradient is almost always zero, giving optimization algorithms no direction for parameter change. Third, using an optimization technique for constructing an optimal action sequence, ignoring stochastic environment responses, is usually not possible in quantum communication tasks. Because different responses are needed depending on measurement outcomes, there is no single action sequence that achieves an optimal protocol, i.e., there is no single optimal point in the parameter space for such an optimization technique. Nevertheless, there is at least one point in the RL agent's memory-parameter space that achieves an optimal protocol, as the RL agent can choose an action depending on the current state of the environment rather than choosing a whole action sequence.

As the learning agent that operates within the RL framework shown in Fig. 1, we use a PS agent [19,54]. PS is a physically motivated approach to learning and decision making, which is based on deliberation in an episodic and compositional memory (ECM). The ECM is organized as an adjustable network of memory units, which provides flexibility in constructing different concepts for learning, e.g., meta-learning [55] and generalization [56,57]. The deliberation in the ECM is based on a random-walk process that is not computationally demanding, and which in addition can be sped up via a quantum walk process

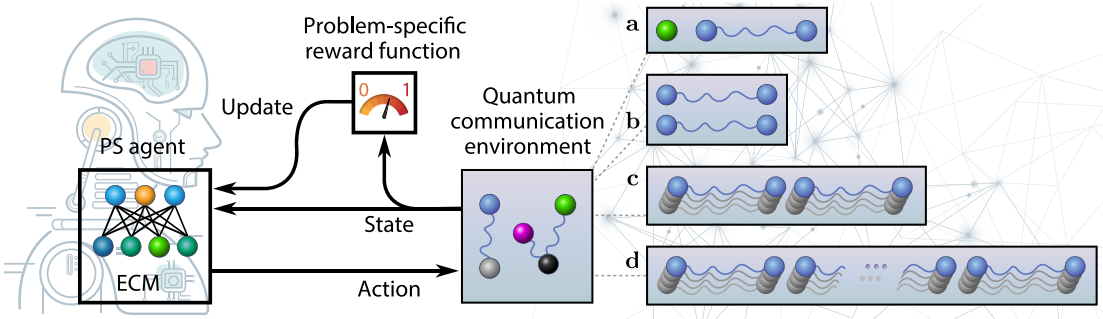

FIG. 1. Illustration of a reinforcement-learning agent interacting with the environment. The agent performs actions that change the state of the environment, while the environment communicates information about its state to the agent. The reward function is customized for each environment. ECM, episodic and compositional memory. The initial states for the different environments that we consider here are illustrated: (a) teleportation of an unknown state, (b) entanglement purification applied recurrently, (c) quantum repeater with entanglement purification and entanglement swapping, (d) scaling of quantum-repeater concepts to distribute long-distance entanglement.

[58,59], leading to a quadratic speedup in deliberation time [60,61], which makes the PS model conceptually attractive. Physical implementations of the basic PS agent and the quantum-enhanced PS agent have been proposed using photonic platforms [62], trapped ions [63], and superconducting circuits [64]. Quantum-enhanced deliberation was recently implemented, as a proof of principle, in a small-scale quantum information processor based on trapped ions [65].

The use of PS in the design of quantum communication protocols has further advantages compared with other approaches, such as standard tabular RL models or deep RL networks. First, the PS agent has been shown to perform well on problems that, from a RL perspective, are conceptually similar to designing communication networks. In problems that can be mapped to a navigation problem [66], such as the design of quantum experiments [24] and the optimization of quantum error-correction codes [40], PS outperformed methods that were used practically for those problems (and were not based on machine learning). In standard navigation problems, such as the grid-world and mountain-car problems, the basic PS agent shows performance qualitatively and quantitatively similar to the standard tabular RL models of SARSA and Q-learning [66]. Second, as was shown in Ref. [66], the computational effort is 1–2 orders of magnitude lower compared with tabular approaches. The reason for this is low model complexity: in static task environments, the basic PS agent has only one relevant model parameter. This makes it easy to set up the agent for a new complex environment, such as a quantum communication network, where model-parameter optimization is costly because of the run time of the simulations. Third, it has been shown that a variant of the PS agent converges to optimal behavior in a large class of Markov decision processes [67]. Fourth, by construction, the decision making of the PS agent can be explained by analyzing the graph properties of its ECM. Because of this intrinsic interpretability of the PS model, we are able to properly analyze the outcomes of the learning process [24].

Next, we show how the PS agent learns quantum communication protocols. The code of the PS agent used in this context is a derivative of a publicly available Python code [68].

III. LEARNING ELEMENTARY PROTOCOLS

We let the agent interact with various environments where the initial states and goals correspond to well-known quantum information protocols. For each of the protocols, we first explain our formulation of the environment and the techniques that we use. Then we discuss the solutions that the agent finds, before finally comparing them with the established protocols. A detailed description of the environments together with additional results can be found in the Appendices.

The learning process follows a similar structure for all of the environments as the agent interacts with the environment over multiple trials. One trial consists of multiple interactions between the agent and the environment. At the beginning of each trial, the environment is initialized in the initial setup, and so each individual trial starts again from a blank slate. The agent selects one of the available actions (which are specific to the environment), and then the environment provides information to the agent about whether the goal has been reached (with a reward $R > 0$) or not ($R = 0$), together with the current percept. The agent then gets to choose the next action, and this repeats until the trial ends either successfully, if the goal is reached, or unsuccessfully, e.g., if a maximum number of actions is exceeded or if there are no more actions left. We call a sequence of actions that the agent used successfully in one trial a protocol.

A. Quantum teleportation

The quantum teleportation protocol [8] is one of the central protocols of quantum information. In the standard version, a maximally entangled state shared between two parties, A and B, is used as a resource and serves to teleport the unknown quantum state of a third qubit, which is also held by party A, from A to B. To achieve this, A performs a Bell measurement and communicates the outcome to B via classical communication, and then B performs a correction operation (a Pauli operation) depending on the measurement outcome. Notice that the same scheme can serve for entanglement swapping when the qubit to be teleported is itself entangled with a fourth qubit held by another party.

1. Basic protocol

The agent is tasked to find a way to transmit quantum information without directly sending the quantum system to the recipient. As an additional resource, a maximally entangled state shared between the sender and the recipient is available. The agent can apply operations from a (universal) gate set locally. This task challenges the agent to find, without any prior knowledge, the best (shortest) sequence of operations out of a large number of possible action sequences, which grows exponentially with the sequence length.

We describe the learning task as follows: There are two qubits A and A' at the sender's station and one qubit B at the recipient's station. Initially, the qubits A and B are in a maximally entangled state $\left|\Phi^+\right\rangle = (1/\sqrt{2})\,(|00\rangle + |11\rangle)$, and A' is in an arbitrary input state $|\Psi\rangle$. The setup is depicted in Fig. 2(a). For this setup, we consider two different sets of actions: The first is a Clifford gate set consisting of the Hadamard gate H and the P-gate $P = \mathrm{diag}(1, i)$, as

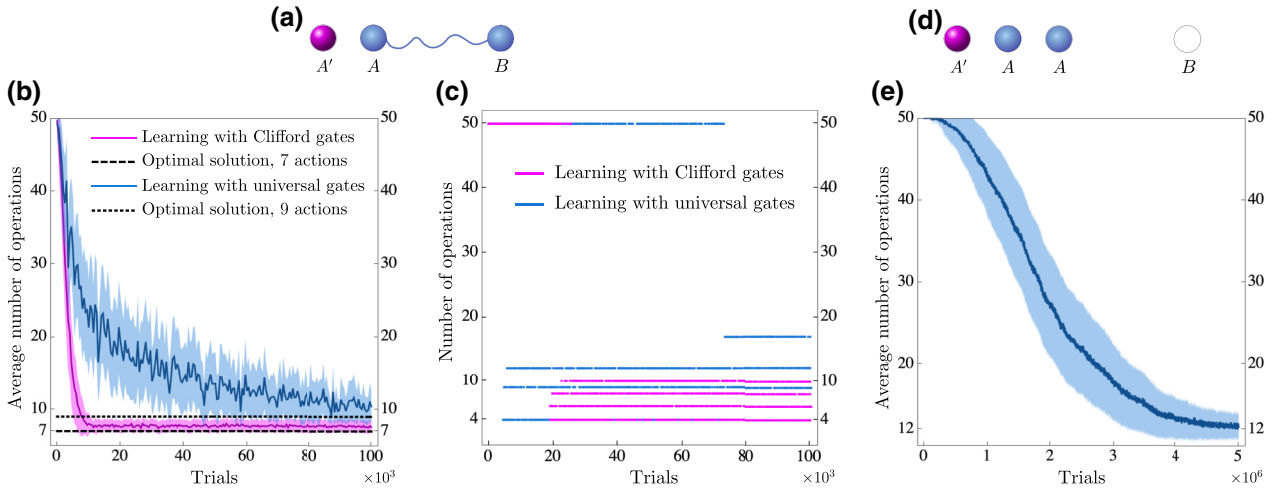

FIG. 2. Reinforcement learning of a teleportation protocol. (a) Initial setup: the agent is tasked to teleport the state of qubit A' to B using a Bell pair shared between A and B. (b) Learning curves of an ensemble of PS agents: average number of actions performed in order to teleport an unknown quantum state when Clifford gates (magenta) and universal gates (blue) are the sets of available actions. (c) Two learning curves (magenta and blue) of two individual PS agents. Four solutions of different length are found by the agents. (d) Initial setup without predistributed entanglement. (e) Learning curve in the learning setting (d): average number of actions performed in order to teleport an unknown quantum state. In (b),(e), the curves represent an average over 500 agents. The shaded areas show the mean squared deviation $\pm\sigma/3$. This deviation appears not only because of different individual histories of the agents, but also because of the difference in the individual solution lengths shown in (c).

well as the controlled-NOT (CNOT) gate (which, as a multiqubit operation, can be applied only to qubits at the same station, i.e., in this case only to A and A'). Furthermore, single-qubit measurements in the Z-basis are allowed, with the agent obtaining a measurement outcome. The second set of actions replaces P with $T = \mathrm{diag}(1, e^{i\pi/4})$, which results in a universal set of quantum gates. A detailed description of the actions and percepts can be found in Appendix B.

The task is considered to be successfully solved if the qubit at B is in state $|\Psi\rangle$. In order to ensure that this works for all possible input states, instead of using random input states, we make use of the Jamiołkowski fidelity [69,70] to evaluate whether or not the protocol proposed by the agent is successful. This means that we require that the overlap of the Choi-Jamiołkowski state [69] $|\Phi^+\rangle_{\widetilde{A'A'}}$ corresponding to the effective map generated by the suggested protocol with the Choi-Jamiołkowski state corresponding to the optimal protocol is equal to 1.

