Deutsches Forschungszentrum für Künstliche Intelligenz
German Research Center for Artificial Intelligence

Universität Bremen

Vorhaben	QuBER-KI - Quantum Deep Reinforcement Learning für einfache robotische Verhalten
Titel	Abschlussbericht
Förderkennzeichen	50RA2207A & 50RA2207B
Zuwendungsempfänger	Deutsches Forschungszentrum für Künstliche Intelligenz GmbH Trippstadter Straße 122, D-67663 Kaiserslautern Universität Bremen Robert-Hooke-Straße 1, 28359 Bremen
Ausführende Stelle	DFKI GmbH – FB Robotics Innovation Center Universität Bremen – Fachbereich 3, AG Robotik
Projektleiter	Lukas Groß (DFKI), Prof. Dr. Frank Kirchner (Universität Bremen)
Bewilligungszeitraum	01.11.2022 – 31.10.2025
Autoren	Lukas Groß (DFKI) & Dirk Heimann (Universität Bremen)
Erstellungsdatum	13.11.2025

Gefördert durch:

Bundesministerium für Wirtschaft und Technologie

aufgrund eines Beschlusses des Deutschen Bundestages

Das diesem Bericht zugrunde liegende Vorhaben wurde mit Mitteln des Bundesministeriums für Wirtschaft und Technologie unter dem Förderkennzeichen 50RA2207A und 50RA2207B gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autoren.

Eingehende Darstellung:
Schlussbericht zum Projekt
QuBER-KI

Sachbericht zum Verwendungsnachweis

Inhaltsverzeichnis

01 Einleitung

Im Rahme des Projekts QuBER-KI - “Quantum Deep Reinforcement Learning für einfache robotische Verhalten” wurde seitens des Robotics Innovation Center des Deutschen Forschungszentrums für Künstliche Intelligenz (DFKI RIC) das Teilvorhaben QuReRo - “Quantengestützte Regelung und Lernen von Roboterverhalten” durchgeführt. Projektpartner war die Arbeitsgruppe Robotik der Universität Bremen, die das Teilvorhaben BeQuL - “Bestärkende quantenmaschinelle Lernverfahren” durchgeführt hat.

Das Team Quantum Computing des DFKI RIC und die AG Robotik der Universität Bremen haben bereits vor Projektbeginn erfolgreich das Projekt QINROS - “Quantencomputing und quantenmaschinelles Lernen für intelligente und robotische Systeme” (gefördert durch das BMWE, Fkz: 50RA2032 und 50RA2033) durchgeführt. Das Ziel von QINROS war es, basierend auf der Künstlichen Intelligenz (KI) in Quantum Computing Agenda der Arbeitsgruppe Robotik und des DFKI RIC, ein Anwendungsfeld des Quantencomputing in der Robotik für Weltraumanwendungen konzeptionell zu erarbeiten und prototypisch umzusetzen. Mit den erfolgten Arbeiten im Bereich Quantum Machine Learning lieferte das Projekt eine gute Grundlage für die Durchführung des vorliegenden Verbundvorhabens. Weiterhin bestand vor Projektbeginn bei den Verbundpartnern Erfahrung in den Bereichen Quantum Computing und Quantum Machine Learning aufgrund des teilweise parallel laufenden Projektes Q3UP! - „Bedarfsorientierte und niederschwellige Qualifikationsbausteine für Quantencomputing und quantenmaschinelles Lernen“ (gefördert durch das BMBF, Fkz: 13 N 15 993 und 13 N 15 779).

In QuBER-KI wurde seitens des DFKI RIC das wissenschaftliche Arbeitsziel der umfassenden Konzeption, Umsetzung und Evaluation von bestärkenden quantenmaschinellen Lernverfahren für das Verhaltenslernen im Rahmen der Weltraumanwendungen verfolgt. Neben Management und Koordination des Gesamtvorhabens war das DFKI RIC im Rahmen des Teilvorhabens QuReRo für die Erarbeitung zweier wesentlicher Bestandteile von QuBER-KI verantwortlich. Zum einen wurden in QuReRo notwendige umfängliche Simulationsumgebungen für ausgewählte Roboterverhalten der Navigation und der Manipulation konzipiert und umgesetzt, um eine breite Evaluation für diese Anwendungsfelder für quantenmaschinelle Lernverfahren bereitzustellen. Zum anderen wurden in den Simulationen Benchmark-Algorithmen der klassischen wie auch quantenbasierten maschinellen Lernverfahren bereitgestellt. Basierend auf diesen Simulationsumgebungen konnten, die in BeQuL notwendigen Arbeiten durchgeführt werden, um die neuartigen bestärkenden quantenmaschinellen Lernverfahren umzusetzen. Mithilfe des QC- und QML-Labors aus der geförderten Maßnahme QuDA-KI konnten in QuReRo umfangreiche Hyperparameteranalysen der ausgewählten klassischen und quanten-hybriden bestärkenden Lernverfahren erstellt werden. Diese Ergebnisse liefern zum einen Erkenntnisse bzgl. der Reproduzier- und Transferierbarkeit von publizierten Ansätzen. Zum anderen zeigen sie deutlich die Limitierungen für die Anwendung in robotischen Weltraumszenarien auf. Daraus lassen sich Rückschlüsse bezüglich der Skalierbarkeit und damit die mögliche Anwendung in komplexeren Lernszenarien ziehen. Darüber hinaus wurden Methoden des Quantum Optimal Controls, insbesondere für die Optimierung von Quantenoperationen, implementiert und verglichen.

Die Universität Bremen verfolgte im Rahmen des Teilvorhabens BeQuL das Ziel der Erarbeitung einer Auswahl an geeigneten Varianten des bestärkenden quantenmaschinellen Lernens auf bestehenden QC-Algorithmen. Weiterhin wurden diese Verfahren evaluiert und eine Teilmenge dieser Varianten gebildet, welche in BeQuL mit neuartigen Ideen weiterentwickelt wurden, um die ausgewählten robotischen Verhalten der Manipulation und der Navigation umzusetzen.

Die Verbundpartner haben die im Projektplan vorgesehenen Ziele unter Einhaltung des Zeit- und

Kostenplans erreicht. Details entnehmen Sie dem zahlenmäßigen Verwendungsnachweis.

02 Notwendigkeit der Projektarbeiten

Das Deutsche Forschungszentrum für Künstliche Intelligenz (DFKI) am Standort Bremen erarbeitet seit mehreren Jahren zusammen mit der Universität Bremen, Arbeitsgruppe Robotik, prototypische Anwendungen im Bereich des quantenmaschinellen, bestärkenden Lernens auf Noisy Intermediate-Scale Quantum (NISQ) Technologien.

