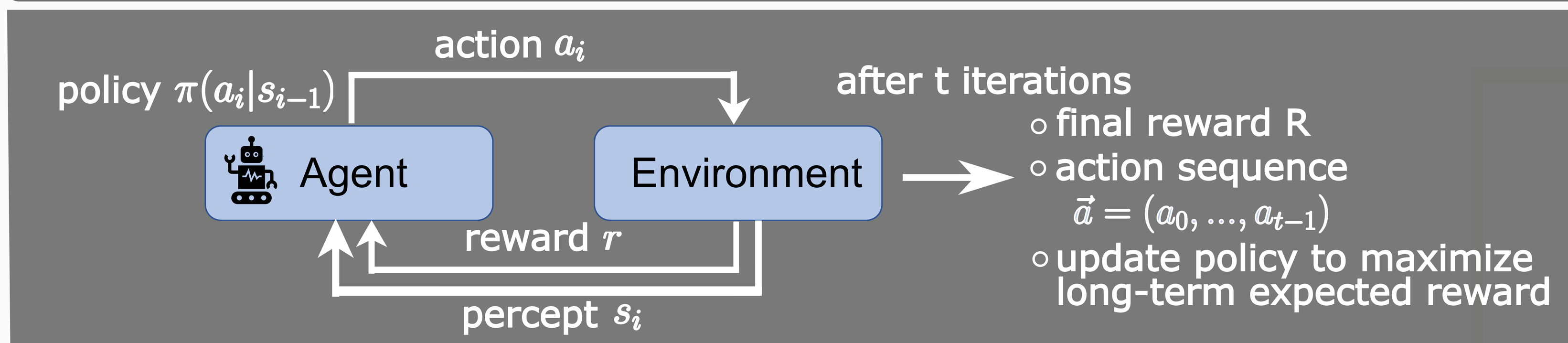
Tic Tac Toe Goes Quantum: Exploring Hybrid Reinforcement Learning On NISQ Devices

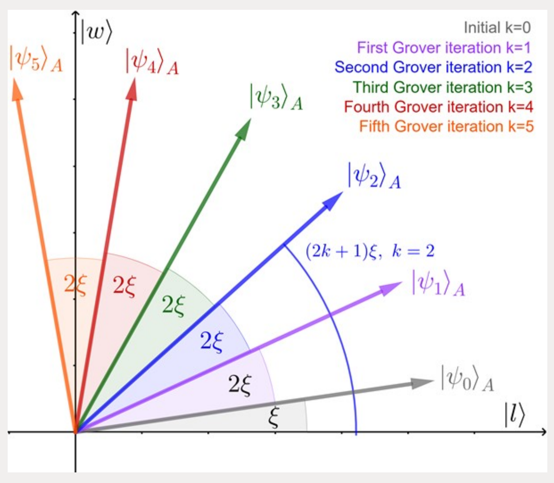
Annette Zapf, Eva Henseler, Sabine Wölk
Institute of Quantum Technologies, German Aerospace Center (DLR), Ulm

Motivation (Reinforcement Learning)



One of the challenges faced by reinforcement learning agents are the significant time resources, i.e. number of interactions with the environment, required for learning. We demonstrate that the learning time can be significantly reduced by implementing a hybrid agent leveraging quantum effects, using the example of a quantum tic tac toe task.

Grover Search



Grover's algorithm [3]:

- Quadratic speed-up possible
- Amplitude amplification of searched states through quantum operations
- Rotation of initial state $|\psi_0\rangle$ closer to the subspace of the winning states $|w\rangle$

The Hybrid Agent [1]

- **Steps of a quantum round:**
  - Quantum epoch:

1. State preparation: Prepare the state $|\psi\rangle_A|-\rangle_R$. *(Agent)*
Superposition of all action states:

$$|\psi\rangle_A = \sum_{\vec{a}} \sqrt{p(\vec{a})}|\vec{a}\rangle_A = \cos(\xi)|l\rangle_A + \sin(\xi)|w\rangle_A = \sqrt{1-\epsilon}|l\rangle_A + \sqrt{\epsilon}|w\rangle_A$$

k x

2. Effect of environment (Oracle): *(Env)*
Apply unitary U_{env} on the prepared state, marking the searched states: $U_{env}|\psi\rangle_A|-\rangle_R = [\sqrt{1-\epsilon}|l\rangle_A - \sqrt{\epsilon}|w\rangle_A]|-\rangle_R$

3. Reflection: Apply a reflection over the initial state:

$$U_R = 2|\psi\rangle\langle\psi|_A - 1_A \qquad (Agent)$$

This leads to an increased amplitude of the winning states [3].