The learning curves, i.e., the number of operations that the agent applies to reach a solution in each trial, are shown as an average over 500 agents in Fig. 2(b). Unsuccessful trials are recorded as 50 operations, as that is the maximum number of operations per trial to which the PS agent is limited. Observing the number of operations decrease below 50 means that the PS agent finds a solution. The decline over time in the averaged learning curve stems not only from an increasing number of agents finding a solution but also from individual agents improving their solutions based on their experience. We observe that the learning curve converges to some average number of operations in both cases, i.e., using a Clifford (magenta) and a universal (blue) gate set. However, the mean squared deviation does not go to zero. This can be explained by looking at the individual learning curves of two example agents in Fig. 2(c): the agent does not arrive at a single solution for this problem setup, but rather four different solutions. These solutions can be summarized as follows (up to different orders of commuting operations):

(a) Apply $H_{A'}\,\mathrm{CNOT}^{A'\to A}$, where H is the Hadamard gate and CNOT is the controlled-NOT operation.

(b) Measure qubits A and A' in the computational basis.

(c) Depending on the measurement outcomes, apply either $\mathbb{1}$, X, Y, or Z (decomposed into the elementary gates of the gate set used) to qubit B.

We see the four different solutions in Fig. 2(c) as four horizontal lines, which appear because of the probabilistic nature of the quantum communication environment. The agent learns different sequences of gates because different operations are needed, depending on measurement outcomes that the agent has no control over. Four appropriate correction operations of different length [as seen in Fig. 2(c)], which are needed in order for the agent to successfully transmit quantum information in each trial, complete the protocol. This protocol found by the agent is identical to the well-known quantum teleportation protocol [8].

Note that because we use the Jamiołkowski fidelity to verify that the protocol implements the teleportation channel for all possible input states, it follows that the same protocol can be used for entanglement swapping if the input qubit at A' is part of an entangled state.

2. Variants without predistributed entanglement

The entangled state shared between two distant parties is the key resource that makes the quantum teleportation protocol possible. Naturally, one could ask whether the agent is still able to find a protocol if not provided with the initial entangled state. To this end, we let the agent solve two variants of this task, with the goal of transferring an input state $|\Psi\rangle$ to the receiving station B without sending it directly.

Variant 1.—Initially, there are two qubits $|0\rangle_{A_1}|0\rangle_{A_2}$ and the input qubit in state $|\Psi\rangle$, all at the sender's station A. Note that in this variant there is no qubit at station B initially. In addition to a universal gate set (multiqubit operations can be applied only to qubits at the same station), the agent now has the additional capability to send a qubit to the recipient's station, with the important restriction that the input qubit may not be sent.

In this case the agent quickly finds a solution, which is, however, different from the standard teleportation protocol and uses only one of the qubits A_1, A_2 provided. The protocol is (up to order of commuting operators and permutations of qubits A_1 and A_2) given by the following: Apply $H_{A'}\mathrm{CNOT}^{A'\to A_1}$, and then send qubit A_1 to the receiving station. Now measure qubit A' in the computational basis. If the outcome is -1, apply the Pauli-Z operator to qubit A_1. This protocol was called *one-bit teleportation* in Ref. [71] and is even simpler, conceptually and in terms of resources, than the standard teleportation.

Variant 2.—Now let us consider the same environment as in Variant 1 but with the additional restriction that no gates or measurements can be applied to the input qubit A' until one of the two other qubits has been sent to the recipient's station. In this case the agent indeed finds the standard quantum teleportation protocol, by first creating a Bell pair via $\mathrm{CNOT}^{A_1\to A_2}H_{A_1}$ and sending A_2 to the second station—the rest is identical to the base case discussed before. In Fig. 2(e), the learning curve averaged over 500 agents is shown. Obviously, it takes significantly more trials for the agent to find solutions, as the action sequence is longer. Nonetheless, this shows that the agent can find the quantum teleportation protocol without being given an entangled state as an initial resource.

B. Entanglement purification

Noise and imperfections are a fundamental obstacle to distributing entanglement over long distances, and so a strategy to deal with these is needed. Entanglement purification is one approach that is integral to enabling long-distance quantum communication. It is a probabilistic protocol that generates out of two noisy copies of a (nonmaximally) entangled state a single copy with increased fidelity. Iterative application of this scheme yields pairs with higher and higher fidelity, and eventually maximally entangled pairs are generated.

In particular, here we investigate a situation that uses a larger amount of entanglement in the form of multiple noisy Bell pairs, each of which may have been affected by noise during the initial distribution, and try to obtain fewer less noisy pairs from them. Again, the agent has to rely on using only local operations at the two different stations that are connected by the Bell pairs.

Specifically, we provide the agent with two noisy Bell pairs $\rho_{A_1B_1}\otimes\rho_{A_2B_2}$ as input, where ρ is of the form $\rho = F\left|\Phi^+\right\rangle\!\left\langle\Phi^+\right| + [(1-F)/3]\left(\left|\Psi^+\right\rangle\!\left\langle\Psi^+\right| + \left|\Phi^-\right\rangle\!\left\langle\Phi^-\right| + \left|\Psi^-\right\rangle\!\left\langle\Psi^-\right|\right)$. Here, $|\Phi^\pm\rangle$ and $|\Psi^\pm\rangle$ denote the standard Bell basis, and F is the fidelity with respect to $|\Phi^+\rangle$. This starting situation is depicted in Fig. 3(a). The agent is tasked with finding a protocol that probabilistically outputs one copy with increased fidelity. However, it is desirable to obtain a protocol that not only results in an increased fidelity when applied once, but also consistently increases the fidelity when applied recurrently, i.e., on two pairs that have been obtained from the previous round of the protocol. In order to make such a recurrent application possible when dealing with probabilistic measurements, identifying the branches that should be reused is an integral part.

To this end, a different technique than before is employed. Rather than simply obtaining a random measurement outcome every time the agent picks a measurement action, instead the agent needs to provide potentially different actions for all possible outcomes. The actions taken on all the different branches of the protocol are then evaluated as a whole. This makes it possible to calculate the result of the recurrent application of that protocol separately for each trial. The agent is rewarded according to both the overall success probability of the protocol and the increase in fidelity obtained.

The agent is provided with a Clifford gate set and single-qubit measurements. Qubits labeled A_i are held by one party, and qubits labeled B_i are held by another party. Multiqubit operations can be applied only to qubits at the same station. The output of each of the branches is forced to be a state with one qubit on side A and one on side B, along with a decision by the agent whether to consider that branch a success or a failure for the purpose of iterating the protocol. Since this naturally needs two single-qubit measurements, with two possible outcomes each, there are four branches that need to be considered.

In Fig. 3(b), we see reward values that 100 agents obtain for the protocols applied to initial states with fidelities of $F = 0.73$. The reward is normalized such that the entanglement-purification protocol presented in Ref. [10] would obtain a reward of 1.0. All the successful protocols found start in the same way (up to permutations

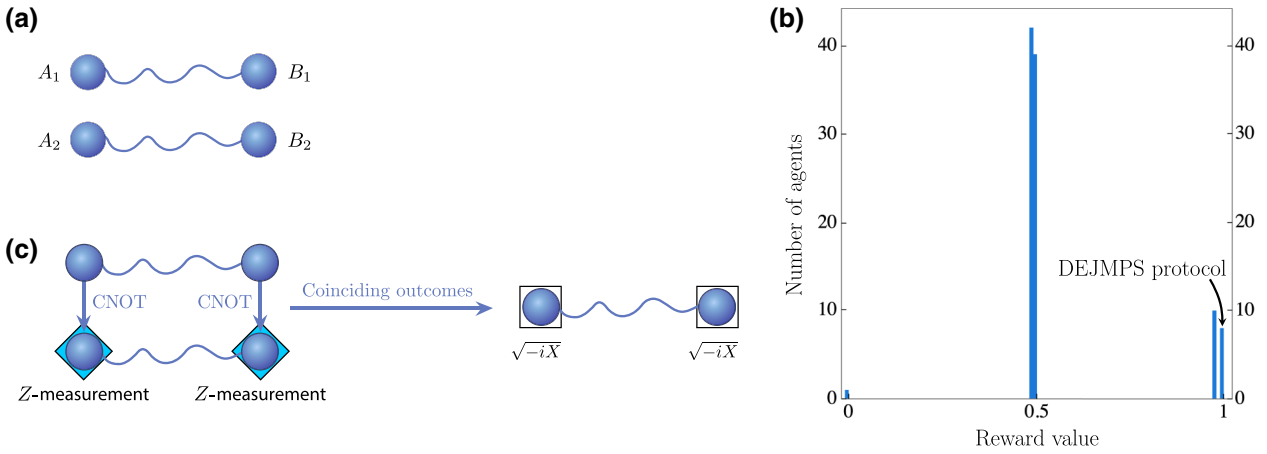

FIG. 3. Reinforcement learning of an entanglement-purification protocol. (a) Initial setup of the quantum communication environment: two entangled noisy pairs shared between two stations A and B. (b) Cumulative reward obtained by 100 agents for the protocols found after 5×10^5 trials. (c) Illustration of the best protocol found by an agent: Apply bilateral CNOT operations and measure one of the pairs. If the measurement outcomes coincide, the protocol is considered successful, and $\sqrt{-iX}$ is applied to both remaining qubits before the next entanglement-purification step.

of commuting operations): they apply $\mathrm{CNOT}^{A_1 \to A_2} \otimes \mathrm{CNOT}^{B_1 \to B_2}$ followed by measuring qubits A_2 and B_2 in the computational basis. In some of the protocols, two of the four branches previously discussed are marked as successful, while others mark only one particular combination of measurement outcomes. The latter therefore have a smaller probability of success, which is reflected in the reward. However, looking closely at the distribution in Fig. 3(b), we can see that these cases correspond to two variants with slightly different rewards. These variants differ in the operations that are applied to the output copies before the next purification step. The variant with slightly lower reward applies the Hadamard gate to both qubits: $H \otimes H$. The protocol that obtains the full reward of 1.0 applies $\sqrt{-iX} \otimes \sqrt{-iX}$ and is depicted in Fig. 3(c). This protocol is equivalent to the well-known DEJMPS protocol [10] for an even number of recurrence steps, but requires a shorter action sequence for the gate set provided to the agent. We discuss this solution in more detail, as well as an additional variant of the environment with automatic depolarization after each recurrence step, in Appendix C.