Auch wenn quantenmaschinelles, bestärkendes Lernen noch ein junges Forschungsfeld ist, sind potenzielle Anwendungsbereiche bereits breit gefächert. Klassisches bestärkendes Lernen (RL) und tiefes bestärkendes Lernen (DRL) als Technologie sind bereits seit einigen Jahren in der Industrie im Einsatz und werden für vielfältige Aufgaben erfolgreich und umfänglich eingesetzt. Die Erfolge in den letzten Jahren sind als einzigartig und weit überdurchschnittlich zu bewerten. Eine Weiterentwicklung von DRL-Algorithmen durch den Einsatz von den jetzt und in Zukunft zur Verfügung stehenden Quantencomputern verspricht daher einen potenziellen Nutzen in vielen Anwendungsfeldern – auch direkt bei den potenziellen industriellen Anwendern. Die Quantencomputer der aktuellen Generation sind in ihrer Rechenkapazität und Benutzbarkeit noch stark eingeschränkt. So erfordern viele theoretisch beschriebene Algorithmen mehr Qubits, längere Kohärenzzeiten und geringere Fehlerraten auf Qubitebene, als zum aktuellen Zeitpunkt auf existierender Quantenhardware verfügbar sind. Dennoch lassen sich diese Quantencomputer bereits nutzen und es ist Teil der aktuellen Forschung, inwieweit bereits mit diesen Quantencomputern eine Verbesserung im Vergleich zu bestehenden klassischen Algorithmen möglich ist. Aufgrund der geringen Anzahl von Qubits beschränken sich viele der Anwendungen, die momentan erforscht werden, auf simple, niedrig-dimensionale Problemstellungen. Für den datengetriebenen Ansatz mit quantengestützten Reinforcement-Learning-Verfahren gibt es bereits einige Vorarbeiten, die zeigen, dass hybride Ansätze geeignet sind, um einfache Verhaltenslernprobleme ähnlich gut wie klassische Methoden zu lösen, und gleichzeitig die Lösung in deutlich kompakteren Modellen repräsentieren. Die Verhalten, die erfolgreich gelernt werden konnten, beschränken sich jedoch bislang auf sehr simple Benchmark-Anwendungen.

In diesem Stand der Forschung sind drei Arbeitsausrichtungen von besonderer Wichtigkeit. Zum einen ist es sehr relevant die Potenziale und Limitierungen der quantenhybriden bestärkenden Lernverfahren zu untersuchen. Dafür ist es notwendig bestehende Quantum Variational Circuit (QVC) Ansätze in einer Vielzahl von komplexen Umgebungen umzusetzen und zu evaluieren. Zum anderen ist es essentiell die notwendigen (klassischen) neuronalen Netze in weiteren erfolgreichen, klassischen und komplexeren Algorithmen des tiefen-bestärkenden Lernens durch QVCs auszutauschen oder zu erweitern. Dies ermöglicht den Vergleich von Quantenergänzungen mit den leistungsstärksten klassischen Algorithmen. Die Analyse auf simpleren Benchmark–Umgebungen und komplexeren Roboterumgebungen ermöglicht der Vergleich von verschiedenen quanten-erweiterte bestärkende Lernverfahren in unterschiedlichen Simulationsumgebungen und liefert Erkenntnisse über die Potenziale und Limitierungen der Methoden. Um die Transferierbarkeit zu analysieren ist es wichtig Umgebungen mit möglichst kleinen Anpassungen am Algorithmus zu lösen und den Einfluss von Hyperparameteränderungen zu untersuchen. Dadurch entsteht ein neues, umfassendes Bild über die Verwendbarkeit von quanten-erweiterten Algorithmen in unterschiedlichen Szenarien. Außerdem ist es von großer Bedeutung, Möglichkeiten zu analysieren und zu konzipieren, die über die Anwendung von QVC-Ansätzen hinausgehen. Diese Fragestellung beinhaltet den höchsten Innovationsgrad, weil quantenmechanische Eigenschaften explizit in bestärkenden Lernverfahren ausgenutzt werden sollen. Darüber hinaus bieten die Methoden des Quantum Optimal Controls ein großes Anwendungsfeld, dessen Untersuchung wertvolle Erkenntnisse liefert.

All diese Arbeiten wurden im Rahmen des Projekts QuBER-KI in den einzelnen Arbeitspaketen

angegangen.

03 Inhaltlicher Ablauf der Arbeitspakete

Das Vorhaben QuBER-KI gliederte sich in fünf Arbeitspakete (s. auch Abbildung 1):

- AP1000: Management und Koordination
- AP2000: Auswahl geeigneter Roboterverhalten für Navigation und Manipulation
- AP3000: Konzeption und Umsetzung von Simulationsumgebungen
- AP4000: Neuartige Optimierungsansätze
- AP5000: Quantenbasiert-bestärkende Lernverfahren für robotische Verhalten

Als Konsortialführer war das DFKI RIC mit dem Arbeitspaket 1000 beauftragt. Weiterhin war das DFKI RIC federführend in den Arbeitspaketen 2000, 3000 und 4000, während das Arbeitspakete 5000 federführend durch die AG Robotik der Universität Bremen bearbeitet wurde.

Im Projekt gab es vier Meilensteine, die allesamt erfolgreich erreicht wurden:

- MS1: Projektbeginn
- MS2: Auswahl und Definition der Roboterverhalten
- MS3: Fertigstellung unterschiedlicher Benchmark-Simulationsumgebungen für bestärkende Lernverfahren
- MS4: Umsetzung quantenmaschineller Lernverfahren für robotische Verhalten

Die Arbeitspakete und Meilensteine sind auch in der folgenden Abbildung dargestellt.

Arbeitsplan				Jahr 1				Jahr 2				Jahr 3			
Arbeitspaket	**Beginn**	**Ende**	**PM**	**Q1**	**Q2**	**Q3**	**Q4**	**Q1**	**Q2**	**Q3**	**Q4**	**Q1**	**Q2**	**Q3**	**Q4**
AP 1000: Management und Koordination	1	36	9												
AP 2000: Auswahl geeignter Roboterveralten für Navigation und Manipulation	1	12	6												
AP 3000: Konzeption und Umsetzung von Simulationsumgebungen	13	36	24												
AP 4000: Neuartige Optimierungsansätze	13	36	36												
AP 5000: Quantenbasierte-bestärkende Lernverfahren für robotische Verhalten	1	36	33												
Meilenstein															
MS 1: Projektbeginn	1														
MS 2: Auswahl und Definition der Roboterverhalten	13														
MS 3: Fertigstellung unterschiedlicher Benchmark-Simulationsumgebungen	25														
MS 4: Umsetzung quantenmaschineller Lernverfahren für robotische Verhalten	36														
AP-Verantwortlicher: DFKI															
AP-Verantwortlicher: Universität Bremen															

Abbildung 1: Darstellung des Plans der Arbeitspakete und Meilensteine.

Im Folgenden werden Ablauf, Inhalte und Ergebnisse der einzelnen Arbeitspakete erläutert.