4. Measurement: A measurement of the action register in the computational basis results in a basis action state $|\vec{a}\rangle_A$ associated with the classical action $\vec{a}$. *(Agent)*

- Classical epoch:

5. State preparation: Prepare the state $|\vec{a}\rangle_A|0\rangle_R$. *(Agent)*

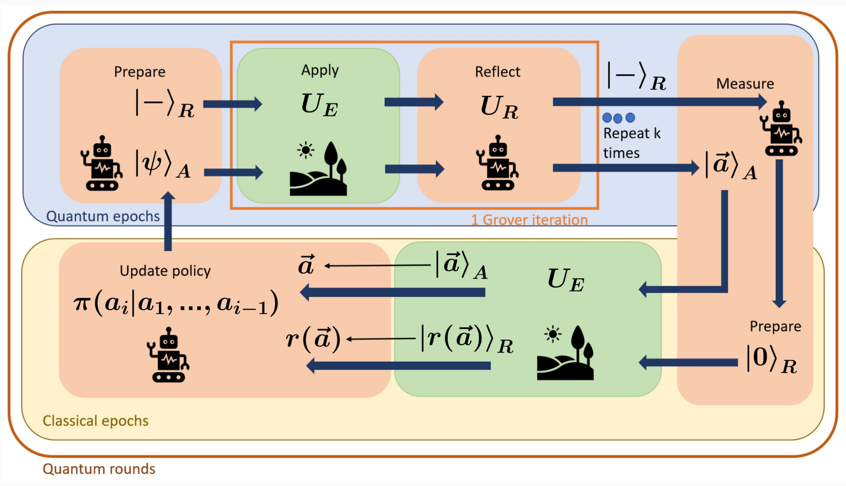
6. Effect of environment (Oracle): Apply the oracle unitary U_{env}:

$$U_{env}|\vec{a}\rangle_A|0\rangle_R = \begin{cases} |\vec{a}\rangle_A|1\rangle_R & \text{if } r(\vec{a}) > 0 \\ |\vec{a}\rangle_A|0\rangle_R & \text{if } r(\vec{a}) \le 0 \end{cases} \qquad (Env)$$

The oracle decides if the chosen action sequence $\vec{a}$ is a rewarded one ($r(\vec{a})>0$).

7. Policy update (learning): *(Agent)*
The basis action states and basis reward states can be associated with the classical action $\vec{a}$ and reward r. The agent updates its policy $\pi(a_i|a_0,\ldots,a_{i-1})$ based on this feedback classically *(e.g. projective simulation method [2])*.

- After some iterations of the Grover search it's beneficial to switch playing only classical epochs