C. Quantum repeater

Entanglement purification alone certainly increases the distance over which one can distribute an entangled state of sufficiently high fidelity. However, the reachable distance is limited because at some point too much noise will accumulate, such that the initial states will no longer have the minimal fidelity required for the entanglement-purification protocol. The insight at the heart of the quantum-repeater protocol [12] is that one can split up the channels into smaller segments and use entanglement purification on short-distance pairs before performing entanglement swapping to create a long-distance pair. In the most extreme case, with very noisy (but still purifiable) short-distance pairs, the requirement of prior purification can easily be understood, since entanglement swapping alone would produce a state that can no longer be purified, but this approach can also be beneficial when considering resource requirements for less extreme noise levels.

While the value of the repeater protocol lies in its scaling behavior, which becomes manifest as the number of links grows, for now the agent has to deal with only two channel segments that distribute noisy Bell pairs with a common station in the middle, as depicted in Fig. 4(a). In this scenario, the challenge for the agent is to use the protocols of the previous sections in order to distribute an entangled state over the whole distance. To this end, the agent may use the previously discovered protocols for teleportation, entanglement swapping, and entanglement purification as elementary actions, rather than individual gates.

The task is to find a protocol for distributing an entangled state between the two outer stations with a threshold fidelity of at least 0.9, all the while using as few initial states as possible. The initial Bell pairs are considered to have initial fidelities of $F = 0.75$. Furthermore, the CNOT gates used for entanglement purification are considered to be imperfect; we model this imperfection as local depolarizing noise, with reliability parameter p, acting on the two qubits involved followed by a perfect CNOT operation [11]. The effective map $\mathcal{M}^{a\to b}_{\mathrm{CNOT}}$ is given by

$$\mathcal{M}^{a\to b}_{\mathrm{CNOT}}(p)\rho = \mathrm{CNOT}_{ab}\left[\mathcal{D}^a(p)\mathcal{D}^b(p)\rho\right]\mathrm{CNOT}^{\dagger}_{ab}, \quad (1)$$

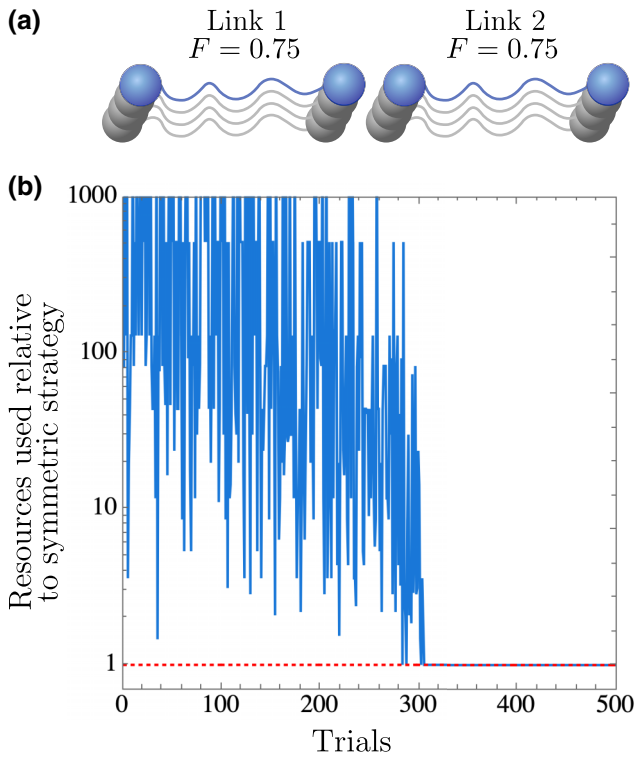

FIG. 4. Reinforcement learning of a quantum-repeater protocol. (a) Initial setup for a length-2 quantum-repeater environment. The agent is provided with many copies of noisy Bell states with initial fidelities $F = 0.75$, which can be purified on each link separately or connected via entanglement swapping at the middle station. (b) Learning curve in terms of resources (initial Bell pairs used) for the best of 128 agents with gate reliability parameter $p = 0.99$. The known repeater solution (red line) is reached.

where $\mathcal{D}^i(p)$ denotes the local depolarizing noise channel with reliability parameter p acting on the ith qubit:

$$\mathcal{D}^i(p)\rho = p\rho + \frac{1-p}{4}\left(\rho + X^i\rho X^i + Y^i\rho Y^i + Z^i\rho Z^i\right), \tag{2}$$

with X^i, Y^i, Z^i denoting the Pauli matrices acting on the ith qubit.

While the point of such an approach begins to show only for much longer distances, which we take a look at in Sec. IV, some key concepts can already be observed on small scales.

The agent naturally tends to find solutions that use a small number of actions in an environment that is similar to a navigation problem. However, this is not necessarily desirable here, because the amount of resources, i.e., the number of initial Bell pairs, is the figure of merit in this scenario rather than the number of actions. Therefore an appropriate reward function for this environment takes the resources used into account.

In Fig. 4(b), the learning curve of the best of 128 agents in terms of resources used is depicted. Looking at the best solutions, the key insight is that it is beneficial to purify the short-distance pairs a few times before connecting them via entanglement swapping, even though this way more actions need to be performed by the agent. This solution is in line with the idea of the established quantum-repeater protocol [12].

IV. SCALING QUANTUM REPEATER

The point of the quantum repeater lies in its scaling behavior, which starts to show only when one considers distances longer than just two links. This means that we have to consider starting situations of variable length, as depicted in Fig. 1(d), using the same error model as described in Sec. III C. In order to distribute entanglement over varying distances, the agent needs to come up with a scalable scheme. However, both the action space and the length of the action sequences required to find a solution would quickly become unmanageable with increasing distance. Furthermore, a RL agent learns a solution for a particular situation and problem size rather than finding a universal concept that can be transferred to similar starting situations and larger scales.

To overcome these restrictions, we provide the agent with the ability to effectively *outsource* finding solutions for distributing an entangled pair over a short distance and reuse them as elementary actions for a larger setting. This means that, as a single action, the agent can instruct multiple subagents to come up with a solution for a small distance and then pick the best action sequence from among those solutions. This process is illustrated in Fig. 5(a).

Again, the aim is to come up with a protocol that distributes an entangled pair over a long distance with sufficiently high fidelity, while using as few resources as possible.

A. Symmetric protocols

First, we take a look at a symmetric variant of this setup: The initial situation is symmetric, and the agent is allowed to do actions in a symmetric way only. If it applies one step of an entanglement-purification protocol to one of the initial pairs, all the other pairs need to be treated in the same way. Similarly, entanglement swapping is always performed at every second station that is still connected to other stations. In Figs. 5(b) and Fig. 5(c), the results for various lengths of Bell pairs with an initial fidelity of $F = 0.75$ are shown. We compare the solutions that the agent finds with a strategy that repeatedly purifies all pairs up to a chosen working fidelity, followed by entanglement swapping (see Sec. E 4). For lengths greater than eight repeater links, the agent still finds a solution with desirable scaling behavior, while using only slightly more resources.

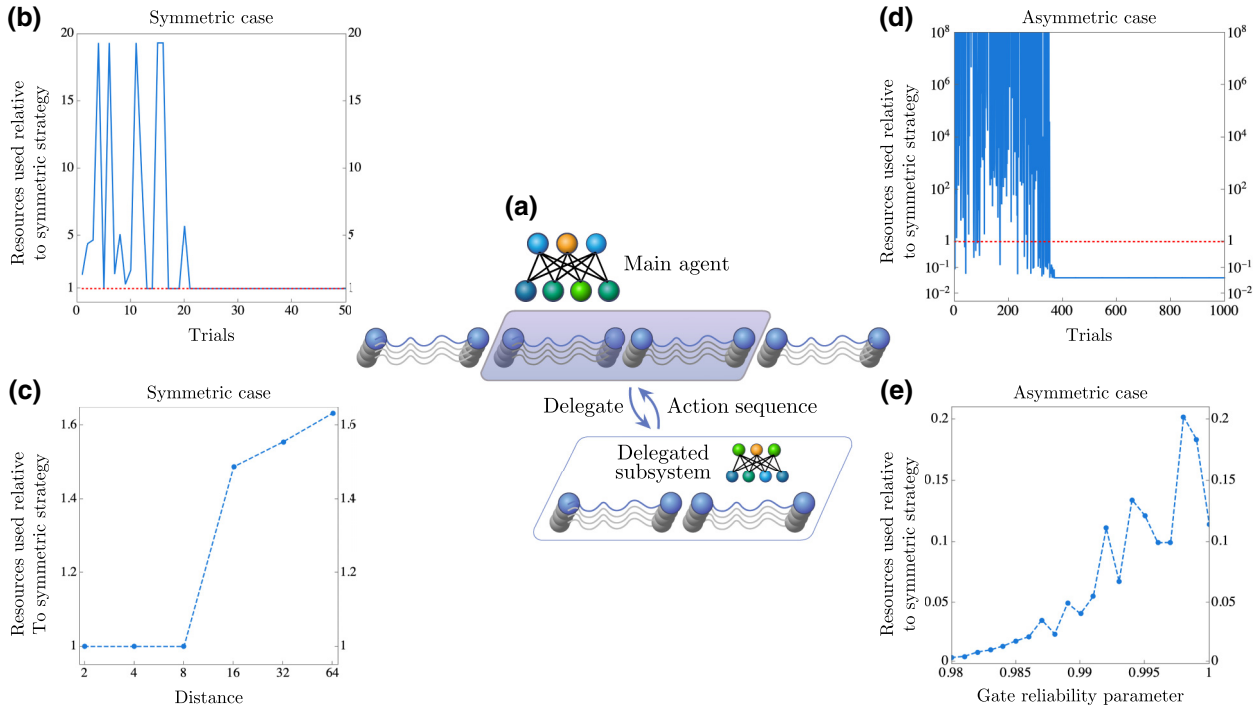

FIG. 5. Reinforcement learning of a scalable quantum-repeater protocol. (a) Illustration of delegation of a block action. The main agent that is tasked with finding a solution to the large-scale problem (repeater length 4 in this example) delegates the solution of a subsystem of length 2 to another agent. That agent comes up with a solution to the smaller-scale problem, and that action sequence is then applied to the larger problem. This counts as one single action for the main agent. For settings even larger then this, the subagent itself can again delegate the solution of a subsubsystem to yet another agent. (b),(c) Scaling repeater with forced symmetric protocols with initial fidelities $F = 0.75$. Gate reliability parameter $p = 0.99$. The threshold fidelity for a successful solution is 0.9. The red line corresponds to a solution with approximately 0.518×10^8 resources used. (b) Best solution found by an agent for repeater length 8. (c) Relative resources used by the agent's solution compared with a symmetric strategy for different repeater lengths. (d),(e) Scaling repeater with asymmetric initial fidelities $(0.8, 0.6, 0.8, 0.8, 0.7, 0.8, 0.8, 0.6)$. The threshold fidelity for a successful solution is 0.9. (d) The best solution found by an agent with gate reliability $p = 0.99$ outperforms a strategy that does not take the asymmetric nature of the initial state into account (red line). (e) Relative resources used by the agent's solution compared with a symmetric strategy for different reliability parameters. (The jumps in the relative resources used are likely due to threshold effects or to agents converging to a very short sequence of block actions that is not optimal.)

