Experimental quantum speed-up in reinforcement learning agents

V. Saggio[1*], B. E. Asenbeck[1], A. Hamann[2], T. Strömberg[1], P. Schiansky[1], V. Dunjko[3], N. Friis[4], N. C. Harris[5,6], M. Hochberg[7], D. Englund[6], S. Wölk[2,8], H. J. Briegel[2,9], and P. Walther[1,10]

[1]*University of Vienna, Faculty of Physics, Vienna Center for Quantum Science and Technology (VCQ), Boltzmanngasse 5, Vienna A-1090, Austria*
[2]*Institut für Theoretische Physik, Universität Innsbruck, Technikerstraße 21a, 6020 Innsbruck, Austria*
[3]*LIACS, Leiden University, Niels Bohrweg 1, 2333 CA Leiden, Netherlands*
[4]*Institute for Quantum Optics and Quantum Information - IQOQI Vienna, Austrian Academy of Sciences, Boltzmanngasse 3, A-1090 Vienna, Austria*
[5]*Lightmatter, 60 State Street, Floor 10, Boston, MA, 02130, USA*
[6]*Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*
[7]*NOKIA Corporation, New York, NY, USA*
[8]*Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR), Institut für Quantentechnologien, Söflingerstr. 100, 89077 Ulm, Germany*
[9]*Fachbereich Philosophie, Universität Konstanz, Fach 17, 78457 Konstanz, Germany and*
[10]*Christian Doppler Laboratory for Photonic Quantum Computer, Faculty of Physics, University of Vienna, Vienna, Austria.*

Increasing demand for algorithms that can learn quickly and efficiently has led to a surge of development within the field of artificial intelligence (AI). An important paradigm within AI is reinforcement learning (RL), where agents interact with environments by exchanging signals via a communication channel. Agents can learn by updating their behaviour based on obtained feedback. The crucial question for practical applications is how fast agents can learn to respond correctly. An essential figure of merit is therefore the learning time. While various works have made use of quantum mechanics to speed up the agent's decision-making process, a reduction in learning time has not been demonstrated yet. Here we present a RL experiment where the learning of an agent is boosted by utilizing a quantum communication channel with the environment. We further show that the combination with classical communication enables the evaluation of such an improvement, and additionally allows for optimal control of the learning progress. This novel scenario is therefore demonstrated by considering hybrid agents, that alternate between rounds of quantum and classical communication. We implement this learning protocol on a compact and fully tunable integrated nanophotonic processor. The device interfaces with telecom-wavelength photons and features a fast active feedback mechanism, allowing us to demonstrate the agent's systematic quantum advantage in a setup that could be readily integrated within future large-scale quantum communication networks.

INTRODUCTION

Rapid advances in the field of machine learning (ML) and in general artificial intelligence (AI) have been paving the way towards intelligent algorithms and automation. An important paradigm within AI is reinforcement learning (RL), where agents interact with an environment and 'learning' is facilitated by feedback exchanges [1]. RL has applications in many sectors, from robotics (e.g. robot control [2], text and speech recognition [3]) to the healthcare domain (e.g. finding optimal treatment policies [4]) to simulation of brain-like computing [5] and neural networks [6, 7]. RL is also at the heart of the celebrated AlphaGo algorithm [8], able to beat even the most skilled human players at the game of Go.

At the same time, thanks to the capability of quantum physics to outperform classical algorithms, the development of quantum technologies has been experiencing remarkable progress [9]. Quantum physics offers improved performance in a variety of applications, from quantum-enhanced sensing [10] to secure communication [11] and quantum information processing [12]. Quantum mechanics also inspires new advantageous algorithms in ML, and

in particular RL [13, 14]. In fact, RL has been successfully implemented to aid in problems encountered in quantum information processing itself, e.g. decoding of errors [15–17], quantum feedback [18], adaptive code-design [19], and even the design of new quantum experiments [20, 21]. Conversely, quantum technology has been employed to enable a quadratically faster decision-making process for RL agents via the quantization of their internal hardware [22–25].

In all of these RL applications, the interaction between the agent and the environment is performed via entirely classical communication. Here we consider a novel RL setting where the agent and the environment can also communicate via a quantum channel [26]. In particular, we introduce a hybrid agent that enables quantum as well as classical information transfer, and thus combines quantum amplitude amplification with a classical update policy via a feedback loop. Within such a scenario, it is possible for the first time to quantify and achieve a quantum speed-up in the agent's learning time with respect to RL schemes based on classical communication only. The learning time is defined as the average number of interactions until the agent receives a reward ('accomplishes
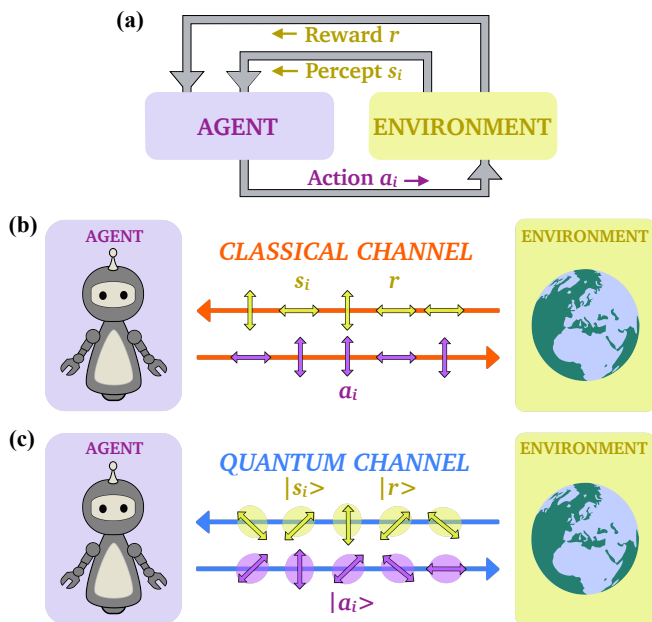
FIG. 1: **Schematic of a learning agent.** **(a)** An agent interacts with an environment by receiving perceptual input s_i and outputting consequent actions a_i. In the case of the correct a_i being performed, the environment issues a reward r that the agent uses to enhance its performance in the next round of interaction. **(b)** Agent and environment interacting via a classical channel, where communication is only possible via a fixed preferred basis (e.g. 'vertical' or 'horizontal'). **(c)** Agent and environment interacting via a quantum channel, where arbitrary superposition states are exchanged.

a specific task'). Such improvement represents a remarkable advantage over previously implemented protocols, and may prove crucial in the development of increasingly complex learning devices [5–7].