AP1000: Management und Kommunikation

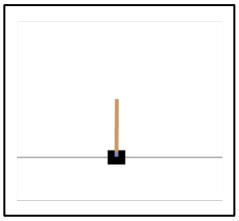
Das DFKI RIC war als Konsortialführer verantwortlich für die Organisation der internen Projektmeetings sowie der regelmäßigen Termine mit dem Fördermittelgeber. Dies beinhaltete insbesondere Koordination und Synchronisation der Arbeitspakete, Berichterstattung an Auftraggeber und interne Stellen sowie die Mittelverwaltung. Die Universität Bremen führte ebenfalls regelmäßige Projektmeetings durch. Auch sie kümmerte sich um Koordination und Synchronisation der Arbeitspakete, Berichterstattung an Auftraggeber und interne Stellen sowie die Mittelverwaltung.

Das Vorhaben wurde seitens des DFKI RIC als Konsortialführer erfolgreich gestartet. Eine Webseite wurde erstellt und der Konsortialvertrag definiert und abgeschlossen. Ein Zwischenmeeting mit Projektträger und Verbundpartnern zum Informationsaustausch fand am 15.06.2023 statt. Am 06.12.2023 wurde der bisherige Stand der Arbeiten zum zweiten Meilenstein in einem Meeting dem Projektträger vorgestellt. Der Meilenstein wurde als erreicht bestätigt. Ebenso wurde am 19.12.2024 in einem Meeting mit dem Projektträger der bisherige Stand der Arbeiten zum dritten Meilenstein vorgestellt. Der Meilenstein wurde als erreicht bestätigt. Abschließend wurden am 25.09.25 in einem Termin mit Projektträger und Verbundpartnern die Arbeiten zum vierten Meilenstein vorgestellt. Der Meilenstein wurde als erreicht bestätigt.

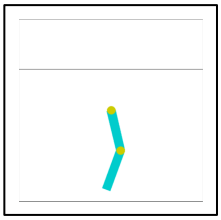
Die öffentliche Kommunikation im Rahmen des AP1000 fand u.a. im Rahmen der jährlichen Tage der offenen Tür des DFKI RIC statt. Hier wurden mehrere Poster präsentiert, die einen Überblick über das Thema Quantum Computing gaben sowie einen Einblick in die aktuelle Forschungsarbeit am DFKI RIC und der Universität Bremen lieferten.

AP2000: Auswahl geeigneter Roboterverhalten für Navigation und Manipulation

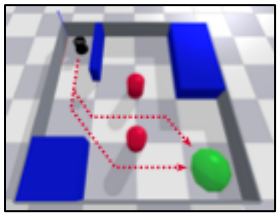
In diesem Arbeitspaket wurde eine ausführliche Recherche, der am DFKI-RIC sowie in einschlägiger Literatur relevanter Roboterverhalten durchgeführt. Das Ziel dieser Recherche war die Festlegung auf eine begrenzte Anzahl geeigneter Roboterverhalten, für die Evaluation der in diesem Projekt betrachteten bestärkenden Lernverfahren. Hierbei wurde besonderes Augenmerk auf drei Merkmale gelegt: Die Komplexität der ausgewählten Verhalten sollte variieren, um Aussagen zur Skalierbarkeit der getesteten Lernverfahren zu ermöglichen. Die Verhalten sollten bereits als Simulation vorliegen oder leicht zu simulieren sein, damit Testläufe oft genug wiederholt werden können, um statistisch signifikante Aussagen zu treffen. Zuletzt wurde auf die Relevanz für die Robotik wert gelegt. Hierbei wurde auch darauf geachtet, verschiedene Bereiche der Robotik abzudecken. Das Ergebnis war die Festlegung auf drei robotische Verhalten (siehe Tabelle 1), welche die in der Robotik relevanten Problemstellungen des Umgangs mit nicht-linearer Dynamik, der Navigation sowie der Manipulation abdecken. Außerdem wurde sichergestellt, dass sich die Verhalten in Ihrer Komplexität unterscheiden. Darüber hinaus wurde weitere besonders simples Verhalten ausgewählt, welche zur ersten Validierung der implementierten Lernverfahren genutzt werden können. Alle Verhalten sind entweder offen zugänglich oder wurden am DFKI entwickelt. Dadurch ist eine solide Dokumentierung und Implementierung der Verhalten in den Umgebungen gewährleistet.



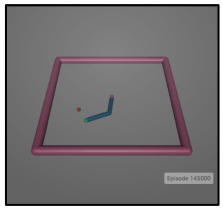
Cartpole



Double Pendulum



Turtlebot



Reacher

Tabelle 1: Auswahl der robotischen Verhalten

Verhalten	Problemstellung	Komplexität	Verfügbarkeit
Cartpole	Validierung	gering	open-source
Pendulum	non-lineare Dynamik	gering bis mittel	open-source
Double-Pendulum	non-lineare Dynamik	mittel	open-source
Turtlebot	Navigation	Hoch (non-lokal)	DFKI
Reacher	Manipulation und Dynamik	Hoch (kombination)	open-source

Zusätzlich wurde die Kompatibilität der gewählten Verhalten mit drei der gängigen klassischen bestärkenden Lernverfahren untersucht, namentlich Deep Q-Networks (DQN), Soft Actor Critic (SAC) und Proximal Policy Optimization (PPO). Hierbei stellte sich heraus, dass vor allem die Beschaffenheit des Aktionsraums ein relevanter Aspekt ist. Während PPO prinzipiell in der Lage ist, sowohl Verhalten mit diskretem als auch mit kontinuierlichem Aktionsraum zu erlernen, sind DQN

und SAC auf Verhalten mit diskretem, respektive kontinuierlichem Aktionsraum beschränkt (siehe Tabelle 2).

Tabelle 2: Kompatibilität mit klassischen bestärkenden Lernverfahren

Verhalten	Aktionsraum	DQN	SAC	PPO
Cartpole	diskret	✓	✗	✓
Pendulum	kontinuierlich	✗	✓	✓
Double-Pendulum	kontinuierlich	✗	✓	✓
Turtlebot	diskret	✓	✗	✓
Reacher	kontinuierlich	✗	✓	✓

Mit der Festlegung auf eine Auswahl geeigneter robotischer Verhalten für die Nutzung mit bestärkenden Lernverfahren wurde das Arbeitspaket 2000 erfolgreich abgeschlossen. Gleichermaßen wurde hiermit der Meilenstein 2 erfolgreich erreicht.

AP3000: Konzeption und Umsetzung von Simulationsumgebungen

Im Rahmen dieses Arbeitspakets wurden, die in AP 2000 ausgewählten Verhalten als Simulationsumgebung zur umfassenden Evaluation verschiedener bestärkender Lernverfahren bereitgestellt. Konkret bedeutete dies die Anpassung vorhandener Implementierungen der Verhalten an eine festgelegte Schnittstelle zu den anzuwendenden Lernverfahren. Die Wahl fiel hierbei zunächst auf die Konventionen der OpenAI-Gym bzw. Gymnasium-API, da diese weit verbreitet ist und bereits hierzu passende Implementierungen der oben genannten Lernverfahren DQN, SAC und PPO vorliegen.