B. Asymmetric setup

The more interesting scenario is when the initial Bell pairs are subjected to different levels of noise, e.g., when the physical channels between stations are of different length or quality. In this scenario, symmetric protocols are not optimal.

We consider the following scenario: nine repeater stations connected via links that can distribute Bell pairs of different initial fidelities $(0.8, 0.6, 0.8, 0.8, 0.7, 0.8, 0.8, 0.6)$. In Fig. 5(d), the learning curve in terms of resources for an agent that can delegate work to subagents is shown. The gate reliability of the CNOT gates used in the entanglement-purification protocol is $p = 0.99$. The solution obtained is compared with the resources needed for a protocol that does not take the asymmetric nature of this situation into account and that is also used as an initial guess for the reward function (see Sec. E 4 for additional details of that approach). Clearly, the solution found by the RL agent is preferable to the protocol tailored to symmetric situations. Figure 5(e) shows how that advantage scales for different gate reliability parameters p.

C. Imperfect memories

One central parameter that influences the performance of a quantum repeater is the quality of the quantum memories available at the repeater stations. It is necessary to store the qubits while the various measurement outcomes of the entanglement swapping and, especially, the entanglement-purification protocols are communicated between the relevant stations.

To this end, we revisit the previous asymmetric setup and assume that the initial fidelities now arise from different channel lengths between the repeater stations. We model the noisy channels as local depolarizing noise [see Eq. (2)] with a length-dependent error parameter $e^{-L/L_{\text{att}}}$, with $L_{\text{att}} = 22$ km [72]. Similarly, we model imperfect

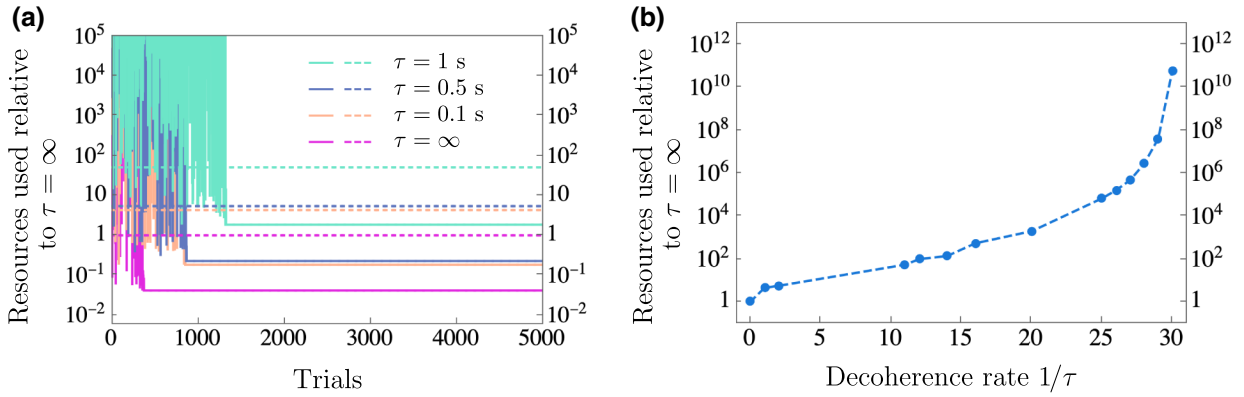

FIG. 6. Reinforcement learning of quantum-repeater protocols with imperfect memories. We consider an asymmetric setup with repeater length 8 and initial fidelities (0.8, 0.6, 0.8, 0.8, 0.7, 0.8, 0.8, 0.6) arising from different distances between the repeater stations. Gate reliability parameter $p = 0.99$. The threshold fidelity for a successful solution is 0.9. (a) Learning curves (relative to the protocol designed for symmetric approaches with perfect memories) for different decoherence times τ of the quantum memories at the repeater stations. The dashed lines show the resources used by symmetric strategies that do not account for the asymmetric situation. (b) Resources (number of initial pairs) required by the protocols found by the agent for different memory times τ (relative to resources required for $\tau = \infty$).

memories by local depolarizing noise with an error parameter $e^{-t/\tau}$, where t is the time for which a qubit is stored and τ is the decoherence time of the quantum memory.

The measurement outcomes of the entanglement-purification protocol need to be communicated between the two stations performing the protocol. This information tells the stations whether or not the purification step is successful, and is needed before the output pair can be used in any further operations. Therefore both qubits of the pair need to be stored for a time $t_{\text{epp}} = l/c$, where the distance between repeater stations is l and the speed of light in glass fiber is $c = 2 \times 10^8$ m/s. The measurement outcomes of the entanglement swapping also need to be communicated from the middle station performing the Bell measurement to the outer repeater stations, so that the correct by-product operator is known. In this case, however, it is not necessary to wait for this information to arrive before proceeding with the next operation. This is the case because all operations used in the protocols are Clifford operations, and thus any Pauli by-product operators can be taken into account retroactively. In our particular case, the information from entanglement swapping may change only the interpretation of the measurement outcomes obtained in the subsequent entanglement-purification protocol, and needs to be available at that time. However, since the information exchange for the entanglement-purification protocol always happens over a longer distance than the associated entanglement swapping, this information will always be available when needed, and so there is no need to account for additional time in memory for this.

Using the same approach as above, we let the agent learn protocols for different values of τ and compare them with the symmetric protocols. We use a gate reliability parameter $p = 0.99$, and the task is considered complete if a pair with a fidelity $F > 0.9$ has been distributed. The learning curves and the advantage over the protocols optimized for symmetric approaches shown in Fig. 6(a) look qualitatively very similar to the result for perfect memories (i.e., $\tau = \infty$). In Fig. 6(b), the numbers of initial pairs required by the protocols found by the agent are shown for different memory times. The required resources increase sharply at a decoherence time of $\tau = 1/30$ s, and the agent is unable to find a protocol for $\tau = 1/31$ s (which could mean either that the threshold has been reached or that the number of actions required for a successful protocol has grown so large that the agent could not find such a protocol in a few thousand trials). It should be noted that this rather demanding memory requirement for this particular setup certainly arises from our very challenging starting situation (e.g., some very low starting fidelities of 0.6).

D. Choosing the location of repeater stations

As an alternative use case, the protocols found by the agent allow us to compare different setups. Let us consider the following situation: We want two distant parties to share entanglement via a quantum repeater using only a small number of intermediate repeater stations. On the path between the terminal stations, there are multiple possible locations where repeater stations could be built. Which combination of locations should be chosen? Furthermore, let us assume that the possible locations are unevenly spaced, so that simply picking a symmetric setup is impossible.

We demonstrate this concept with the following example setup using the same error model as in Sec. IV C for both length-dependent channel noise and imperfect memories ($\tau = 0.1$ s). We consider possible locations (numbered

TABLE I. Resource requirements for various choices for constructing repeater stations in order to connect two distant parties 20 km apart. There are seven possible locations at which one can place asymmetrically spaced repeater stations (see Sec. E 3). Gate reliability parameter $p = 0.99$, decoherence time for quantum memories $\tau = 0.1$ s, attenuation length $L_{\text{att}} = 22$ km.

Station at	Resources	Stations at	Resources	Stations at	Resources
4	9.60×10^4	3, 6	1.06×10^5	2, 4, 7	9.95×10^4
5	1.60×10^5	3, 5	1.39×10^5	3, 5, 7	1.07×10^5
3	1.60×10^5	2, 5	1.59×10^5	2, 4, 6	1.08×10^5
6	4.97×10^5	3, 7	1.64×10^5	1, 3, 6	1.10×10^5
2	5.02×10^5	2, 6	1.64×10^5	1, 3, 5	1.11×10^5
7	4.31×10^6	4, 7	1.80×10^5	2, 3, 5	1.11×10^5
1	5.90×10^8	3, 4	1.90×10^5	3, 4, 6	1.12×10^5

1 to 7) for stations between the two end points that are located at positions that correspond to the asymmetric setup in the previous subsection but are scaled down to a total distance of $L_{\text{tot}} = 20$ km. The positions of all locations are listed in Appendix E. In Table I, we show the best combinations of locations for placing either one, two, or three repeater stations and the amount of resources the agent's protocol takes for these choices.

Naturally, this analysis could be repeated for different initial setups and error models of interest. This particular example can be understood as a proof of principle that using the agent in this way can be useful as well.

V. SUMMARY AND OUTLOOK

We have demonstrated that reinforcement learning can serve as a highly versatile and useful tool in the context of quantum communication. When provided with a sufficiently structured task environment, including an appropriately chosen reward function, the learning agent can retrieve (effectively rediscover) basic quantum communication protocols such as teleportation, entanglement purification, and the quantum repeater. We have developed methods to state challenges that occur in quantum communication as RL problems in a way that offers very general tools to the agent while ensuring that relevant figures of merit are optimized.

We have shown that stating the challenges considered above as a RL problem is beneficial and offers advantages over using optimization techniques, as discussed in Sec. II.

Regarding the question of the extent to which programs can help us in finding genuinely new schemes for quantum communication, it has to be emphasized that a significant part of the work consists in asking the right questions and identifying the relevant resources, and both of these elements are central to the formulation of the task environment and are provided by researchers. However, it should also be noted that not every aspect of designing the environment is necessarily a crucial addition, and many details of the implementation are simply an acknowledgment of practical limitations such as computational run times. When provided with a properly formulated task, a learning agent can play a helpful assisting role in exploring the possibilities.

In fact, we use the PS agent in this way to demonstrate that the application of machine learning techniques to quantum communication is not limited to rediscovering existing protocols. The PS agent finds adapted and optimized solutions in situations that lack certain symmetries assumed by the basic protocols, such as the qualities of the physical channels connecting different stations. We extend the PS model to include the concept of delegating parts of the solution to other agents, which allows the agent to deal effectively with problems of larger size. Using this new capability for long-distance quantum repeaters with asymmetrically distributed channel noise, the agent comes up with novel and practically relevant solutions.

We are confident that the approach presented here can be extended to more complex scenarios. We believe that reinforcement learning can become a practical tool to be applied to quantum communication problems, such as designing quantum networks, that do not have a rich spectrum of existing protocols, especially if the underlying network structure is irregular. Alternatively, machine learning could be used to investigate other architectures for quantum communication, such as constructing cluster states for all-photonic quantum repeaters [73].