We demonstrate this protocol using a fully-programmable nanophotonic processor interfaced with photons at telecom wavelengths. The setup enables the implementation of an active feedback mechanism, thus proving suitable for demonstrations of RL algorithms. Moreover, such a photonic platform holds the potential of being fully interfaceable with future quantum communication networks thanks to the photons' telecom wavelengths. In fact, a long-standing goal of the current development of quantum communication technologies lies in establishing a form of 'quantum internet' [27, 28] — a highly interconnected network able to distribute and manipulate complex quantum states via fibres and optical links (ground-based or even via satellites). We therefore envisage AI and RL to play an important role in future large-scale quantum communication networks, including a potential quantum internet, much in the same way that AI forms integral parts of the internet today. Our results thus additionally demonstrate the feasibility of integrating quantum mechanical RL speed-ups in future complex quantum networks.

QUANTUM ENHANCEMENT IN REINFORCEMENT LEARNING

The conceptual idea of RL is shown in Fig. 1(a). Here an agent (e.g. physical or web-based) interacts with an environment by receiving perceptual input, called 'percepts' s_i, and outputting specific 'actions' a_i accordingly. 'Rewards' r issued by the environment for correct combinations of percepts and actions incentivize agents to improve their decision-making, and thus to learn [1].

Although RL has already been shown amenable to quantum enhancements, in practical demonstrations the interaction between agents and environments has so far been restricted exclusively to classical communication, meaning that agents are limited to exchanges in a fixed preferred basis, e.g. 'vertical' or 'horizontal' as shown in Fig. 1(b). In general, it has been shown that granting an agent access to quantum hardware (while still restricting it to classical exchanges with the environment) does not reduce the average number of interactions that the agent needs to accomplish its task, although it allows it to output actions quadratically faster [22, 25]. To achieve a reduction in learning times, and thus in sample complexity, fully quantum-mechanical interactions between agent and environment must be considered.

Our goal is to experimentally demonstrate such a quantum-enhanced RL scenario. We start with considering an agent and environment with access to internal quantum (as well as classical) hardware that can interact by coherently exchanging quantum states $|a_i\rangle$, $|s_i\rangle$, and $|r\rangle$, representing actions a_i, percepts s_i, and rewards r, respectively, via a joint quantum channel, as illustrated in Fig. 1(c). In this case, communication is no longer limited to a fixed preferred basis, but allows for an exchange of arbitrary superpositions. Agents react to (sequences of) percepts $|s_{i-1}\rangle$ with (sequences of) actions $|a_i\rangle$ according to a policy $\pi(a_i|s_{i-1})$ that is updated during the learning process. In particular, this quantum framework includes so-called deterministic strictly epochal (DSE) learning scenarios, which we focus on. Here, 'epochs' consist of strings of percepts $\vec{s} = (s_0, \cdots, s_{L-1})$ and actions $\vec{a} = (a_1, \cdots, a_L)$ of fixed length L, and a final reward r, and both $\vec{s} = \vec{s}(\vec{a})$ and $r = r(\vec{a})$ are completely determined by $\vec{a}$ (see Methods, Sec. I). Many interesting environments are epochal, e.g. in applications of RL to quantum physics [21, 29–31] or popular problems such as playing Go [8]. A non-trivial feature of the DSE scenario is that the effective behaviour of the environment can be modelled via a unitary U_E on the action and reward quantum registers (indicated with subscripts A and R, respectively):

$$U_E|\vec{a}\rangle_\mathrm{A}|0\rangle_\mathrm{R} = \begin{cases} |\vec{a}\rangle_\mathrm{A}|1\rangle_\mathrm{R} & \text{if } r(\vec{a}) > 0 \\ |\vec{a}\rangle_\mathrm{A}|0\rangle_\mathrm{R} & \text{else} \end{cases}, \quad (1)$$

which identifies rewarded sequences of actions. The quantum-enhanced agent can use U_E to perform a quantum search for rewarded action sequences and teach the found sequences to a classical agent [26]. We push this
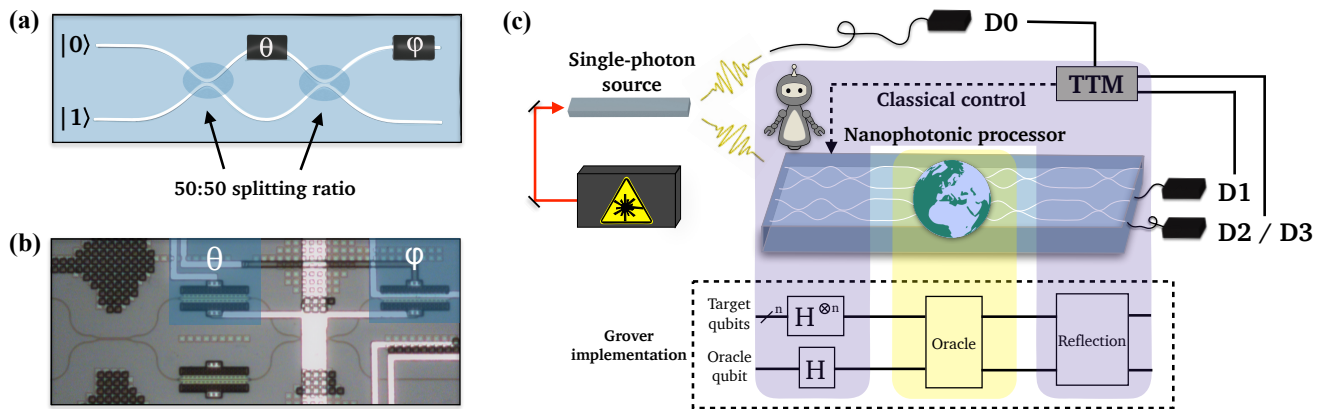
FIG. 2: **Experimental setup. (a)** Single programmable unit consisting of a Mach-Zehnder interferometer (MZI) equipped with two fully tunable phase shifters θ and ϕ. **(b)** Real-life picture of a single MZI in the processor. The third phase shifter in the bottom arm of the interferometer is not used. **(c)** Overview of the setup. A single-photon source generates single-photon pairs at telecom wavelength. One photon is sent straight to a single-photon detector D0, while the other one is coupled into the processor and undergoes the desired computation. It is then detected, in coincidence with the photon in D0, either in detector D1 or D2/D3 after the agent plays the classical/quantum strategy (refer to Fig. 3 for more details). The coincidence events are recorded with a custom-made Time Tagging Module (TTM). Different areas of the processor are assigned to either the agent or the environment. The former one also comprises a classical control updating its policy. They alternately play a Grover-like search to look for the rewarded sequence of actions.

idea further and grant the quantum and the classical part of the agent complete access to each other. In this way, it is possible to create a hybrid agent with a feedback loop between quantum search and classical policy update. This approach allows us to quantify the speed-up in learning time, which is not possible in the general setting discussed in [26] (see Methods, Sec. I).

It is known that near-term quantum devices are generally subject to limited coherence times, restricting the number of quantum gates that can be performed before essential quantum features are lost. However, already a few or even a single coherent query to U_E are enough to construct a quantum-enhanced RL framework that can outperform a classical one.