Darüber hinaus wurde ein Framework implementiert, welches dazu dient, ausführliche numerische Experimente unter Nutzung verschiedener Umgebungen und Lernverfahren strukturiert durchführen und analysieren zu können.

Erste Implementierungen der gewählten Lernverfahren (DQN, SAC, PPO) sowie die entsprechenden auf parametrisierten Quantenschaltkreisen basierenden Varianten wurden im Laufe des zweiten Jahres bereitgestellt. Es wurden erste Schritte gemacht, um diese Implementierungen mit den Simulationsumgebungen zu integrieren und das resultierende Framework für das Training auf Hochleistungshardware zu optimieren.

Es konnte die Funktion der Algorithmen Deep Q-Network (DQN), Proximal Policy Optimization (PPO) und Soft Actor-Critic (SAC) in verschiedenen Umgebungen, unter anderem Pendulum, Cartpole, Reacher, und Double-Pendulum getestet werden. Erste Ergebnisse zeigten, dass RL-Algorithmen, die auf parametrisierten Quantenschaltkreisen (PQC) basieren, in Einzelfällen vergleichbare Leistungen, z.B. in der Konvergenzgeschwindigkeit und Stabilität der Lern-Performance, wie ihre klassischen RL-Algorithmen erbringen können.

Im Rahmen der Testläufe wurden zwei zentrale Faktoren identifiziert, die die Weiterentwicklung in diesem Bereich positiv beeinflussen können. So stellte sich heraus, dass die verwendete Kombination aus TensorFlow Agents und TensorFlow Quantum keine optimale Nutzung der vorhandenen High Performance Hardware gewährleisten konnte. Zweitens stellte sich die Frage der Vergleichbarkeit zweier Reinforcement Learning Ansätze im Hinblick ihrer Sensitivität auf Hyperparameter. So konnten Beispiele beobachtet werden bei denen optimale spezifische Hyperparameter in beiden Ansätzen zu optimaler Performance führten, die klassischen Ansätze aber deutlich robuster auf Änderungen in den Hyperparametern reagierten als das Quantum Äquivalent.

Als Lösungsansatz zur Performancesteigerung wurde zum Ende des zweiten Jahres damit begonnen ein Framework basierend auf JAX zu entwickeln, welches durch weitreichende Parallelisierung eine deutlich verbesserte Ausnutzung des zur Verfügung stehenden CPU- und GPU-RAMs gewährleistet. Hierdurch konnten Laufzeitverbesserungen von drei bis vier Größenordnungen erzielt werden.

Die Performancesteigerung ermöglichte eine Erste umfassendere Hyperparametersuche für verschiedene Algorithmen durchzuführen. Erste Erkenntnisse bestätigten die Vermutung, dass QRL-Ansätze empfindlicher auf Änderung der Hyperparameter reagieren. Auf Grundlage dieser Experimente wurde das Paper „A Case Study on The Effects of Hyperparameters in Quantum Deep Reinforcement Learning“ erstellt und bei der IEEE Quantum Week 2024 eingereicht. Darüber hinaus wurde die Notwendigkeit wohldefinierter Metriken der Hyperparametersensitivität für die Vergleichbarkeit unterschiedlicher RL-Ansätze festgestellt. Aufgrund fehlender etablierter Arbeiten hierzu wurde mit der Recherche und Entwicklung entsprechender Metriken begonnen.

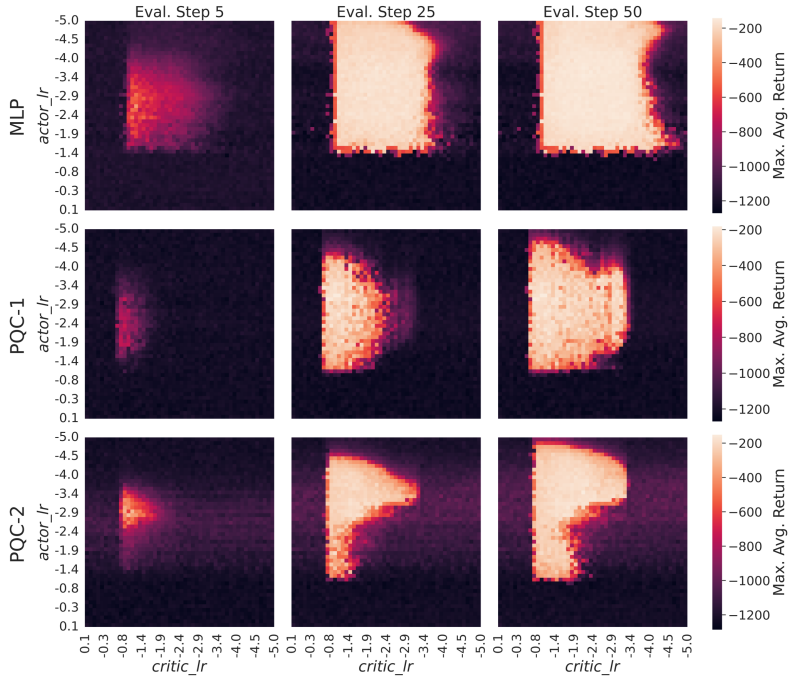


Abbildung 2: Maximum Average Return in Abhängigkeit der Actor und Critic Learning-Rate

Abbildung 2 stellt eine der Haupterkenntnisse aus oben genannter Arbeit dar. Gezeigt ist die Maximum Average Return, welche einen Haupt-Performance-Indikator darstellt, in Abhängigkeit zur Actor und Critic Learning Rate. Trainiert wurde PPO auf der Pendulum Umgebung mittels eines klassischen neuronalen Netzes und zwei PQC Varianten. Die hellen Bereiche stellen eine hohe Performance dar, während dunkle Bereiche ein Versagen des Algorithmus darstellen. Es ist klar ersichtlich, dass die beiden PQC Varianten für ausgewählte Hyperparameter gleiche Performance erreichen, der Bereich der optimalen Performance des klassischen NNs allerdings deutlich größer ist.

Darüber hinaus wurde die Arbeit „Quantum Deep Reinforcement Learning for Robot Navigation Tasks“ im Berichtszeitraum weiter überarbeitet und das Reviewer Feedback integriert. In dieser Arbeit werden unterschiedliche Konfigurationen der PQC hybriden double deep Q network Reinforcement Learning Methode bei dem Lösen meherer Navigationsaufgaben für einen einfachen Roboter (TurtleBot) in simulierten Umgebungen mit zunehmender Komplexität miteinander verglichen. Die Arbeit konnte abschließend im Journal IEEE Access veröffentlicht werden (https://doi.org/10.1109/ACCESS.2024.3417808).