ACKNOWLEDGMENTS

J.W. and W.D. were supported by the Austrian Science Fund (FWF) through Grants No. P28000-N27 and No. P30937-N27. J.W. acknowledges funding by Q.Link.X from the BMBF in Germany. A.A.M. acknowledges funding by the Swiss National Science Foundation (SNSF) through Grant No. PP00P2-179109 and by the Army Research Laboratory Center for Distributed Quantum Information via the SciNet project. H.J.B. was supported by the FWF through SFB BeyondC, Grant No. F7102, and by the Ministerium für Wissenschaft, Forschung, und Kunst Baden-Württemberg (AZ: 33-7533.-30-10/41/1).

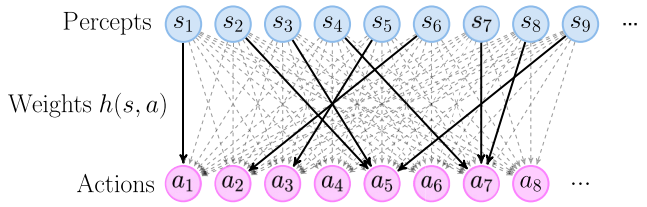

FIG. 7. The PS network in its basic form, which corresponds to a weighed directed bipartite graph. Clips corresponding to eight states of the environment and seven possible actions are shown. They are connected by directed edges. The thick edges correspond to high probabilities of choosing action a_k in the state s_m.

APPENDIX A: INTRODUCTION TO PROJECTIVE SIMULATION

This work is based on using RL for quantum communication. As a RL model, we use the model of PS, which was first introduced in Ref. [19]. In the main text, we provide background information on RL and PS, and explain our motivation for using PS. In this section, we give an introduction to the working principles of the PS agent. The decision making of the PS agent is realized in its episodic and compositional memory, which is a network of memory units, or clips. Each clip encodes a percept, an action, or a sequence thereof. There are several different mechanisms for connecting clips in the PS network, details of which can be found, e.g., in Refs. [56,66]. In this paper, we use a two-layered PS network construction, similar to one used for designing quantum experiments [24]. The two-layered network is schematically shown in Fig. 7.

The first layer of clips corresponds to percepts s, which are states of the (quantum communication) environment. The layer of percepts is connected to the layer of actions a by edges with weights $h(s,a)$. These weights determine the probabilities of choosing the corresponding action: the agent perceives a state s_m and performs action a_k with probability

$$p_{mk} = \frac{e^{h_{mk}}}{\sum_l e^{h_{ml}}}. \tag{A1}$$

The decision-making process within the PS model, using the two-layered network, is hence a one-step random-walk process. One trial, consisting of n agent-environment interaction steps, is an n-step random-walk process defined by the weight matrix h_{mk}. The weights of the network are updated at the end of each agent-environment interaction step i according to the learning rule

$$h_{mk}^{(i+1)} = h_{mk}^{(i)} - \gamma\left(h_{mk}^{(i)} - 1\right) + g_{mk}^{(i+1)} r^{(i)}, \tag{A2}$$

for all m and k; $r^{(i)}$ is the reward, and $g^{(i+1)}$ is a coefficient that distributes this reward in proportion to how much a given edge (m,k) contributes to the sequence of actions rewarded. To be more specific, $g^{(i+1)}$ is set to 1 once the edge has been used in a random-walk process, and goes back to its initial value of 0 with a rate η afterwards:

$$g_{mk}^{(i+1)} = \begin{cases} 1, & \text{if } (m,k) \text{ was traversed,} \\ (1-\eta)\, g_{mk}^{(i)}, & \text{otherwise.} \end{cases} \tag{A3}$$

The time-independent parameter η is set to a value in the interval $[0,1]$. The second metaparameter, $0 \le \gamma \le 1$, of the PS agent is a damping parameter that helps the agent to forget, which is beneficial in cases where the environment changes.

APPENDIX B: QUANTUM TELEPORTATION

1. Description of environment

Figure 2(a) depicts the setup. Qubit A' is initialized as part of an entangled state $|\Phi^+\rangle_{\tilde{A}'A'}$ in order to facilitate measuring the Jamiołkowski fidelity later on. Qubits A and B are in a state $|\Phi^+\rangle_{AB}$.

Goal: The state of A' should be teleported to B. We measure this by calculating the Jamiołkowski fidelity of the effective channel applied by the action sequences. This means that we calculate the overlap of $|\Phi^+\rangle_{\tilde{A}'B}$ with the reduced state $\rho_{\tilde{A}'B}$ to determine whether the goal has been reached.

Actions: The following actions are allowed:

(a) Depending on the specification of the task, either $P = \left(\begin{smallmatrix} 1 & 0 \\ 0 & i \end{smallmatrix}\right)$ for a Clifford gate set or $T = \left(\begin{smallmatrix} 1 & 0 \\ 0 & e^{i\pi/4} \end{smallmatrix}\right)$ for a universal gate set, on each qubit (3 actions).

(b) The Hadamard gate $H = (1/\sqrt{2})\left(\begin{smallmatrix} 1 & 1 \\ 1 & -1 \end{smallmatrix}\right)$ on each qubit (3 actions)

(c) The CNOT gate $\text{CNOT}^{A'\to A}$ on the two qubits at location A (1 action).

(d) Z-measurement on each qubit (3 actions).

In total, there are 10 actions. Note that the Clifford group is generated by P, H, and CNOT [74], and replacing P with T makes it a universal gate set [75]. The measurements are modeled as destructive measurements, which means that operations acting on a particular qubit are no longer available after a measurement has been performed on that qubit, thereby reducing the number of actions that the agent can choose.

Percepts: The agent uses only the previous actions of the current trial as a percept. Z-measurements with different outcomes produce different percepts.

Reward: If the goal is reached, $R = 1$ and the trial ends. Otherwise, $R = 0$.

2. Discussion

One central complication in this scenario is that the entanglement is a limited resource. If the entanglement

is destroyed without anything being accomplished (e.g., qubit A is measured as the first action), then the goal can no longer be reached no matter what the agent tries afterwards. This is a feature that distinguishes this setup and other quantum environments with irreversible operations from a simple problem of navigation in a maze. Instead, this is more akin to a navigational problem where there are numerous cliffs that the agent can fall over and can never get back up again, which means that the agent could be permanently separated from the goal.

An alternative formulation that would make the goal reachable in each trial even if a wrong irreversible action was taken would be to provide the agent with a reset action that resets the environment to the initial state. A different class of percepts would need to be used in this case.

With prior knowledge of the quantum teleportation protocol, it is easy to understand why this problem structure favors a RL approach. The shortest solution for one particular combination of measurement outcomes takes only four actions, and these actions are to be performed regardless of which correction operation needs to be applied. This means that once this simple solution has been found, it significantly reduces the complexity of finding the other solutions, as now only the correction operations need to be found.

Compare this with searching for this solution by brute-forcing action sequences. For the universal gate set, we know that the most complicated of the four solutions takes at least 14 actions. Ignoring the measurement actions for now, as they reduce the number of available actions anyway, there are 7^{14} possible action sequences. So, we would have to try *at least* $7^{14} > 6.7 \times 10^{11}$ sequences, which is vastly more than the few hundred thousand trials needed by the agent.

3. Environment variants without predistributed entanglement

Variant 1.—Figure 2(d) shows the initial setup. The qubits A_1 and A_2 are both initialized in the state $|0\rangle$. As before, the input qubit A' is initialized as part of an entangled state $|\Phi^+\rangle_{\tilde{A}'A'}$ in order to later obtain the Jamiołkowski fidelity.

Goal: A qubit at location B is in the initial state of qubit A'. As before, the Jamiołkowski fidelity is used to determine whether this goal has been reached.

Actions: The following actions are available:

(a) The T-gate $T = \left(\begin{smallmatrix} 1 & 0 \\ 0 & e^{i\pi/4} \end{smallmatrix}\right)$ on every qubit (3 actions).

(b) The Hadamard gate $H = (1/\sqrt{2})\left(\begin{smallmatrix} 1 & 1 \\ 1 & -1 \end{smallmatrix}\right)$ on each qubit (3 actions).

(c) A NOT gate on each pair of qubits as long as they are at the same location (initially $\text{CNOT}^{A'\to A_1}$, $\text{CNOT}^{A'\to A_2}$, and $\text{CNOT}^{A_1\to A_2}$).

(d) A Z-measurement on each qubit (3 actions).

(e) Send a qubit A_1 or A_2 to location B (2 actions).

There are 14 actions initially available. Measuring a qubit removes all actions that involve that qubit from the pool of available actions. Sending a qubit to location B removes the action used itself (as that qubit is now at location B) and also triggers a check of which CNOT actions are now possible.

Percepts: The agent uses as a percept only the information about which of the previous actions of the current trial were taken. Z-measurements with different outcomes produce different percepts.

Reward: If the goal is reached, $R = 1$ and the trial ends. Otherwise, $R = 0$.

Variant 2.—As above, but initially no action involving the input qubit A' is available, i.e., there are no single-qubit gates, no measurement, nor a CNOT gate. After one of the sending actions is used, these actions on qubit A' become available. Furthermore, only one qubit may be sent in total, i.e., after a qubit is sent, neither of the two sending actions may be chosen.

Comment on the variant environments.—Variant 1 serves as a good example of why specifying the task correctly is such an important part of RL problems. As discussed in the main text, the agent immediately spots the loophole and finds a protocol that uses fewer actions than the standard teleportation protocol. Another successful way to circumvent the restriction of not sending A' directly is to construct a SWAP operation from the gate set. It is then possible to simply swap the input state with the state of one of the qubits that can be sent. However, in the given action set, this solution consists of a longer sequence of actions and is therefore deemed more expensive than the one that the agent finds.

APPENDIX C: ENTANGLEMENT PURIFICATION

1. Description of environment

Figure 3(a) shows the initial setup, with A and B sharing two Bell pairs with initial fidelity $F = 0.73$.

Goal: Find an action sequence that results in a protocol that improves the fidelity of the Bell pair when applied recurrently. This means that two copies of the resulting two-qubit state after one successful application of the protocol are taken, and the protocol uses them as input states.

Actions: The following actions are available:

(a) $P_x = HPH$ on each qubit (4 actions).

(b) H, the Hadamard gate, on each qubit (4 actions),

(c) The CNOT gates $\text{CNOT}^{A_1\to A_2}$ and $\text{CNOT}^{B_1\to B_2}$ on qubits at the same location (2 actions).

(d) Z-measurements on each qubit (4 actions).

(e) Accept/reject (2 actions).