In a classical epoch, the agent prepares the state $|\vec{a}\rangle_A |0\rangle_R$ where $\vec{a}$ is determined by directly sampling from a classical probability distribution $p(\vec{a})$. From classically interacting with the environment, the agent determines the corresponding reward $r(\vec{a})$ and percepts $\vec{s}(\vec{a})$, and uses the obtained information to update its policy π, thus learning.

In a quantum epoch, the agent prepares the state $|\psi\rangle_A |-\rangle_R$, with $|\psi\rangle = \sum_{\vec{a}} \sqrt{p(\vec{a})} |\vec{a}\rangle$ being determined according to $p(\vec{a})$ dictated by its current classical policy π, and $|-\rangle = (|0\rangle - |1\rangle)/\sqrt{2}$. Both states are sent to the environment, which acts on them via U_E inducing a phase of -1 for all rewarded sequences of actions. The states are sent back to the agent which performs a reflection $U_{Ref} = (2|\psi\rangle\langle\psi|_A - \mathbb{1}_A)$ over the initial superposition state. This leads to an amplitude amplification of rewarded action sequences similar to the Grover's algorithm [32]. Parametrising the classical probability to find a reward via $\sin^2(\xi)$ with $\xi \in [0, 2\pi]$, the steps above followed by a final measurement on the action reg-

ister result with an increased probability $\sin^2(3\xi)$ in a rewarded action sequence. However, this procedure does not reveal the reward or the corresponding sequence of percepts. These can be determined only during classical test epochs, where the measured sequence of actions is used as input state.

The quantum-enhanced agent plays in a hybrid way by alternating between these quantum and classical test epochs. Such agents find rewarded sequences of actions faster, and hence usually learn faster than entirely classical agents. The learning speed-up manifests in a reduced average quantum learning time $\langle T \rangle_Q$, that is, the average number of epochs necessary to achieve a predefined winning probability Q_L. In general, a quadratic improvement in the learning time given by $\langle T \rangle_Q \leq \alpha \sqrt{J \langle T \rangle_C}$ for the hybrid agent can be achieved if the maximal number of coherent interactions between the agent and the environment scales with the problem size. Here, J quantifies the number of rewards the agent needs to find in order to learn and $\langle T \rangle_C$ the average classical learning time. However, already a single (coherent) query to U_E is enough to improve the learning time albeit by a constant factor, as demonstrated in the Methods, Sec. I.

EXPERIMENTAL IMPLEMENTATION

Quantized RL protocols can be compactly realized using state-of-the-art photonic technology [33]. In particular the path towards miniaturization of photonic platforms holds the advantage of providing scalable architectures where many elementary components can be accommodated on small devices. Here we use a programmable nanophotonic processor with dimensions of 4.9 x 2.4 mm, comprising 26 waveguides fabricated to form 88 Mach-

Zehnder interferometers (MZIs) in a trapezoidal configuration. A quantum gate is implemented by a MZI equipped with two thermo-optic phase shifters, one internal to allow for a scan of the output distribution over $\theta \in [0, 2\pi]$ and one external dictating the relative phase $\phi \in [0, 2\pi]$ between the two output modes, as shown in Fig. 2(a). This makes each MZI act as a fully-tunable beam splitter and allows for coherent implementation of sequences of quantum gates. Information is spatially encoded onto two orthogonal modes $|0\rangle = (1, 0)^{\mathrm{T}}$ and $|1\rangle = (0, 1)^{\mathrm{T}}$, which constitute the computational basis. Fig. 2(b) shows how a MZI looks like in our integrated device.

Pairs of single photons are generated in the telecom wavelength band from a single-photon source pumped by laser light at 789.5 nm. One photon is coupled into one waveguide in the processor to perform a particular computation, while the other one is sent straight to a single-photon detector D0 for heralding. The detectors are superconducting nanowires that combine extremely high efficiency (up to ∼90%) with a very short dead time (< 100 ns), thus proving the best candidates for fast feedback at telecom wavelengths. Detection events recorded in D0 and at the output of the processor falling into a temporal window of 1.3 ns are registered with a time tagging module (TTM) as coincidence events (see Methods, Sec. II for more details). Different areas of the processor are assigned to the agent and the environment, who alternately perform the aforementioned steps of the Grover-like amplitude amplification. The agent is further equipped with a classical control mechanism that updates its learning policy. All of this can be seen in Fig. 2(c).

In our experiment, we represent all possible sequences of actions by a single qubit ($|1\rangle_{\mathrm{A}} \hat{=}$ rewarded, $|0\rangle_{\mathrm{A}} \hat{=}$ not rewarded) and use another qubit to encode the reward ($|1\rangle_{\mathrm{R}}$, $|0\rangle_{\mathrm{R}}$). This results in a four-level system where each level is a waveguide path in our processor, as shown in Fig. 3 (which illustrates only the part of the processor needed for our computation).

The probability for the agent to choose a correct sequence of actions is initially set to $\sin^2(\xi) = \varepsilon = 0.01$, representing, for instance, a single rewarded sequence of actions out of 100. After a single photon is coupled into the mode $|0_{\mathrm{A}}0_{\mathrm{R}}\rangle$, the agent puts it into the superposition state $|\psi\rangle_{\mathrm{A}} = (\cos\xi|0\rangle_{\mathrm{A}} + \sin\xi|1\rangle_{\mathrm{A}})|0\rangle_{\mathrm{R}}$ via the unitary U_p. Next, it can decide to play classically or quantum-mechanically.

Classical strategy. The environment flips the reward qubit only if the action qubit is in the correct state via the unitary U_E, as shown in Fig. 3(a). Next, the photon is coupled out and detected in either D1 or D2 with probability $\cos^2(\xi)$ and $\sin^2(\xi)$, respectively. Only if D2 is triggered (i.e. the agent has been rewarded), a classical feedback mechanism updates the policy π, and thus $p(\vec{a})$ and ε, via rules given by the projective simulation framework (see Methods, Sec. I).
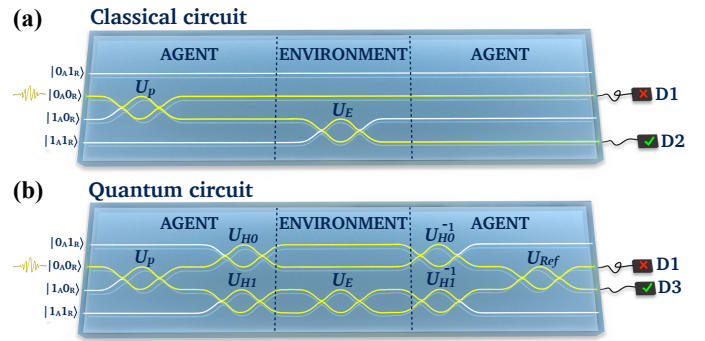
Quantum strategy. Here, the theoretical framework de-