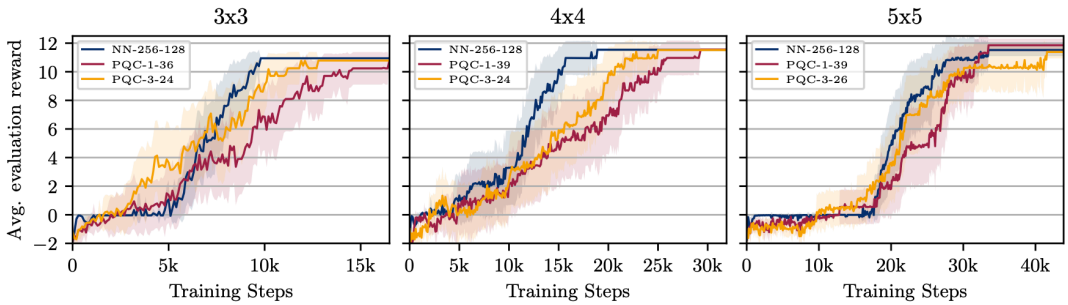


Abbildung 3: Lernerfolg von PQC-DQN in der TurtleBot Umgebung in 3 verschiedenen Größen

Abbildung 3 zeigt ein Hauptergebnis aus dieser Arbeit. Verglichen wird der Reward zweier PQC-DQN Algorithmen mit einem auf klassischen NN basierenden DQN in der TurtleBot Umgebung. Exemplarisch ist hier wieder zu sehen, dass PQCs ähnliche Leistungen wie NNs aufzeigen können.

Die obigen Erkenntnisse bildeten die Basis für die Entscheidung, den Fokus im letzten Jahr auf die Entwicklung eines hochoptimierten Frameworks zur parallelisierten Hyperparametersuche auf High-Performance Computern zu legen. In Folge dessen wurde das Framework hyperlax entwickelt und die zu betrachtenden Algorithmen und Umgebungen integriert.

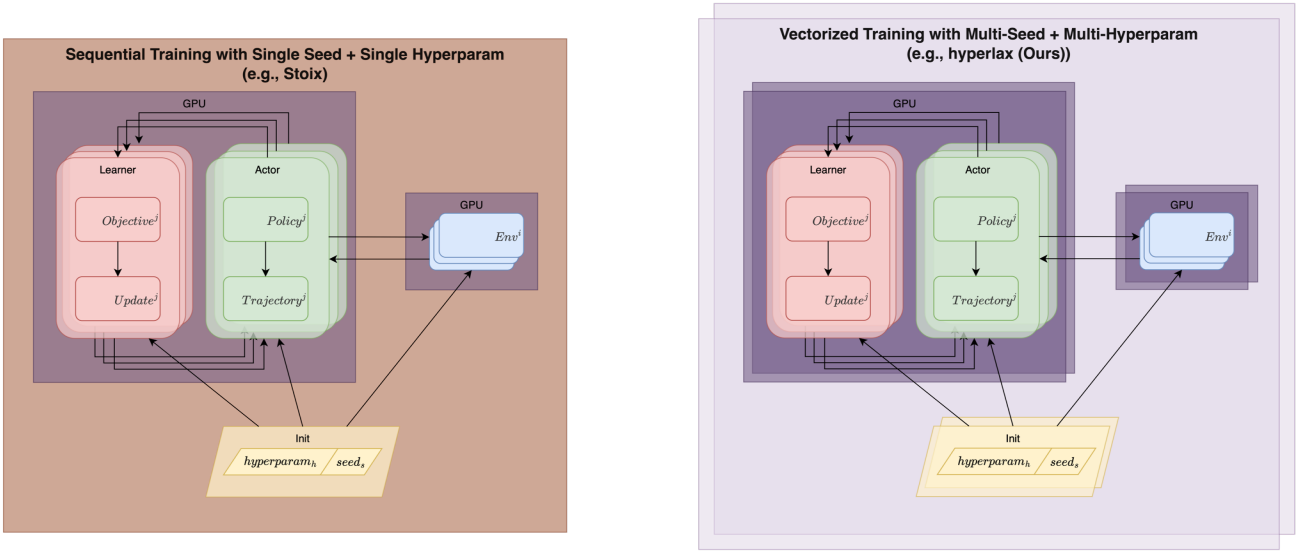


Abbildung 4: Konzeptidee für hyperlax

Abbildung 4 zeigt konzeptionell die Idee von hyperlax. hyperlax erweitert die Ideen von PureJaxRL und Stoix, zwei früheren JAX-RL-Systemen. PureJaxRL demonstrierte, wie man die gesamte Trainingsschleife mit jax.jit kompiliert und lax.scan verwendet, um alle Umgebungsschritte auf dem Beschleuniger zu halten und CPU-GPU-Engpässe zu beseitigen. Stoix baute darauf auf, indem es modulare Abstraktionen für das Training mit mehreren Geräten einführte und pmap für synchronisierte Gradientenaktualisierungen über GPUs hinweg verwendete. Hyperlax führt diese Prinzipien weiter, indem es mehrere Hyperparameterkonfigurationen parallel innerhalb einer einzigen kompilierten Berechnung ausführt und so von der Optimierung einzelner Experimente zu vollständig gebatchten Experimenten übergeht. Es parallelisiert ganze Hyperparametersätze, behandelt unterschiedliche Rollout-Längen und Umgebungsanzahlen durch Padding, Maskieren oder Gruppieren und verteilt die Daten gleichmäßig auf die Hardware. Sein Trainer, der Phaser, verwaltet Experimente, die zu unterschiedlichen Zeitpunkten enden, indem er nur die aktiven neu kompiliert und so die Ressourcennutzung effizient hält. Dieses Framework bewahrt die statische Form und das funktionale Design von JAX und ermöglicht gleichzeitig groß angelegtes, paralleles Training und Benchmarking für klassische und quantenbasierte Reinforcement-Learning-Modelle. Das Framework wurde als open-soruce software zur verfügung gestellt (https://github.com/dfki-ric-quantum/hyperlax-quantum).

Mittels hyperlax war es möglich umfassende Experimente durchzuführen, die im Rahmen von AP5000 analysiert wurden.

AP4000: Neuartige Optimierungsansätze

Das Ziel von AP4000 ist es neuartige Ansätze der mehrdimensionalen Optimierung zu erforschen, welche zu einer verbesserten Regelung führen.

Dazu wurde mit der Literaturrecherche zu Zusammenhängen zwischen Verfahren der klassischen optimalen Regelung robotischer System und den Möglichkeiten des Quantencomputing begonnen. Eine erste Erkenntnis hierbei ist die Tatsache, dass nicht nur die Anwendung von Quantenalgorithmen auf Probleme der optimalen Steuerung, sondern auch der umgekehrte Fall denkbar ist. Also das Nutzen von Kenntnissen aus der optimalen Steuerung zur Optimierung der Interaktion mit den Qubits eines Quantencomputers.