In total, there are 16 actions. Note that these gates generate the Clifford group. (We tried different variants of the method of generating the gate set, as the choice of basis is not fundamental; the one with P_x gave the best results.) The measurements are modeled as destructive measurements, which means that operations acting on a particular qubit are no longer available after a measurement has been performed on that qubit, thereby reducing the number of actions that the agent can choose. In order to reduce the huge action space further, the requirement that the final state of one sequence of gates needs to be a two-qubit state shared between A and B is enforced by removing actions that would destructively measure all qubits on one side. The accept and reject actions are essential because they allow identification of successful branches.

Percepts: The agent uses only the previous actions of the current trial as a percept. Z-measurements with different outcomes produce different percepts.

Reward: The protocol suggested by the agent is performed recurrently ten times. This is done to ensure that the solution found is a viable protocol for recurrent application, because it is possible that a single step of the protocol might increase the fidelity but further applications of the protocol could undo that improvement. The reward function is given by $R = \max\left(0, \text{const} \times \sqrt[10]{\prod_{i=1}^{10} p_i}\Delta F\right)$, where p_i is the success probability (i.e., the combined probability of the accepted branches) of the ith step, ΔF is the increase in fidelity after ten steps, and the constant is chosen such that the known protocols [[9] or [10]] would receive a reward of 1.

Problem-specific techniques: To evaluate the performance of an entanglement-purification protocol that is applied in a recurrent fashion, it is necessary to know which actions are performed and, especially, whether the protocol should be considered successful for all possible measurement outcomes. Therefore, it is not sufficient to use the same approach as for the teleportation challenge and simply consider one particular measurement outcome for each trial. Instead, the agent is required to choose actions for all possible measurement outcomes every time it chooses a measurement action. This means that we keep track of multiple separate branches (and the associated probabilities) with different states of the environment. The average density matrix of the branches that the agent decides to keep is the state that is used for the next purification step. We choose to do things this way because it allows us to obtain a complete protocol that can be evaluated in each trial, and the agent is rewarded according to the performance of the whole protocol.

2. Discussion

As discussed in the main text, the agent finds an entanglement-purification protocol that is equivalent to the DEJMPS protocol [10] for an even number of purification steps.

Let us briefly recap how the DEJMPS protocol works: Initially, we have two copies of a state ρ that is diagonal in the Bell basis and can be written with coefficients λ_{ij}:

$$\rho = \lambda_{00}\left|\Phi^+\right\rangle\!\left\langle\Phi^+\right| + \lambda_{10}\left|\Phi^-\right\rangle\!\left\langle\Phi^-\right| + \lambda_{01}\left|\Psi^+\right\rangle\!\left\langle\Psi^+\right| + \lambda_{11}\left|\Psi^-\right\rangle\!\left\langle\Psi^-\right|. \tag{C1}$$

The effect of the multilateral CNOT operation $\text{CNOT}^{A_1\to A_2}\otimes \text{CNOT}^{B_1\to B_2}$, followed by measurements in the computational basis on A_2 and B_2 and postselected for coinciding measurement results, is

$$\widetilde{\lambda}_{00} = \frac{\lambda_{00}^2+\lambda_{10}^2}{N}, \quad \widetilde{\lambda}_{10} = \frac{2\lambda_{00}\lambda_{10}}{N}, \qquad \widetilde{\lambda}_{01} = \frac{\lambda_{01}^2+\lambda_{11}^2}{N}, \quad \widetilde{\lambda}_{11} = \frac{2\lambda_{01}\lambda_{11}}{N}, \tag{C2}$$

where $\widetilde{\lambda}_{ij}$ denotes the new coefficient after the procedure, and $N = (\lambda_{00}+\lambda_{10})^2 + (\lambda_{01}+\lambda_{11})^2$ is a normalization constant and also the probability of success. Without any additional intervention, if this map is applied recurrently, not only is the desired coefficient λ_{00} (the fidelity) amplified, but also both λ_{00} and λ_{10}.

To avoid this and amplify only the fidelity with respect to $|\Phi^+\rangle$, the DEJMPS protocol calls for the application of $\sqrt{-iX}\otimes\sqrt{iX}$ to both copies of ρ before applying the multilateral CNOTs and performing the measurements. The effect of this operation is to exchange the two coefficients λ_{10} and λ_{11}, thus preventing the unwanted amplification of λ_{10}. So, the effective map in each entanglement-purification step is the following:

$$\widetilde{\lambda}_{00} = \frac{\lambda_{00}^2+\lambda_{11}^2}{N}, \quad \widetilde{\lambda}_{10} = \frac{2\lambda_{00}\lambda_{11}}{N}, \qquad \widetilde{\lambda}_{01} = \frac{\lambda_{01}^2+\lambda_{10}^2}{N}, \quad \widetilde{\lambda}_{11} = \frac{2\lambda_{01}\lambda_{10}}{N}, \tag{C3}$$

with $N = (\lambda_{00}+\lambda_{11})^2 + (\lambda_{01}+\lambda_{10})^2$.

In contrast, the solution found by the agent calls for $\sqrt{-iX}\otimes\sqrt{-iX}$ to be applied, which exchanges two different coefficients λ_{00} and λ_{01} instead, for an effective map

$$\widetilde{\lambda}_{00} = \frac{\lambda_{01}^2+\lambda_{10}^2}{N}, \quad \widetilde{\lambda}_{10} = \frac{2\lambda_{01}\lambda_{10}}{N}, \qquad \widetilde{\lambda}_{01} = \frac{\lambda_{00}^2+\lambda_{11}^2}{N}, \quad \widetilde{\lambda}_{11} = \frac{2\lambda_{00}\lambda_{11}}{N}, \tag{C4}$$

and $N = (\lambda_{01}+\lambda_{10})^2 + (\lambda_{00}+\lambda_{11})^2$. Note that the maps given by Eqs. (C3) and (C4) are identical except that the

roles of $\widetilde{\lambda}_{k0}$ and $\widetilde{\lambda}_{k1}$ are exchanged. It is clear that applying the agent's map twice has the same effect as applying the DEJMPS protocol twice, which means that they are equivalent for an even number of recurrence steps.

As a side note, the other protocol found by the agent as described in the main text applies such an additional operation before each entanglement-purification step as well: applying $H \otimes H$ to ρ exchanges λ_{10} and λ_{01}. This also yields a successful entanglement-purification protocol, but with slightly worse performance.

3. Automatic-depolarization variant

We also investigate a variant where, after each purification step, the state is automatically depolarized before the protocol is applied again. This means that if the first step brings the state up to the new fidelity F', it is then brought to the form $F'\left|\Phi^+\right\rangle\!\left\langle\Phi^+\right| + [(1-F')/3]\left(\left|\Psi^+\right\rangle\!\left\langle\Psi^+\right| + \left|\Phi^-\right\rangle\!\left\langle\Phi^-\right| + \left|\Psi^-\right\rangle\!\left\langle\Psi^-\right|\right)$. This can always be achieved without changing the fidelity [9].

In Fig. 8, the reward obtained by 100 agents in this alternative scenario is shown. The successful protocols consist of applying $\mathrm{CNOT}^{A_1\to A_2} \otimes \mathrm{CNOT}^{B_1\to B_2}$ followed by measuring qubits A_2 and B_2 in the computational basis. Optionally, some additional local operations that do not change the fidelity itself can be added, as the effect of those is undone by the automatic depolarization. Similarly to the scenario described in the main text, there are some solutions that accept only one branch as successful, which means that they get only half the reward, as the success probability in each step is halved (center peak in Fig. 8). The protocols for which two relevant branches are accepted are equivalent to the entanglement-purification protocol presented in Ref. [9].

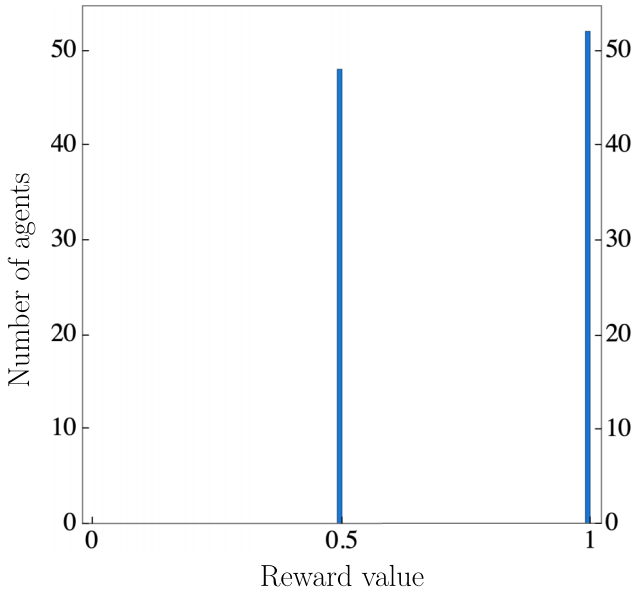

FIG. 8. Entanglement-purification environment with automatic depolarization after each purification step. The figure shows the rewards obtained by 100 agents for the protocols found after 5×10^5 trials.

APPENDIX D: QUANTUM REPEATER

1. Description of environment

The setup is depicted in Fig. 4(a). The repeater stations share entangled states with their neighbors via noisy channels, which results in pairs with an initial fidelity $F = 0.75$. The previous two protocols are now available as the elementary actions in this more complex scenario.

Goal: Entangled pair between the leftmost and the rightmost station with fidelity above threshold fidelity $F_{\mathrm{th}} = 0.9$.

Actions:

(a) Purify a pair with one entanglement-purification step. (Left pair, right pair, the long-distance pair that arises from entanglement swapping.)

(b) Entanglement swapping at the middle station.

We use the protocol given in Ref. [9] for this, as it is computationally easier to handle. For practical applications, it would be advisable to use a more efficient entanglement-purification protocol.

Percepts: Current position of the pairs and fidelity of each pair.

Reward function: $R = (\mathrm{const}/\mathrm{resources})^2$. The whole path is rewarded in full. The reward constant is obtained from an initial guess using the working-fidelity strategy described in Sec. E 4.

APPENDIX E: SCALING REPEATER

1. Description of environment

In addition to the elementary actions from the distance-2 quantum repeater discussed above, we provide the agent with the ability to delegate solving smaller-scale problems of the same type to other agents, therefore splitting the problem into smaller parts. Then, the sequence of actions found is applied as one *block action*, as illustrated in Fig. 5(a).

Goal: Entangled pair between the leftmost and the rightmost station with fidelity above the threshold fidelity $F_{\mathrm{th}} = 0.9$.

Actions:

(a) Purify a pair with one entanglement-purification step.

(b) Entanglement swapping at a station.

(c) *Block actions* of shorter lengths.