FIG. 3: **Circuit implementation.** One photon is coupled into the $|0_{\mathrm{A}}0_{\mathrm{R}}\rangle$ spatial mode and undergoes different unitaries depending on whether a classical **(a)** or a quantum **(b)** epoch is implemented. The waveguides highlighted in yellow show the photon's possible paths. Identity gates are represented with straight waveguides.

scribed above is exploited to speed up the learning process. As illustrated in Fig. 3(b), after the reward qubit is rotated to $|-\rangle_{\mathrm{R}}$ via the unitaries U_{H0} and U_{H1}, the environment acts as an oracle via U_E. Consecutively, the agent reverses the effect of U_{H0} and U_{H1}, and then it performs the reflection U_{Ref}. Next, a measurement in the computational basis of the action register is performed, leading to detection of a rewarded sequence of actions in D3 with increased probability $\sin^2(3\xi)$. The classical test epoch needed to determine the reward is not experimentally implemented for practical reasons, since the circuit visibility is bigger than 0.99.

In general, any Grover-like algorithm faces a drop in the amplitude amplification after the optimal point is reached. Keeping in mind that each agent will reach this optimal point at different epochs, one can identify the probability $\varepsilon = 0.396$ (see Methods, Sec. I) until which it is beneficial for all agents to use a quantum strategy, as on average they will observe more rewards than with the classical strategy. As soon as this probability is surpassed, it is advantageous for the agents to switch to an entirely classical strategy. This combined quantum-classical strategy thus avoids the winning probability drop without introducing any additional overheads in terms of experimental resources.

RESULTS

Here, we show the experimental comparison between quantum and classical strategies.

At the end of each classical epoch, we record outcomes 1 and 0 for the rewarded and not rewarded behaviour, obtaining a binary sequence whose length equals the number of played epochs for the classical learning strategy, and half of the number of epochs for the quantum strategy. To enable a fair comparison between the quantum and classical scenarios, the reward in the quantum case is distributed — i.e. averaged — over the two epochs needed to determine a test sequence of actions $\vec{a}$ and its
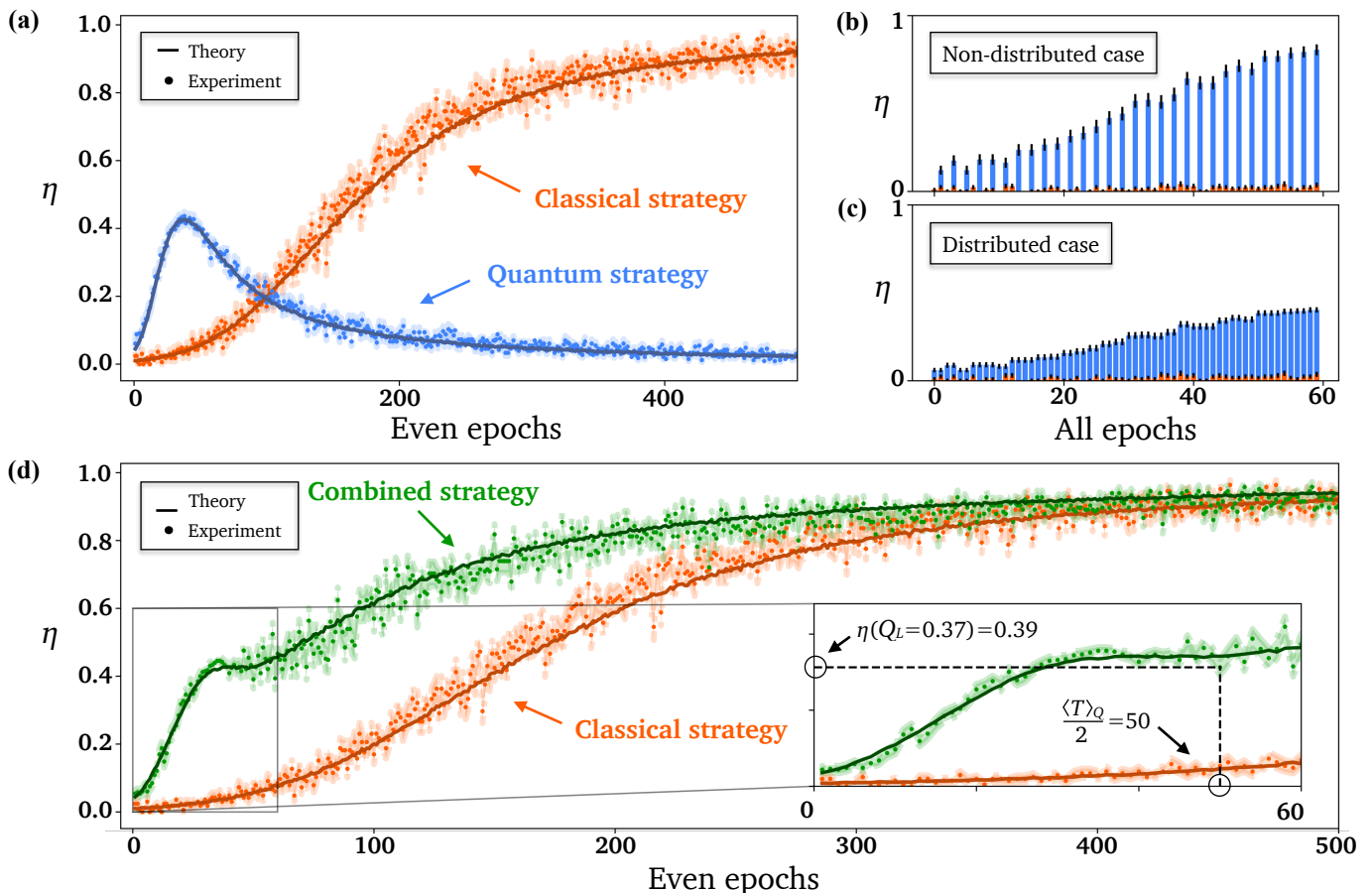
FIG. 4: **Behaviour of the average reward η for different learning strategies.** The solid line represents the theoretical data simulated with $n = 10,000$ agents, while the dots represent the experimental data measured with $n = 165$ agents. The shaded regions indicate the errors associated to each single data point. **(a)** η of agents playing a quantum (blue) or classical (orange) strategy. **(b)** η accounting, in the quantum case, for rewards obtained only every second epoch, compared to **(c)** the case where the reward is distributed over the two epochs an agent needs to acquire it. **(d)** Comparison between the classical and combined quantum-classical case, where an advantage over the classical case is visible. Here, the agents stop the quantum strategy at their best performance (at $\varepsilon = 0.396$) and continue playing classically. The inset shows the point at which the agent playing the quantum strategy reaches the predefined winning probability $Q_L = 0.37$, after $\langle T \rangle_Q = 100$ epochs.

reward. The reward is then averaged over many different agents playing independently of one another. Fig. 4 shows the average reward η for the different learning strategies.