Eine verbreitete technische Umsetzung von supraleitenden Quantencomputern sind Transmon Qubits deren quantenmechanische Zustände mit Mikrowellenpulsen geändert werden können, um Operationen (Gates) auszuführen. Das quantum optimal control Feld forscht unter anderem daran, wie die Amplituden dieser Mikrowellenpulse für das Verändern von dynamischen quantenmechanischen Zuständen genutzt werden können. Eine ausführliche Aufbereitung relevanter Arbeiten bis 2016 (https://dx.doi.org/10.1088/0953-8984/28/21/213001) und bis 2022 (https://doi.org/10.1140/epjqt/s40507-022-00138-x) befindet sich den beiden Arbeiten von Koch et. al. Ein wichtiger Unterbereich ist das Optimieren der Amplitude des Mikrowellenpulses für präzise Basisoperationen auf dem verwendeten supraleitenden Bauteil. Die Methoden lassen sich grob in drei Kategorien unterteilen: modellbasiert, modelfrei, und hybrid. Hybride Methoden zeichnen sich dadurch aus, dass die initiale Amplitude anhand eines Models ermittelt wird, aber mit direkten Hardwaredaten an das reale Bauteil angepasst wird. Methoden aus diesem Bereich, die mit analytischen Näherungen eine einfache Pulseform finden, die mit Hardware Daten verbessert wird, sind besonders erfolgreich gewesen. Die DRAG Methode (https://link.aps.org/doi/10.1103/PhysRevLett.103.110501) ermöglicht das Kalibrieren von 1- Qubit Operationen wie der X-Operation. Die echoed cross-resonance Methode ermöglicht das Kalibrieren von XZ- zwei-Qubit Operation (https://link.aps.org/doi/10.1103/PhysRevA.93.060302) auf fixed-frequency Transmons. Mit der PALEA Methode (https://arxiv.org/abs/2508.16437) können Controlled-Z Operationen für Transmons mit veränderbaren Frequenzen und Coupler-Frequenzen implementiert werden. Diese hybriden Methoden haben den Nachteil, dass sie auf analytischen Lösungen basieren, die nur mithilfe von Annäherungen erzielt werden können. Modelfreie Verfahren wie überwachte (https://link.aps.org/doi/10.1103/PhysRevApplied.21.044012) und bestärkende Lernverfahren (https://link.aps.org/doi/10.1103/PRXQuantum.2.040324) sind sehr Daten intensiv, weshalb Kalibrierungen an echten Systemen lange dauern können.
Modell-basierte Methoden wie GRAPE (https://link.aps.org/doi/10.1103/PhysRevA.63.032308) können auch mit Echtdaten fein-kalibriert werden (https://link.aps.org/doi/10.1103/PhysRevA.97.042122). Zusammen mit fortschreitender Technik der klassischen Technologie zur Erzeugung von Mikrowellenpulsen, zeigt dieser Fakt, dass modelbasierte Verfahren das Potenzial besitzen, zukünftig in der Optimierung von Pulsen für Operationen auf Qubits eingesetzt zu werden.

Wir haben die iterative linear quadratic regulator (iLQR) Methode, die in der Trajektorienoptimierung von komplexer robotischer Systeme erfolgreich war, für die Optimierung von geglätteten Pulseamplituden für Quantenoperationen angepasst.
Es wurde eine Simulationsumgebung erstellt, in der der Quantenzustand eines 1- und 2-Transmonsysteme mit Amplituden eines Mikrowellenpulses manipuliert werden kann.
Diese Mikrowellenpulse werden durch iLQR ermittelt. Die erhaltenen Amplituden konnten über die frei zugängliche Qiskit Pulse API an echte IBM-Quantensysteme geschickt werden. Die Qualität der Mikrowellenpulse kann auf echten Quantensystemen mit unterschiedlichen Verfahren evaluiert werden. Zwei weit verbreitete Methoden sind die Quantum Process Tomography und das Randomized Benchmarking. Quantum Process Tomography ermöglicht es die Auswirkung eines Pulses als Channel über Messungen zu charakterisieren, indem bestimmte Anfangszustände und Messbasen gewählt werden. Dieses Verfahren wurde für 1- und 2-Qubit Operationen mit Qiskit

implementiert und steht nun zur Charakterisierung zur Verfügung. Das Randomized Benchmark Verfahren eignet sich besonders, um die Qualität von Basisoperationen zu ermitteln. Dazu werden Schaltkreise aus einer unterschiedlichen Anzahl hintereinander ausgeführten beliebigen Clifford Gates erstellt. Abschließend wird das Clifford Gate bestimmt, dass zu der vorherigen Kette inverse ist und dem Schaltkreis hinzugefügt. So wäre die gesamte Kette aus Operationen, wenn die Operationen keinen Fehlern unterlägen, gleich der Identität. Die Abweichung zur Identität mit ansteigender Länge der Kette aus Operationen ist deshalb eine Gütemessung für die Pulse der Operationen. Dieses Verfahren wurde für 1-Qubit Operationen in Qiskit implementiert, sodass eigene Pulse für 1-Qubit Operationen untersucht werden können.

Wir analysieren die Einsatzmöglichkeiten der iterative Linear Quadratic Regulator Methode, die Echtzeitsteuerung von komplexeren Robotern ermöglicht, für die Optimierung der Mikrowellenpulse der Basisoperationen von supraleitenden fixed-frequency transmon Systemen. Die Pulse für 1-Qubit Operationen zeigen eine Fidelity von über 99.8 % in den Experimenten von Quantum Process Tomography und Randomized Benchmarking auf. Dies liegt hinter den state-of-the-art-Lösungen, die in unserem Experiment eine Güte von über 99.9 % erzielen. Die Randomized-Benchmark Ergebnisse von dem echten IBM System sind in Abbildung 5 dargestellt. Die Abweichung lässt sich dadurch erklären, dass wir ein vereinfachtes Modell ohne Berücksichtigung von Fehlern des realen Systems verwenden. Die Ergebnisse zu 1-Qubit Operationen wurden als Poster auf dem Workshop „Quantum Optimal Control - From Mathematical Foundations to Quantum Technologies“ in Berlin vorgestellt.

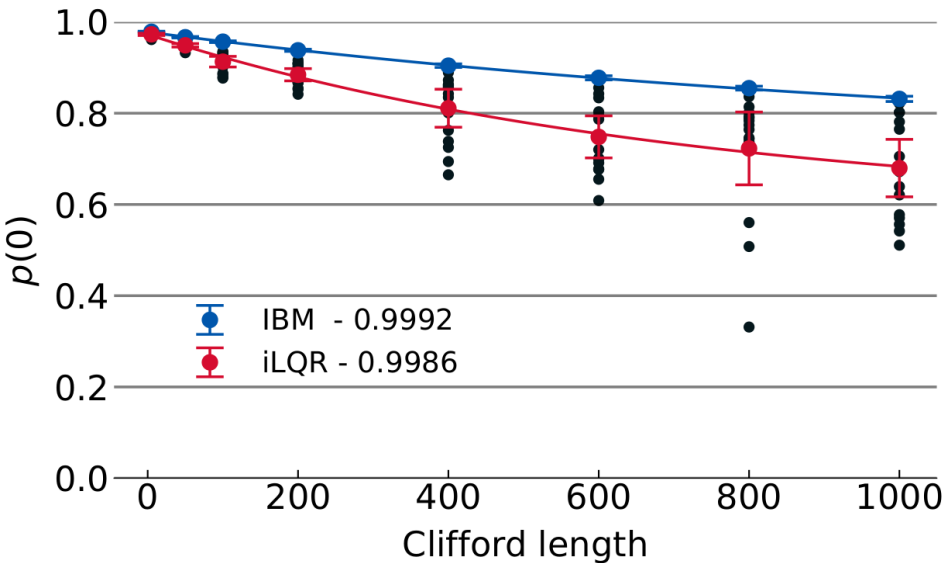


Abbildung 5: Randomized Benchmark Experiment für 1-qubit Operation erhalten durch unsere iLQR Methode im Vergleich zu der Standard IBM Methode.