So, for a setup with L repeater links, initially there are L purification actions and $L-1$ entanglement-swapping actions. Of course, the purification actions have to be adjusted every time an entanglement swapping is performed to include the new, longer-distance pair. The block actions can be applied at different locations; for example, a length-2 block action can initially be applied at $L-1$ different positions (which also have to be adjusted to include longer-distance pairs as entanglement-swapping actions are chosen). So, it is easy to see how the action space quickly becomes much larger as L increases.

Percepts: Current position of the pairs and fidelity of each pair.

Reward: Again we use the resource-based reward function $R = (\text{const}/\text{resources})^2$, as this is the metric that we would like to optimize. The whole path is rewarded in full. The reward constant is obtained from an initial guess (see Sec. E 4) and adjusted downward once a better solution is found such that the maximum possible reward from one trial is 1.

Comment on block actions: The main agent can use block actions for a wide variety of situations at different stages of the protocol. This means that the subagents are tasked with finding block actions for a wide variety of initial fidelities, and so a new problem needs to be solved for each new situation. In order to speed up the trials, we save situations that have already been solved by subagents in a large table, and reuse the action sequence found if a similar situation arises.

Symmetric variant.—We force a symmetric protocol by modifying the actions as follows:

Actions:

(a) Purify all pairs with one entanglement-purification step.

(b) Entanglement swapping at every second active station.

(c) *Block actions* of shorter length that have been obtained in the same symmetrized manner.

2. Additional results and discussion

We also investigate different starting situations for this setup. Here we discuss two of them:

First, we also apply an agent that is not restricted to symmetric protocols to a symmetric starting situation. The results for initial fidelities $F = 0.7$ can be found in Figs. 9(a) and 9(b). In general, the agent finds solutions that are very close but not equal to those obtained with the working-fidelity strategy described in Sec. E 4. Remarkably, for some reliability parameters p the agent even finds a solution that is slightly better, by switching the order of operations around a little, or a threshold effect, where a solution that omits an entanglement-purification step on one of the pairs is still enough to reach the desired threshold fidelity.

Finally, we also look at a situation that is highly asymmetric, with starting fidelities (0.95, 0.9, 0.6, 0.9, 0.95, 0.95, 0.9, 0.6). Thus there are high-quality links on most connections, but two links suffer from very high levels of noise. The results depicted in Figs. 9(c) and 9(d) show that the advantage over a working-fidelity strategy is even more pronounced.

3. Repeater stations for setups with memory errors

As mentioned in the main text, in order to properly account for imperfections in the quantum memories, we need to know the distance between repeater stations.

In Sec. IV C, we look at a total distance of just below 78.9 km with seven intermediate repeater stations located at the positions shown in Table II. Together with the distance-dependent error model introduced in that section, these give rise to the asymmetric initial fidelities (0.8, 0.6, 0.8, 0.8, 0.7, 0.8, 0.8, 0.6), which we also use for a setup with perfect memories.

In Sec. IV D, we consider a list of seven possible locations to position repeater stations at, which we obtain by scaling down the previous scenario to a total distance of 20 km. We choose such a comparatively short distance in

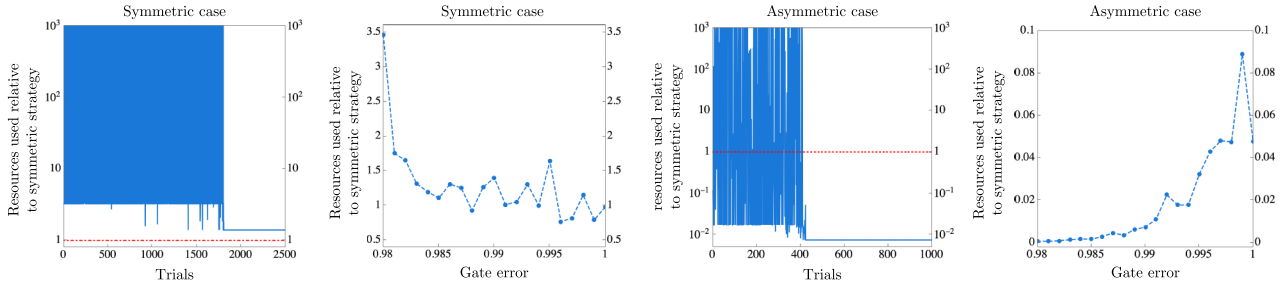

FIG. 9. (a),(b) Scaling repeater with eight repeater links with symmetric initial fidelities of 0.7. (a) Best solution found by an agent for gate reliability $p = 0.99$. (b) Relative resources used by the agent's solution compared with the working-fidelity strategy for different gate reliability parameters. (c),(d) Scaling repeater with eight repeater links with very asymmetric initial fidelities (0.95, 0.9, 0.6, 0.9, 0.95, 0.95, 0.9, 0.6). (c) Best solution found by an agent for gate reliability $p = 0.99$. (d) Relative resources used by the agent's solution compared with the working-fidelity strategy for different gate reliability parameters.

TABLE II. Positions of the repeater stations in Sec. IV C. The terminal stations between which shared entanglement is to be established are located at positions 0 and 78.9 km.

Repeater-station index	Position
1	6.8 km
2	23.6 km
3	30.4 km
4	37.2 km
5	48.5 km
6	55.3 km
7	62.1 km

order to make protocols with only one added repeater station a viable solution. The positions of the seven possible locations are listed in Table III.

4. Working-fidelity strategy

This is the strategy that we use to determine the reward constants for the quantum-repeater environments, and was presented in Ref. [12]. This strategy leads to a resource requirement per repeater station that grows logarithmically with the distance.

For repeater lengths with 2^k links, the working-fidelity strategy is a fully nested scheme and can therefore be stated easily:

(1) Pick a working fidelity F_w.
(2) Purify all pairs until their fidelity is $F \geq F_w$.
(3) Perform entanglement swapping at every second active station such that there are half as many repeater links left.
(4) Repeat from step 2 until only one pair remains (and therefore the outermost stations are connected).

We then optimize the choice of F_w such that the resources are minimized for the given scenario.

As we have to deal with repeater lengths that are not a power of 2 as part of the delegated subsystems discussed in the main text, the strategy is adjusted as follows for such cases:

(1) Pick a working fidelity F_w.
(2) Purify all pairs until their fidelity is $F \geq F_w$.
(3) Perform entanglement swapping at the station with the smallest combined distance of their left and right pairs (e.g., 2 links + 3 links). If multiple stations are equal in this regard, pick the leftmost station.
(4) Repeat from step 2 until only one pair remains (and therefore the outermost stations are connected).

Then, we again optimize the choice of F_w such that the resources are minimized for the given scenario.

TABLE III. Possible locations at which repeater stations can be positioned in Sec. IV D. The terminal stations between which shared entanglement is to be established are located at positions 0 and 20 km.

Possible location index	Position
1	1.73 km
2	5.98 km
3	7.71 km
4	9.44 km
5	12.29 km
6	14.02 km
7	15.75 km

TABLE IV. Run times for our scenarios on a machine with four Sixteen-Core AMD Opteron 6274 CPUs.

Scenario	Run time
Teleportation [Fig. 2(b) or 2(c)]	Clifford, $\sim$ 5 h; universal, $\sim$ 3 d
Teleportation variant 1	$<$ 1 h
Teleportation variant 2 [Fig. 2(e)]	$\sim$ 93 h
Entanglement purification (Fig. 3)	$\sim$ 7 h
Quantum repeater (Fig. 4)	$<$ 1 h
Scaling symmetric case [Figs. 5(b) and 5(c)]	$\sim$ 3 h
Scaling asymmetric case [Figs. 5(d) and 5(e)]	$\sim$ 24 h per data point
Scaling with memory errors (Fig. 6)	$\sim$ 25 h per data point
Repeater-station positions (Table I)	$\sim$ 4.5 h

APPENDIX F: COMPUTATIONAL RESOURCES

All numerical calculations are performed on a machine with four Sixteen-Core AMD Opteron 6274 CPUs. In Table IV, we provide the run times of the calculations presented in this paper. Memory requirements are insignificant for all of our scenarios.

[1] H. J. Kimble, The quantum internet, Nature **453**, 1023 (2008).
[2] S. Wehner, D. Elkouss, and R. Hanson, Quantum internet: A vision for the road ahead, Science **362**, eaam9288 (2018).
[3] P. W. Shor and J. Preskill, Simple Proof of Security of the BB84 Quantum key Distribution Protocol, Phys. Rev. Lett. **85**, 441 (2000).
[4] R. Renner, PhD thesis, ETH Zurich, 2005.
[5] Y.-B. Zhao and Z.-Q. Yin, Apply current exponential de Finetti theorem to realistic quantum key distribution, Int. J. Mod. Phys.: Conf. Ser. **33**, 1460370 (2014).

[6] D. Gottesman and H.-K. Lo, Proof of security of quantum key distribution with two-way classical communications, IEEE Trans. Inf. Theory **49**, 457 (2003).

[7] H.-K. Lo, A simple proof of the unconditional security of quantum key distribution, J. Phys. A: Math. Gen. **34**, 6957 (2001).

[8] C. H. Bennett, G. Brassard, C. Crépeau, R. Jozsa, A. Peres, and W. K. Wootters, Teleporting an Unknown Quantum State via Dual Classical and Einstein-Podolsky-Rosen Channels, Phys. Rev. Lett. **70**, 1895 (1993).

[9] C. H. Bennett, G. Brassard, S. Popescu, B. Schumacher, J. A. Smolin, and W. K. Wootters, Purification of Noisy Entanglement and Faithful Teleportation via Noisy Channels, Phys. Rev. Lett. **76**, 722 (1996).

[10] D. Deutsch, A. Ekert, R. Jozsa, C. Macchiavello, S. Popescu, and A. Sanpera, Quantum Privacy Amplification and the Security of Quantum Cryptography Over Noisy Channels, Phys. Rev. Lett. **77**, 2818 (1996).

[11] W. Dür and H. J. Briegel, Entanglement purification and quantum error correction, Rep. Prog. Phys. **70**, 1381 (2007).

[12] H.-J. Briegel, W. Dür, J. I. Cirac, and P. Zoller, Quantum Repeaters: The Role of Imperfect Local Operations in Quantum Communication, Phys. Rev. Lett. **81**, 5932 (1998).

[13] S. Russel and P. Norvig, *Artificial Intelligence: A Modern Approach* (Prentice Hall, New Jersey, 2010), 3rd ed.

[14] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA, 2017), 2nd ed.

[15] M. Wiering and M. van Otterlo, eds., *Reinforcement Learning: State of the Art*. Adaptation, Learning, and Optimization Vol. 12 (Springer, Berlin, Germany, 2012).

[16] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, Human-level control through deep reinforcement learning, Nature **518**, 529 (2015).

[17] D. Silver, A. Huang, C. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, Mastering the game of Go with deep neural networks and tree search, Nature **529**, 484 (2016).