The theoretical data is simulated for $n = 10,000$ agents and the experimental data obtained from $n = 165$. Fig. 4(a) visualizes the quantum improvement originating from the use of amplitude amplification in comparison with a purely classical strategy. For completeness, we also show in Fig. 4(b) and (c) the comparison between not distributing and distributing the reward over two epochs in the quantum case.

As soon as $\varepsilon = 0.396$ is reached, η starts decreasing. Our setup allows the agents to always choose the favorable strategy by switching from quantum to classical as soon as the second becomes more advantageous. In this way, such combined strategy always outperforms the purely classical scenario, as shown in Fig. 4(d).

We define $Q_L = 0.37$ as the winning probability. Note however that any probability below $\varepsilon = 0.396$ can be defined as the winning probability Q_L. The learning time $\langle T \rangle$ for Q_L decreases from $\langle T \rangle_C = 270$ in the classical case to $\langle T \rangle_Q = 100$ in the combined quantum-classical case. This implies a reduction of 63%, which fits well to the theoretical values $T_C^{theory} = 293$ and $T_Q^{theory} = 97$, taking into account small experimental imperfections.

As the combined configuration prevents the average reward from dropping, it is particularly relevant for real-life applications when a Grover search (with its intrinsic overshooting drawback) is implemented. An average reward that saturates at a high niveau without a subsequent drop can also be achieved by employing other algorithms like the fixed-point algorithm [34]. However, they show a less favorable speed-up than Grover-like amplitude amplification, especially considering the limited size of our integrated processor.

In general, an agent can experience a quadratic speed-up in its learning time if it can perform an arbitrary number of coherent Grover iterations [35] even if the number of actual rewarded sequences is unknown [36].

CONCLUSIONS

We have demonstrated a novel RL protocol where information is alternately communicated via a quantum and a classical channel. This makes it possible to evaluate the agent's performance resulting in a speed-up in the learning time, and gain optimal control of the learning process. Learning agents relying on purely classical communication are therefore outperformed. At the same time, emerging photonic circuit technology provides the advantages of compactness, full tunability and low-loss communication, thus proving suitable for RL algorithms where active feedback mechanisms, even over long distances, need to be implemented.

We highlight that although photonic architectures prove particularly suitable for these types of learning algorithms, the presented demonstration is based on a novel theoretical background that is general and applicable to any quantum platform. Moreover, as the field of integrated optics is moving towards the fabrication of increasingly large devices, this demonstration could be extended to more complex quantum circuits allowing for processing of high-dimensional states, and thus paving the way for achieving superior performance in increasingly complex learning devices.

*Corresponding author: valeria.saggio@univie.ac.at.

[1] Sutton, R. S. & Barto, A. G., *Reinforcement Learning: An Introduction* (MIT press, Cambridge, 1998).

[2] Johannink, T. et al. Residual Reinforcement Learning for Robot Control. In *2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada*, 6023-6029 (IEEE, 2019). URL https://doi.org/10.1109/ICRA.2019.8794127.

[3] Tjandra, A., Sakti, S. & Nakamura, S. Sequence-to-Sequence ASR Optimization via Reinforcement Learning. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada*, 5829-5833 (IEEE, 2018). URL https://doi.org/10.1109/ICASSP.2018.8461705.

[4] Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C. & Faisal A. A. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nat. Med.* **24,** 1716-1720 (2018). URL http://doi.org/10.1038/s41591-018-0213-5.

[5] Thakur, C. S. et al. Large-scale neuromorphic spiking array processors: A quest to mimic the brain. *Frontiers in neuroscience* **12,** 891 (2018). URL https://doi.org/10.3389/fnins.2018.00891.

[6] Shen, Y. et al. Deep learning with coherent nanophotonic circuits. *Nat. Photon.* **11,** 441 (2017). URL https://doi.org/10.1038/nphoton.2017.93.

[7] Steinbrecher, G. R., Olson, J. P., Englund, D. & Carolan, J. Quantum optical neural networks. *npj Quantum Information* **5,** 1-9 (2019). URL https://doi.org/10.1038/s41534-019-0174-7.

[8] Silver, D. et al. Mastering the game of Go without human knowledge. *Nature* **550,** 354-359 (2017). URL http://dx.doi.org/10.1038/nature24270.

[9] Arute, F. et al. Quantum supremacy using a programmable superconducting processor. *Nature* **574,** 505-510 (2019). URL https://doi.org/10.1038/s41586-019-1666-5.

[10] Giovannetti, V., Lloyd, S. & Maccone, L. Advances in quantum metrology. *Nat. Photon.* **5,** 222 (2011). URL https://doi.org/10.1038/nphoton.2011.35.

[11] Sergienko, A. V. (ed.) *Quantum Communications and Cryptography* (CRC press, Boca Raton, 2018). URL https://doi.org/10.1201/9781315221120.

[12] Monroe, C. Quantum information processing with atoms and photons. *Nature* **416,** 238-246 (2002). URL https://doi.org/10.1038/416238a.

[13] Dong. D., Chen, C., Li, H & Tarn, T.-J. Quantum Reinforcement Learning. *IEEE T. Syst, Man Cy. B* **38,** 1207-1220 (2008). URL https://doi.org/10.1109/TSMCB.2008.925743.

[14] Dunjko, V. & Briegel, H. J. Machine learning & artificial intelligence in the quantum domain: a review of recent progress. *Rep. Progr. Phys.* **81,** 074001 (2018). URL https://doi.org/10.1088/1361-6633/aab406.

[15] Baireuther, P., O'Brien, T. E., Tarasinski, B. & Beenakker, C. W. J. Machine-learning-assisted correction of correlated qubit errors in a topological code. *Quantum* **2,** 48 (2018). URL https://doi.org/10.22331/q-2018-01-29-48.

[16] Breuckmann, N. P. & Ni, X. Scalable Neural Network Decoders for Higher Dimensional Quantum Codes. *Quantum* **2,** 68-92 (2018). URL https://doi.org/10.22331/q-2018-05-24-68.

[17] Chamberland, C. & Ronagh, P. Deep neural decoders for near term fault-tolerant experiments. *Quant. Sci. Techn.* **3,** 044002 (2018). URL https://doi.org/10.1088/2058-9565/aad1f7.

[18] Fösel, T., Tighineanu, P., Weiss, T. & Marquardt, F. Reinforcement Learning with Neural Networks for Quantum Feedback. *Phys. Rev. X* **8,** 031084 (2018). URL https://doi.org/10.1103.PhysRevX.8.031084.