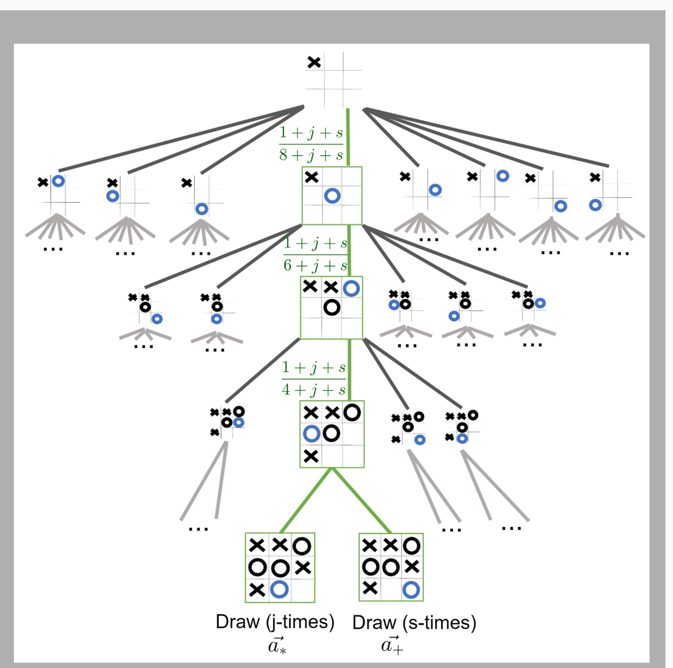


Q-Tic Tac Toe Environment

- **Assumptions:**
  - Deterministic strictly epochal environment (DSE)
  - Effect of environment can be modeled as quantum oracle

Strictly epochal:
- Fixed length of action sequence: $\vec{a} = (a_0, a_1, a_2, a_3)$
- Final reward R: lose R=0, draw R=1

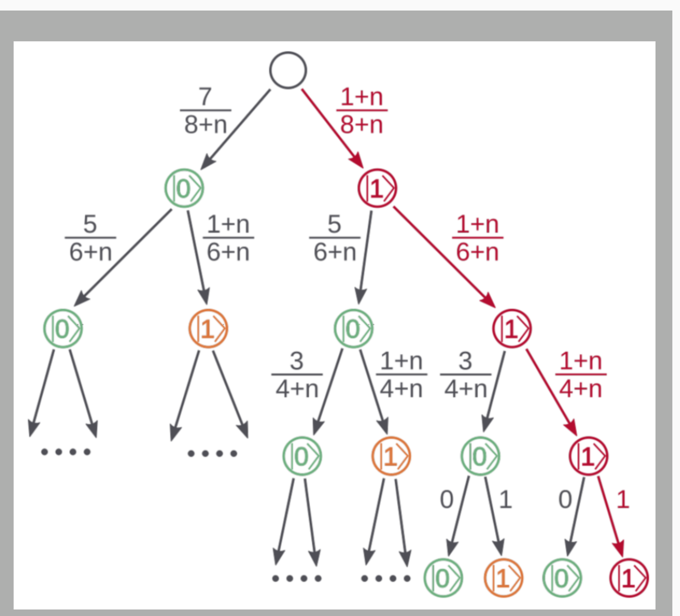
Deterministic:
- The selection of an action a_i always leads to the same state s_i
  → Policy: $\pi(a_i|a_0,\ldots,a_{i-1})$
- Opponent (env) plays optimal strategy (rules see [4])



Decision tree of tic tac toe (DSE)

Simplified version with product states

- The actions are condensed: winning actions $|1\rangle$, losing actions $|0\rangle$
- Instead of increasing only the probability of the rewarded action sequence $\vec{a}$ the agent increases all probabilities to choose $|1\rangle$ over $|0\rangle$
  → $|\psi\rangle_A$ is a product state



Decision tree for product states

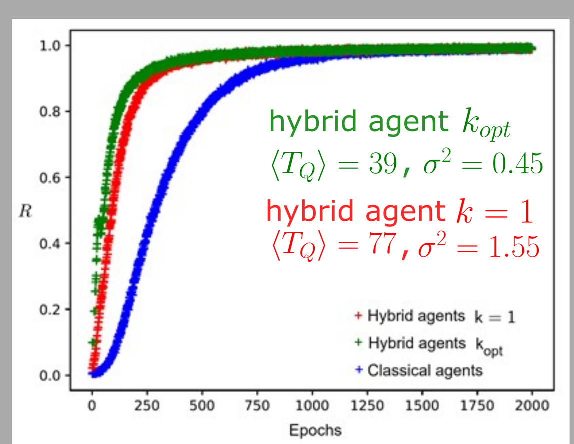
Ressource requirements

- **Standard tic tac toe:** action qubits:
  $5\cdot6\cdot4\cdot2 = 340 < 2^8$ → n=8 qubits (with symmetry)
  1 reward qubit and $2\cdot(n-1)$ ancilla qubits
  → CNOT count for n action qubits:
  $2^n - 2 + 2\cdot(n-1)\cdot6 + 1 + 2\cdot(2^n-2) + 2\cdot(n-1)\cdot6 + 1 \overset{n=8}{=} 932$
  (+ gates from swapping)