[18] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis, A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play, Science **362**, 1140 (2018).

[19] H. J. Briegel and G. De las Cuevas, Projective simulation for artificial intelligence, Sci. Rep. **2**, 400 (2012).

[20] S. Hangl, E. Ugur, S. Szedmak, and J. Piater, in *Proceeding of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 2016*, p. 2799.

[21] S. Hangl, V. Dunjko, H. J. Briegel, and J. Piater, Skill learning by autonomous robotic playing using active learning and exploratory behavior composition, Front. Robot. AI **7**, 42 (2020).

[22] L. Orseau, T. Lattimore, and M. Hutter, in *Algorithmic Learning Theory*, edited by S. Jain, R. Munos, F. Stephan, and T. Zeugmann (Springer Berlin Heidelberg, 2013), p. 158.

[23] H. J. Briegel, On creative machines and the physical origins of freedom, Sci. Rep. **2**, 522 (2012).

[24] A. A. Melnikov, H. Poulsen Nautrup, M. Krenn, V. Dunjko, M. Tiersch, A. Zeilinger, and H. J. Briegel, Active learning machine learns to create new quantum experiments, Proc. Natl. Acad. Sci. USA **115**, 1221 (2018).

[25] V. Dunjko and H. J. Briegel, Machine learning & artificial intelligence in the quantum domain: A review of recent progress, Rep. Prog. Phys. **81**, 074001 (2018).

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

[27] V. Dunjko and P. Wittek, A non-review of quantum machine learning: Trends and explorations, Quantum Views **4**, 32 (2020).

[28] L. O'Driscoll, R. Nichols, and P. A. Knott, A hybrid machine learning algorithm for designing quantum experiments, Quantum Mach. Intell. **1**, 5 (2019).

[29] M. Krenn, M. Erhard, and A. Zeilinger, Computer-inspired quantum experiments, arXiv:2002.09970.

[30] A. A. Melnikov, P. Sekatski, and N. Sangouard, Setting up experimental Bell test with reinforcement learning, arXiv:2005.01697.

[31] L. Cincio, Y. Subaşı, A. T. Sornborger, and P. J. Coles, Learning the quantum algorithm for state overlap, New J. Phys. **20**, 113022 (2018).

[32] J. Bang, J. Ryu, S. Yoo, M. Pawłowski, and J. Lee, A strategy for quantum algorithm design assisted by machine learning, New J. Phys. **16**, 073017 (2014).

[33] A. A. Melnikov, L. E. Fedichkin, and A. Alodjants, Predicting quantum advantage by quantum walk with convolutional neural networks, New J. Phys. **21**, 125002 (2019).

[34] A. A. Melnikov, L. E. Fedichkin, R.-K. Lee, and A. Alodjants, Machine learning transfer efficiencies for noisy quantum walks, Adv. Quantum Technol. **3**, 1900115 (2020).

[35] C. Moussa, H. Calandra, and V. Dunjko, To quantum or not to quantum: Towards algorithm selection in near-term quantum optimization, arXiv:2001.08271.

[36] T. Fosel, P. Tighineanu, T. Weiss, and F. Marquardt, Reinforcement Learning with Neural Networks for Quantum Feedback, Phys. Rev. X **8**, 031084 (2018).

[37] H. Xu, J. Li, L. Liu, Y. Wang, H. Yuan, and X. Wang, Generalizable control for quantum parameter estimation through reinforcement learning, npj Quantum Inf. **5**, 82 (2019).

[38] F. Schäfer, M. Kloc, C. Bruder, and N. Lörch, A differentiable programming method for quantum control, arXiv:2002.08376.

[39] M. Tiersch, E. J. Ganahl, and H. J. Briegel, Adaptive quantum computation in changing environments using projective simulation, Sci. Rep. **5**, 12874 (2015).

[40] H. Poulsen Nautrup, N. Delfosse, V. Dunjko, H. J. Briegel, and N. Friis, Optimizing quantum error correction

codes with reinforcement learning, Quantum **3**, 215 (2019).
[41] R. Sweke, M. S. Kesselring, E. P. van Nieuwenburg, and J. Eisert, Reinforcement learning decoders for fault-tolerant quantum computation, arXiv:1810.07207.
[42] A. Valenti, E. van Nieuwenburg, S. Huber, and E. Greplova, Hamiltonian learning for quantum error correction, Phys. Rev. Res. **1**, 033092 (2019).
[43] G. Liu, M. Chen, Y.-X. Liu, D. Layden, and P. Cappellaro, Repetitive readout enhanced by machine learning, Mach. Learn.: Sci. Technol. **1**, 015003 (2020).
[44] S. Yu, F. Albarran-Arriagada, J. C. Retamal, Y.-T. Wang, W. Liu, Z.-J. Ke, Y. Meng, Z.-P. Li, J.-S. Tang, E. Solano, L. Lamata, C.-F. Li, and G.-C. Guo, Reconstruction of a photonic qubit state with reinforcement learning, Adv. Quantum Technol. **2**, 1800074 (2019).
[45] J. Carrasquilla, G. Torlai, R. G. Melko, and L. Aolita, Reconstructing quantum states with generative models, Nat. Mach. Intell. **1**, 155 (2019).
[46] G. Torlai, B. Timar, E. P. L. van Nieuwenburg, H. Levine, A. Omran, A. Keesling, H. Bernien, M. Greiner, V. Vuletić, M. D. Lukin, R. G. Melko, and M. Endres, Integrating Neural Networks with a Quantum Simulator for State Reconstruction, Phys. Rev. Lett. **123**, 230504 (2019).
[47] G. Carleo and M. Troyer, Solving the quantum many-body problem with artificial neural networks, Science **355**, 602 (2017).
[48] X. Gao and L.-M. Duan, Efficient representation of quantum many-body states with deep neural networks, Nat. Commun. **8**, 662 (2017).
[49] A. Canabarro, S. Brito, and R. Chaves, Machine Learning Nonlocal Correlations, Phys. Rev. Lett. **122**, 200401 (2019).
[50] R. S. Sutton, in *Proceedings of the 7th International Conference on Machine Learning, 1990*, p. 216.
[51] P. Mirowski, R. Pascanu, F. Viola, H. Soyer, A. Ballard, A. Banino, M. Denil, R. Goroshin, L. Sifre, K. Kavukcuoglu, D. Kumaran, and R. Hadsell, Learning to navigate in complex environments, arXiv:1611.03673.
[52] T. Mannucci and E.-J. van Kampen, in *Proceedings of the IEEE Symposium Series on Computational Intelligence, 2016*.
[53] In the teleportation task, the shortest possible sequence of gates is equal to 14, and in each step of this sequence there are at least seven possible gates that can be applied.
[54] J. Mautner, A. Makmal, D. Manzano, M. Tiersch, and H. J. Briegel, Projective simulation for classical learning agents: A comprehensive investigation, New Gener. Comput. **33**, 69 (2015).
[55] A. Makmal, A. A. Melnikov, V. Dunjko, and H. J. Briegel, Meta-learning within projective simulation, IEEE Access **4**, 2110 (2016).
[56] A. A. Melnikov, A. Makmal, V. Dunjko, and H. J. Briegel, Projective simulation with generalization, Sci. Rep. **7**, 14430 (2017).
[57] K. Ried, B. Eva, T. Müller, and H. J. Briegel, How a minimal learning agent can infer the existence of unobserved variables in a complex environment, arXiv:1910.06985.
[58] J. Kempe, Quantum random walks: An introductory overview, Contemp. Phys. **44**, 307 (2003).
[59] S. E. Venegas-Andraca, Quantum walks: A comprehensive review, Quantum Inf. Process. **11**, 1015 (2012).
[60] G. D. Paparo, V. Dunjko, A. Makmal, M. A. Martin-Delgado, and H. J. Briegel, Quantum Speed-Up for Active Learning Agents, Phys. Rev. X **4**, 031002 (2014).
[61] S. Jerbi, H. Poulsen Nautrup, L. M. Trenkwalder, H. J. Briegel, and V. Dunjko, A framework for deep energy-based reinforcement learning with quantum speed-up, arXiv:1910.12760.
[62] F. Flamini, A. Hamann, S. Jerbi, L. M. Trenkwalder, H. Poulsen Nautrup, and H. J. Briegel, Photonic architecture for reinforcement learning, New J. Phys. **22**, 045002 (2020).
[63] V. Dunjko, N. Friis, and H. J. Briegel, Quantum-enhanced deliberation of learning agents using trapped ions, New J. Phys. **17**, 023006 (2015).
[64] N. Friis, A. A. Melnikov, G. Kirchmair, and H. J. Briegel, Coherent controlization using superconducting qubits, Sci. Rep. **5**, 18036 (2015).
[65] T. Sriarunothai, S. Wölk, G. S. Giri, N. Friis, V. Dunjko, H. J. Briegel, and C. Wunderlich, Speeding-up the decision making of a learning agent using an ion trap quantum processor, Quantum Sci. Technol. **4**, 015014 (2018).
[66] A. A. Melnikov, A. Makmal, and H. J. Briegel, Benchmarking projective simulation in navigation problems, IEEE Access **6**, 64639 (2018).
[67] J. Clausen, W. L. Boyajian, L. M. Trenkwalder, V. Dunjko, and H. J. Briegel, On the convergence of projective-simulation-based reinforcement learning in Markov decision processes, arXiv:1910.11914.
[68] Projective simulation Github repository, github.com/qic-ibk/projectivesimulation (retrieved April 8, 2020).
[69] A. Jamiołkowski, Linear transformations which preserve trace and positive semidefiniteness of operators, Rep. Math. Phys. **3**, 275 (1972).
[70] A. Gilchrist, N. K. Langford, and M. A. Nielsen, Distance measures to compare real and ideal quantum processes, Phys. Rev. A **71**, 062310 (2005).
[71] X. Zhou, D. W. Leung, and I. L. Chuang, Methodology for quantum logic gate construction, Phys. Rev. A **62**, 052316 (2000).
[72] This value is usually used for photon loss in a glass fiber, but here we use it for the decay in fidelity instead, which results in extremely noisy channels.
[73] K. Azuma, K. Tamaki, and H.-K. Lo, All-photonic quantum repeaters, Nat. Commun. **6**, 6787 (2015).
[74] D. Gottesman, Theory of fault-tolerant quantum computation, Phys. Rev. A **57**, 127 (1998).
[75] P. O. Boykin, T. Mor, M. Pulver, V. Roychowdhury, and F. Vatan, in *Proceedings of the 40th Annual Symposium on Foundations of Computer Science*, FOCS '99 (IEEE Computer Society, Washington, DC, 1999), p. 486.