[19] Poulsen Nautrup, H., Delfosse, N., Dunjko, V., Briegel, H. J. & Friis, N. Optimizing quantum error correction codes with reinforcement learning. *Quantum* **3,** 215 (2019). URL https://doi.org/10.22331/q-2019-12-16-215.

[20] Krenn, M., Malik, M., Fickler, R., Lapkiewicz, R. & Zeilinger, A. Automated Search for new Quantum Experiments. *Phys. Rev. Lett.* **116,** 090405 (2016). URL http://doi.org/10.1103/PhysRevLett.116.090405.

[21] Melnikov, A. A. et al. Active learning machine learns to create new quantum experiments. *Proc. Natl. Acad. Sci. U.S.A.* **115,** 1221-1226 (2018). URL http://dx.doi.org/10.1073/pnas.1714936115.

[22] Paparo, G. D., Dunjiko, V., Makmal, A., Martin-

Delgrado, M. A. & Briegel, H. J. Quantum Speedup for Active Learning Agents. *Phys. Rev. X* **4,** 031002 (2014). URL http://dx.doi.org/10.1103/PhysRevX.4.031002.

[23] Dunjko, V., Friis, N. & Briegel, H. J. Quantum-enhanced deliberation of learning agents using trapped ions. *New J. Phys.* **17,** 023006 (2015). URL https://doi.org/10.1088/1367-2630/17/2/023006.

[24] Jerbi, S., Poulsen Nautrup, H., Trenkwalder, L. M., Briegel, H. J. & Dunjko, V. A framework for deep energy-based reinforcement learning with quantum speed-up (2019). Preprint at https://arxiv.org/abs/1910.12760.

[25] Sriarunothai, T. et al. Speeding-up the decision making of a learning agent using an ion trap quantum processor. *Quantum Sci. Technol.* **4,** 015014 (2019). URL http://dx.doi.org/10.1088/2058-9565/aaef5e.

[26] Dunjko, V., Taylor, J. M., & Briegel, H. J. Quantum-Enhanced Machine Learning. *Phys. Rev. Lett.* **117,** 130501 (2016). URL http://dx.doi.org/10.1103/PhysRevLett.117.130501.

[27] Kimble, H. J. The quantum internet. *Nature* **453,** 1023-1030 (2008). URL https://doi.org/10.1038/nature07127.

[28] Cacciapuoti, A. S. et al. Quantum Internet: Networking Challenges in Distributed Quantum Computing. *IEEE Network* **34,** 137-143 (2020). URL https://doi.org/10.1109/MNET.001.1900092.

[29] Denil, M. et al. Learning to Perform Physics Experiments via Deep Reinforcement Learning (2016). Preprint at https://arxiv.org/abs/1611.01843.

[30] Bukov, M. et al. Reinforcement Learning in Different Phases of Quantum Control. *Phys. Rev. X* **8,** 031086 (2018). URL http://dx.doi.org/10.1103/PhysRevX.8.031086.

[31] Poulsen Nautrup, H. et al. Operationally meaningful representations of physical systems in neural networks (2020). Preprint at https://arxiv.org/abs/2001.00593.

[32] Grover, L. K. Quantum mechanics helps in searching for a needle in a haystack. *Phys. Rev. Lett.* **79,** 325 (1997). URL https://doi.org/10.1103/PhysRevLett.79.325.

[33] Flamini, F. et al. Photonic architecture for reinforcement learning. *New. J. Phys.* **22,** 045002 (2020). URL https://doi.org/10.1088/1367-2630/ab783c.

[34] Yoder, T. J., Low, G. H. & Chuang, I. L. Fixed-Point Quantum Search with an Optimal Number of Queries. *Phys. Rev. Lett.* **113,** 210501 (2014). URL https://doi.org/10.1103/PhysRevLett.113.210501.

[35] Hamann, A. et al. A hybrid agent for quantum-accessible reinforcement learning (2020). In preparation.

[36] Boyer, M., Brassard, G., Hoyer, P. & Tappa, A. Tight bounds on quantum searching. *Fortschr. Phys.* **46,** 493 (1998). URL https://doi.org/10.1002/3527603093.ch10.

[37] Briegel, H. J. & De las Cuevas, G. Projective simulation for artificial intelligence. *Sci. Rep.* **2,** 1-16 (2012). URL http://doi.org/10.1038/srep00400.

[38] Kim, T., Fiorentino, M. & Wong, F. N. C. Phase-stable source of polarization-entangled photons using a polarization Sagnac interferometer. *Phys. Rev. A* **73,** 012316 (2006). URL http://doi.org/10.1103/PhysRevA.73.012316.

[39] Saggio, V. et al. Experimental few-copy multipartite entanglement detection. *Nat. Phys.* **15,** 935-940 (2019). URL http://doi.org/10.1038/s41567-019-0550-4.

[40] Harris, N. C. et al. Quantum transport simulations in a programmable nanophotonic processor. *Nat. Photon.* **11,** 447-452 (2017). URL

http://doi.org/10.1038/nphoton.2017.95.

[41] Marsili, F. et al. Detecting single infrared photons with 93% system efficiency. *Nat. photon.* **7,** 210-214 (2013). URL https://doi.org/10.1038/nphotonc.2013.13.

METHODS

I. Quantum-enhanced learning agents

In this section, we develop an explicit method for combining a classical reinforcement learning agent with quantum amplitude amplification. Our approach for such hybrid agents goes beyond the ideas of Ref. [26] by introducing a feedback loop between classical policy update and quantum amplitude amplification. The developed model allows us to determine achievable improvements in sample complexity, and thus learning time. In addition, the final policy of our agent has similar properties as the policy of the underlying classical agent, leading to a comparable behavior as discussed in more detail in [35] (a paper dedicated to discuss the theoretical background of the hybrid agent presented here more specifically).

In the following, we concentrate on simple deterministic, strictly epochal (DSE) environments. That is, the interaction between the agent and the environment is structured into epochs where each epoch starts with the same percept s_0, and at each time step i an action-percept pair (a_i, s_i) is exchanged. At the end of each epoch, after L action-percept pairs are communicated, a reward $r \in \{0, 1\}$ is given to the agent. The rules of the game are deterministic and time independent, such that performing a specific action a_i after receiving a percept s_{i-1} always leads to the same following percept s_i.