- **Simplified tic tac toe + optimized gate decomposition:**
  $8 = 2^3$ → n=3 action qubits ($\vec{a} = (a_0, a_1, a_2)$)
  1 reward qubit and 1 ancilla qubit
  → CNOT count:
  $2\cdot(n-2)\cdot6 + 6 + 2\cdot(2^n-2) + 2\cdot(n-2)\cdot6 + 1 \overset{n=3}{=} 43$
  (+ gates from swapping)

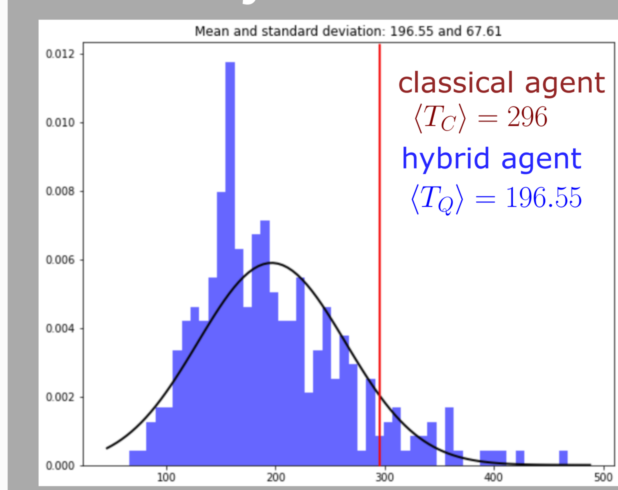
Results

Noiseless



Average reward of 2500 agents per epoch. The reward after k+1 epochs of the hybrid agent is equally distributed over each epoch.

Noisy Simulation



Histogram of the learning times of 300 hybrid agents of a noisy simulation (Qiskit FakeMontreal).
→ 4 action qubits

Noisy Quantum Device

Average learning time of 600 hybrid agents on ibmq_ehningen for simplified task and optimized gate decomposition:

$$\langle T_Q\rangle = 123.77$$

→ 5 qubits
(3 action, 1 reward, 1 ancilla)

noise level (device)	configuration	agents count	$\langle T_Q\rangle$	σ^2	CNOT count
noiseless	9 action qubits	2500	**77**	1.55	(1724)
noiseless	4 action qubits	300	**78.27**	51.86	222
noisy simulation	4 action qubits	300	**196.55**	67.61	222
noisy simulation	3 action qubits, product state	300	**101.84**	44.78	67
noisy QC	3 action qubits, product state	600	**123.77**	67.61	67

Challenges on NISQ Devices

1.) **Noisy** devices: gate errors, measurement errors
2.) **Limited number of qubits**
3.) **Limited connectivity of qubits**

[1] A. Hamann and S. Wölk, "Performance analysis of a hybrid agent for quantum-accessible reinforcement learning", New Journal of Physics,vol. 24, 033044 (2022).
[2] H. Briegel, G. De las Cuevas, "Projective simulation for artificial intelligence", Sci Rep 2, 400 (2012).
[3] L.K. Grover, "A fast quantum mechanical algorithm for database search", Proceedings of the 28th Annual ACM Symposium on the Theory of Computing, New York, 212-219 (1996).
[4] M. Abu Dala et al.,"Tic-tac-toe learning using artificial neural networks", International Journal of Engineeringand Information Systems, vol. 3, pp. 919 (2019).

This work is supported by the QuEST project (DLR-TT, Fraunhofer-IWM, DLR-QT) which is funded by the Baden-Württemberg Ministry of Economic Affairs, Labour and Housing.
It contributes to the QLearning ("Quantenprozessoren für das bestärkende Lernen") project (DLR-QT).

DLR Deutsches Zentrum für Luft- und Raumfahrt
Institut für Quantentechnologien
annette.mueller@dlr.de