Die Quantum Process Tomography Resultate für unsere Pulse für 2-Qubit Operationen weichen noch deutlich von dem erwarteten Ergebnis ab, was darauf hindeutet, dass das Modell das reale System noch nicht ausreichend gut beschreibt. In der Zwischenzeit hat IBM ihren Pulszugang eingestellt, sodass wir keine weiteren Experimente auf deren Systemen mehr durchführen konnten und uns auf die Simulation beschränkt haben.

Wir haben die fixed-frequency Transmon Modelle für ein- und zwei-Qubit Operationen mit zwei und drei Level, implementiert in Simulation implementiert. In (https://doi.ieeecomputersociety.org/10.1109/QCE57702.2023.00144) wurden die isomorphe Repräsentation von komplexwertigen Matrizen, die Inkludierung der Ableitung der Pulseamplituden als Optimierungsvariablen und die Pade Approximation verwendet, um direct collocation für die Optimierung für Pulse für Operationen zu verwenden.
Wir haben diese Methoden verwendet, um das Optimierungsproblem so umschreiben zu können,

dass es mit iLQR gelöst werden kann. In dem einfachsten Fall, eine X Operation für 1 Qubit mit 2 Level, erreichen wir mit der Methode eine Güte von 10^{-9} in der idealen Simualtion und einen glatten Verlauf der u^X Pulsamplitude. Dieses steht im Einklang mit der analytischen Berechnung, die für diesen einfachsten Fall noch möglich ist.
In dem 1 Qubit mit 3 Level Fall wird Leakage, also das Erreichen des unerwünschten zweiten angeregten Zustandes, ermöglicht. Die Ziel-X-Operation wird um ein drittes Level erweitert. Die Güte, die wir mit unserem iLQR Setup erhalten beträgt 10^{-7}. Die Lösung hängt stark von den gewählten Kostenmatrizen ab. In dieser vorgestellten Lösung ist der X-Pulse Gauß-förmig und die u^Y Pulskomponente ist ungefähr proportional zu der Ableitung der u^X Pulsekomponente. Auch dies ist deshalb im Einklang mit den theoretischen Schlussfolgerungen der DRAG Methode für 1-Qubit Operationen. Für cross-resonance Operationen auf 2 Qubits erhalten wir in dem 2 Level Fall einen sehr glatten Verlauf der Pulsamplituden, der zu glatten Zustandsübergängen führt. Die Güte von 10^{-8} ist sehr hoch. Für den 3 Level Fall erreichen wir eine Güte von 10^{-5} und einen chaotischeren Zustandsübergang. Das ist der komplizierteste Fall für den mit der iLQR Implementierung eine gute Lösung gefunden werden konnte.
Die vier Resultate mit den jeweiligen Zustandsverläufen sind in der Abbildung 6 dargestellt:

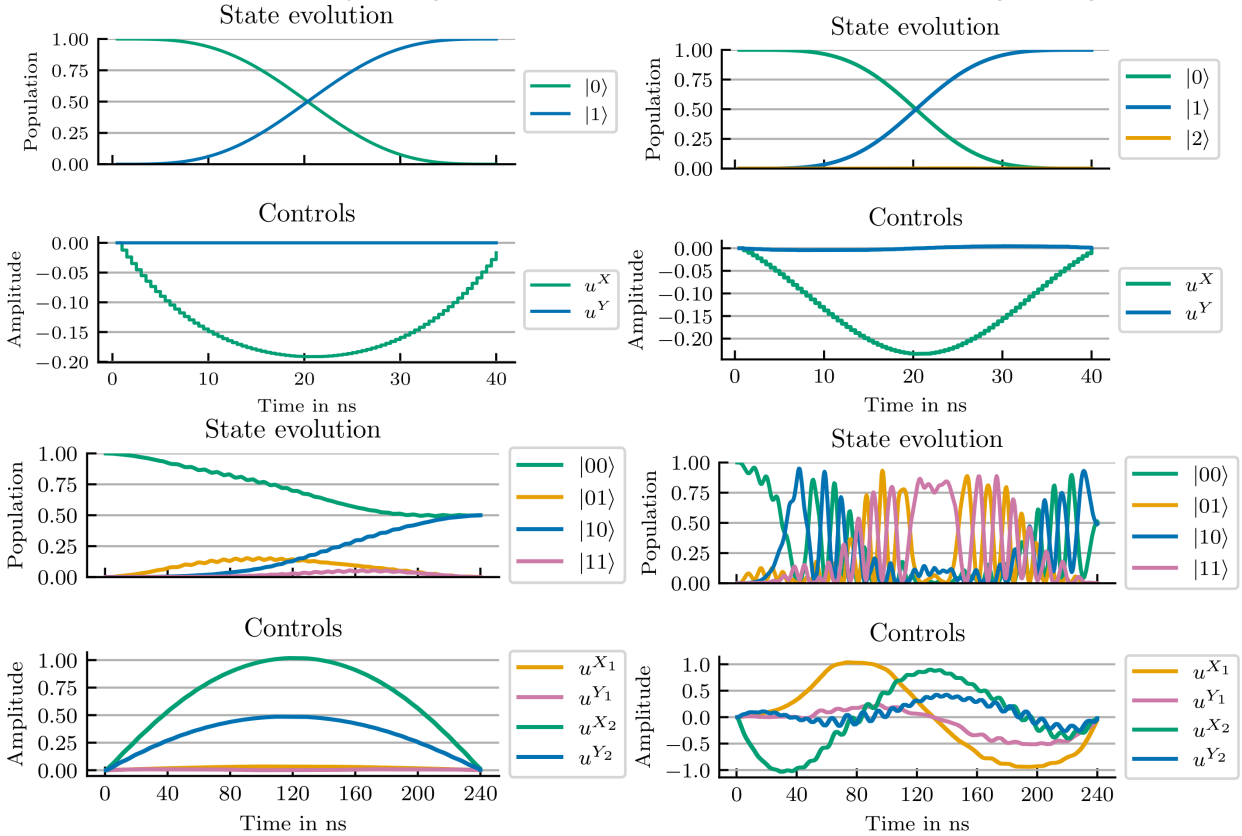


Abbildung 6: In der ersten Zeile sind die 1-Qubit X-Operation und in der zweiten Zeile die 2-Qubit cross-resonance Resultate dargestellt. In der ersten Spalte sind die 2 Level und in der zweiten Spalte die 3 Level Resultate dargestellt.