The behaviour of an agent is determined by its policy described by the probability $\pi(a_i|s_{i-1})$ to perform the action a_i given the percept s_{i-1}. In deterministic settings, the percept s_i is completely determined by all previous performed actions $a_1, \cdots, a_i$ such that $\pi(a_i|s_{i-1}) = \pi(a_i|a_1, \cdots, a_{i-1})$. Thus, the behaviour of the agent within one epoch is described by sequences of actions $\vec{a} = (a_1, \cdots, a_L)$ and their corresponding probabilities

$$p(\vec{a}) = \prod_{i=1}^{L} \pi(a_i|a_1, \cdots, a_{i-1}). \tag{A.1}$$

The learning agent implemented here uses a policy based on projective simulation [37], where each sequence of actions $\vec{a}$ is associated with a weight factor $h(\vec{a})$ initialized to $h = 1$. Its policy is defined via the probability distribution

$$p(\vec{a}) = \frac{h(\vec{a})}{\sum_{\vec{a}'} h(\vec{a}')}. \tag{A.2}$$

If the agent has played the sequence $\vec{a}$, it updates the corresponding weight factor via

$$h(\vec{a}) \rightarrow h(\vec{a}) + \lambda r(\vec{a}), \tag{A.3}$$

where $\lambda = 2$ in our experiment. In general, the update method for quantum-enhanced agents is not limited to projective simulation and can be used to enhance any

classical learning scenario, provided that $p(\vec{a})$ exists and that the update rule is solely based on the observed rewards.

We generalize the given learning problem to the quantum domain by encoding different sequences of actions $\vec{a}$ into orthogonal quantum states $|\vec{a}\rangle$ defining our computational basis. In addition, we create a fair unitary oracular variant $\tilde{U}_E$ of the environment [26], whose behaviour is described by

$$\tilde{U}_E|\vec{a}\rangle = \begin{cases} |\vec{a}\rangle & \text{if } r(\vec{a}) = 0 \\ -|\vec{a}\rangle & \text{else} \end{cases}. \tag{A.4}$$

The unitary oracle $\tilde{U}_E$ can be used to perform, for instance, a Grover search or amplitude amplification for rewarded sequences of actions by performing Grover iterations

$$U_G = \left(2|\psi\rangle\langle\psi| - \mathbb{1}\right)\tilde{U}_E \tag{A.5}$$

on an initial state $|\psi\rangle$. A quantum-enhanced agent with access to $\tilde{U}_E$ can thus find rewarded sequences of actions faster than a corresponding classical agent, defined by the same initial policy $\pi(a_i|s_{i-1})$ and update rules, without access to $\tilde{U}_E$.

In general, the optimal number k of Grover iterations $U_G^k|\psi\rangle$ depends on the winning probability

$$q = \sum_{\{\vec{a}|r(\vec{a})\neq 0\}} p(\vec{a}) \tag{A.6}$$

via $k \sim 1/\sqrt{q}$ [32]. In the following, we assume that q is known at least to a good approximation. This is for instance possible if the number of winning sequences $|\vec{a}\rangle$ is known. However, a similar quantum-enhanced learning agent can be also developed if q is unknown by adapting methods from [36] as described in [35].

I.1. Description of the agent

A hybrid agent that learns faster than a classical one can be constructed by alternating between quantum amplitude amplification and classical policy update by repeatedly performing the following steps:

1. Given the classical probability distribution $p(\vec{a})$, determine the success probability q based on the current policy and prepare the quantum state

$$|\psi\rangle = \sum_{\{\vec{a}\}} \sqrt{p(\vec{a})}|\vec{a}\rangle. \tag{A.7}$$

2. Apply the optimal number of Grover iteration $k(\sqrt{q})$ leading to

$$|\psi'\rangle = U_G^k|\psi\rangle \tag{A.8}$$

and perform a measurement on $|\psi'\rangle$ in the computational basis to determine a test sequence of actions $\vec{a}$.

3. Play one classical epoch by using the test sequence $\vec{a}$ determined in step 2 and record the corresponding sequence of percepts $\vec{s}(\vec{a})$ and the reward $r(\vec{a})$.

4. Update the classical policy $\pi(a_k|s_{k-1})$ and the resulting probability distribution $p(\vec{a})$ according to A.3.

There exists a limit Q on q determining whether it is more advantageous for the agent to perform a Grover iteration with $k(\sqrt{q}) \geq 1$ or sample directly from $p(\vec{a})$ (therefore $k(\sqrt{q}) = 0$) to determine $\vec{a}$. In the latter case the agent would only interact classically (as in step 3) with the environment.

After each epoch, a classical agent receives a reward with probability $q = \varepsilon = \sin^2(\xi)$. We assume that the agent can use one epoch to either perform one Grover iteration (step 2) or to determine the reward of a given test sequence $\vec{a}$ (step 3). Thus, for $k = 1$ it receives after every second epoch a reward with probability $\sin^2(3\xi)$. As a consequence, we define the expected average reward of an agent playing a classical strategy as $\eta_C = 2\sin^2(\xi)$ and of an agent playing a quantum strategy with $k = 1$ as $\eta_Q = \sin^2(3\xi)$. For $q < Q$, $\eta_Q > \eta_C$, meaning that the quantum strategy proves advantageous over the classical case. However, as soon as $\eta_C = \eta_Q$ (at $Q = 0.396$), a classical agent starts outperforming a corresponding quantum agent which still performs Grover iterations.

Determining the winning probability q exactly such that $q = \varepsilon$ as in the example presented here is not always possible. In general, additional information like the number of possible solutions and model building helps to perform this task. Note that a Q smaller than 0.396 should be chosen if q can only be estimated up to some range. To circumvent this problem, methods like Grover search with unknown reward probability [36], or fixed-point search [34], can be used to determine if and how many steps of amplitude amplification should be performed [35].

I.2. Learning time

We define the learning time T as the number of epochs an agent needs on average to reach a predefined winning probability Q_L. The above described quantum-enhanced agent can reach the probability Q on average with less epochs than its classical counterpart. However, once both reach Q, they need on average the same number of epochs to reach $Q + \Delta_Q$ with $0 \leq \Delta_Q < 1 - Q$. Therefore, we choose $Q_L \leq Q$ in order to quantify the achievable improvement of a hybrid agent compared to its classical counterpart. In our experiment, we choose $Q_L = 0.37$ to define the learning time.

Let $\ell_J = \{\vec{a}_1, \cdots, \vec{a}_J\}$ be a time-ordered list of all the rewarded sequences of actions an agent has found until it reaches Q_L. Note that the actual policy π_j, and thus p_j, of our agents depend only on the list ℓ_j of observed rewarded sequences of actions, and this is independent of whether it has found them via classical sampling or quantum amplitude amplification. As a result, a classical agent and its quantum-enhanced version are described by the same policy $\pi(\ell_j)$ and behave similar if they have found the same rewarded sequences of actions. However, the quantum-enhanced agent finds them faster.

In general, the actual policy and overall success probability might depend on the found rewarded action sequence. Thus the number J of observed rewarded action sequences necessary to learn might vary. However, this is not the case for the here reported experiment. The learning time in this case can be determined via

$$T(J) = \sum_{j=1}^{J} t_j \tag{A.9}$$

where t_j determines the number of epochs necessary for the agent to find the next rewarded sequence $\vec{a}_j$ after it has observed $j - 1$ rewards. For a purely classical agent, the average time is given by

$$\langle t_j \rangle_C = \frac{1}{q_j} \tag{A.10}$$

where q_j is the actual success probability. This time is quadratically reduced to

$$\langle t_j \rangle_Q = \frac{\alpha}{\sqrt{q_j}} \tag{A.11}$$

for the quantum-enhanced agent. Here, α is a parameter depending only on the number of epochs needed to create one oracle query $\tilde{U}_E$ [26] and on whether q_j is known. In the case considered here, we find $\alpha = \pi/4$. As a consequence, the average learning time for the quantum-enhanced learning agent is given by

$$\langle T(J) \rangle_Q = \sum_{j=1}^{J} \frac{\alpha}{\sqrt{q_j}} \tag{A.12}$$

$$\leq \alpha\sqrt{J}\sqrt{\sum_{j=1}^{J} \frac{1}{q_j}} \tag{A.13}$$

$$\leq \alpha\sqrt{J}\sqrt{\langle T(J) \rangle_C} \tag{A.14}$$

where we used the Cauchy-Schwarz inequality in the second step. The classical learning time typically scales with $\langle T \rangle_C \sim A^K$ for a learning problem with episode length K and the choice between A different actions in each step. The number J of observed rewarded sequences in order to learn depends on the specific policy update and sometimes also on the list ℓ_J of observed rewarded sequences of actions. For an agent sticking with the first rewarded action sequence, we would find $J = 1$. However, typical learning agents are more explorative, and common scalings are $J \sim K$ such that we find for these cases

$$\langle T(J) \rangle_Q \sim \sqrt{\log(\langle T(J) \rangle_C)}\sqrt{\langle T(J) \rangle_C}. \tag{A.15}$$

This is equivalent to a quasi-quadratic speed-up in the learning time if it is possible to perform arbitrary numbers of Grover iterations.

In more general settings, there exist several possible ℓ_J with different length J such that the learning time $\langle T(J) \rangle$ needs to be averaged over all possible ℓ_J, which leads again to a quadratic speed-up in learning [35].

I.3. Limited coherence times

In general, all near-term quantum devices allow for coherent evolution only for a limited time and thus a maximal number n of Grover iterations. For winning probabilities $q = \sin^2 \xi$ with $(2n+1)\xi \le \pi/2$, performing n Grover iterations leads to the highest probability of finding a rewarded action.

Again, we assume that the actual policy of an agent only depends on the number of observed rewards an agent has found. As a consequence, the average time a quantum-enhanced agent limited to n Grover iterations needs to achieve the success probability $Q < \sin^2(\pi/(4n+2))$ is given by

$$\langle T(J,n) \rangle_Q = \sum_{j=1}^{J} \frac{\alpha_0 n + 1}{\sin[(2n+1)\xi_j]^2} \qquad (A.16)$$

with $\sin^2 \xi_j = q_j$ and α_0 determining the number of epochs necessary to create one oracle query $\tilde{U}_E$. For $\alpha_0 = 1$, $n >> 1$ and $(2n+1)\xi_J \ll \pi/2$ we can approximate the learning time for the quantum-enhanced agent via

$$\langle T(J,n) \rangle_Q \approx \sum_{j=1}^{J} \frac{1}{4nq_j} = \frac{\langle T(J) \rangle_C}{4n} \qquad (A.17)$$

where we used $\sin x \approx x$ for $x \ll 1$. In general, it can be shown [35] that the success probability $Q_n = \sin^2(\pi/(4n+2))$ can be reached by a quantum-enhanced agent limited to n Grover iterations in a time

$$\langle T(n) \rangle_Q \le \gamma \frac{\langle T \rangle_C}{n} \qquad (A.18)$$

where γ is a factor depending on the specific setting.
In our case Eq. A.16 can be used to compute the lower bound for the average quantum learning time, with $\alpha_0 = n = 1$. For the classical strategy Eq. A.10 is used.

II.   Experimental details

A continuous wave laser (Coherent Mira HP) is used to pump a single-photon source producing photon pairs in the telecom wavelength band. The laser light has a central wavelength of 789.5 nm and pumps the single photon source at a power of approximately 100 mW. The source is a periodically poled KTiOPO$_4$ non-linear crystal placed in a Sagnac interferometer [38, 39], where the emission of single photons occurs via a type II Spontaneous Parametric Down-Conversion (SPDC) process. The crystal (produced by Raicol) is 30 mm long, set to a temperature of 25 °C, has a poling period of 46.15 μm and is quasi-phase matched for degenerate emission of photons at 1570 nm when pumping with coherent laser light at 785 nm. As the processor is calibrated for a wavelength of 1580 nm, we shift the wavelength of the laser light to 789.75 nm in order to produce one photon at 1580 nm (that is then coupled into the processor) and another one at 1579 nm (the heralding photon).

The processor used for the experiment is a silicon-on-insulator (SOI) type, designed by the Quantum Photonics Laboratory at MIT (Massachusetts Institute of Technology) [40]. Each programmable unit on the device acts as a tunable beam splitter implementing the unitary

$$U_{\theta,\phi} = \begin{pmatrix} e^{i\phi} \sin \frac{\theta}{2} & e^{i\phi} \cos \frac{\theta}{2} \\ \cos \frac{\theta}{2} & -\sin \frac{\theta}{2} \end{pmatrix} \qquad (A.19)$$

where θ and ϕ are the internal and external phases (as in Fig. 2(b)) set via thermo-optical phase shifters controlled by a voltage supply. The achievable precision for phase settings is higher than 250 μrad. The bandwidth of the phase shifters is around 130 kHz. The waveguides, spatially separated from one another by 25.4 μm, are designed to admit one linear polarization only. The high contrast in refractive index between the silicon and silica (the insulator) allows for waveguides with very small bend radius (less than 15 μm), thus enabling a high component density (in our case 88 MZIs) on small areas (in our case 4.9 x 2.4 mm). Given the small dimensions, the in- (and out-)coupling is realized with the help of $Si_3N_4 - SiO_2$ waveguide arrays (produced by Lionix International), that shrink (and enlarge) the 10 μm optical fibers' mode to match the 2 μm mode size of the waveguides in the processor. The total input-output loss is around 7 dB. The processor is stabilized to a temperature of 28 °C and calibrated at 1580 nm for optimal performance. To reduce the black-body radiation emission due to the heating of the phase shifters when voltage is applied, wavelength division multiplexers with a transmission peak centered at 1571 nm and bandwidth of 13 nm are used before the photons are sent to the detection apparatus. In our processor, two external phase shifters in the implemented circuits were not responding to the supplied voltage. This defects were accounted for by deploying an optimization procedure.

The single-photon detectors are multi-element superconducting nanowires (produced by photonSpot) with efficiencies up to 90% in the telecom wavelength band. They have a dark count rate of $\backsim$ 100 c.p.s, low timing jitter (hundreds of ps) and a reset time < 100 ns [41].