Die Ergebnisse zeigen, dass iLQR auf diese Weise für geglättete Pulsamplituden verwendet werden kann und Ergebnisse mit hoher Güte in der fehlerfreien Simulation erzielt werden können.
Die dargestellten Ergebnisse sind Beispiele, die aus einer ausgiebigen Hyperparametersuche ausgewählt wurden. Die Ergebnisse wurden auf der IEEE Quantum Week 2025 in Albuquerque, NM vorgestellt und werden demnächst in den Proceedings veröffentlicht. Die aktuelle Implementierung basiert auf dem euklidischen, quadratischen Abstand zur Zieloperation und beinhaltet nicht die Fidelity selbst. Es existieren Arbeiten zur klassischen Nutzung von iLQR unter zur Hilfenahme der Eigenschaften von Lie Gruppen. Diese Methoden sollten sich auch auf den Quantenfall (Lie Gruppe

SU(n)) erweitern lassen und sollten die Konvergenzrate erhöhen.

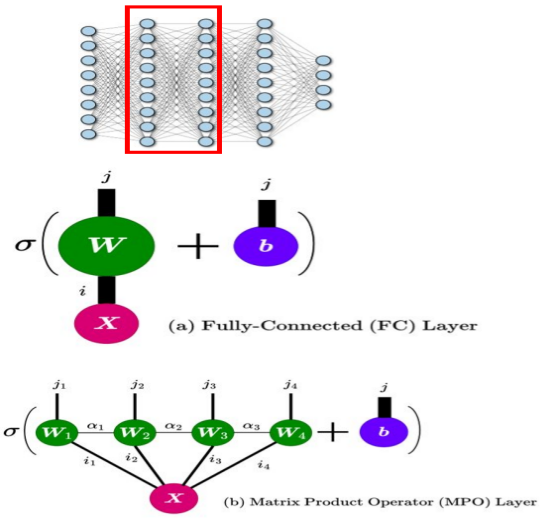
AP5000: Quantenbasiert-bestärkende Lernverfahren für robotische Verhalten

Im laufe des ersten Projektjahres wurde eine Einwertung zur Einsetzbarkeit von bestimmten QC-Algorithmen für bestärkenden Lernverfahren vorgenommen. In erster Linie wurden die hybriden quanten reinforcement learning ansätze auf Basis von PQCs als weit verbreitete Methode identifiziert. So bestanden bereits arbeiten zu PQC erweitertem DQN (https://doi.org/10.22331/q-2022-05-24-720), PPO (https://arxiv.org/abs/2212.07932, https://arxiv.org/abs/2203.14348) und SAC (https://arxiv.org/abs/2112.11921)

World models (https://doi.org/10.48550/arXiv.1809.01999) eröffnen eine Möglichkeit die Umgebung mitzulernen. Diese Methode könnte dabei helfen Umgebungsinformationen als Quantenzustände zur Verfügung zu stellen auf denen Oracle basierende Quantum Algorithmen nach Lösungen suchen. Auf der QIP 2022 in Gent wurde anhand eines Posters vorgestellt wie PQCs genutzt werden können, um Zustandstransitionen von simplen MDPs in Quantenzuständen zu repräsentieren.

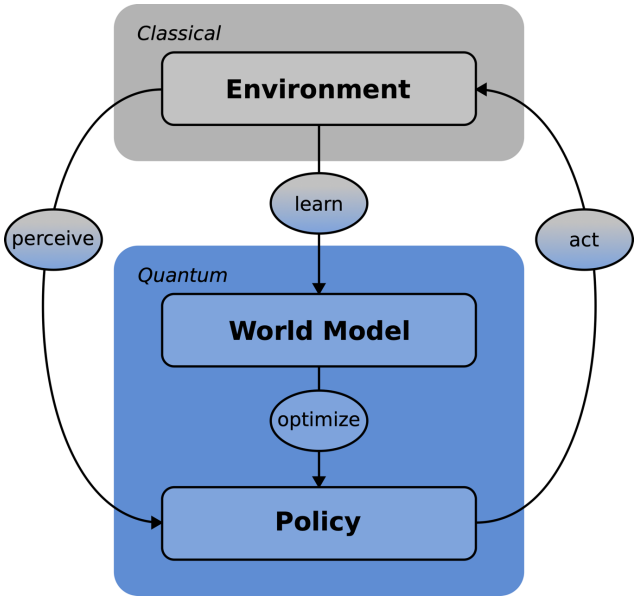
Zum Abschluss des ersten Jahres wurde die ausführliche Literaturrecherche zu bekannten Ansätzen für quantenbasiert-bestärkende Lernverfahren finalisiert. Das Ergebnis dieser Recherche ist die Identifikation von drei relevanten Kategorien. Diese sind die Quanten-inspirierten, die hybriden, sowie die voll-quantengestützten bestärkenden Lernverfahren.

Als Quanten-inspirierte Methode sind vor allem Tensornetzwerke von Interesse. Diese für Quanten-Vielteilchen-Systeme entwickelte Methode lässt sich auf neuronale Netze übertragen, um hier für eine effektive Reduktion der trainierbaren Parameter zu ermöglichen. Darüber hinaus haben sich quanten inspirierte Tensor Networks als interessante Schnittstelle zwischen klassischen und quantum Computing gezeigt. Ergebnisse aus Quantum Tensor Networks for Variational Reinforcement Learning (https://tensorworkshop.github.io/NeurIPS2020/accepted_papers/NIPS_2020_Workshop_Yiming%20(1).pdf) wurden mit dem bereitgestellten Code nachvollzogen und reproduziert. In ersten weiterführenden Tests wurden im Rahmen dieses Arbeitspakets hierzu bereits veröffentlichte Methoden implementiert und Ergebnisse reproduziert. Darüber hinaus wurden Ansätze zur Nutzung dieser Methoden in kontinuierlichen Aktionsräumen getestet. Diese Arbeiten wurden als Poster auf der QTML 2023 vorgestellt.



Die hybriden Methoden sind im Vergleich am besten etabliert. Unsere Recherche hat ergeben, dass zu drei der wichtigsten Algorithmen des bestärkenden Lernens, DQN, PPO und SAC, jeweils bereits Quanten-hybride Erweiterungen existieren. In diesen Erweiterungen werden hauptsächlich neuronale Netze, welche als allgemeine Funktionsapproximatoren dienen, gegen parametrisierte Quantenschaltkreise ausgetauscht. Diese hybriden Erweiterungen sollen im Folgenden im Rahmen des AP 3000 DFKI-seitig implementiert und evaluiert werden. Erste Implementierungen zur technischen Validierung ohne reproduzierbaren Lernerfolg, aufgrund von fehlender Hyperparameteroptimierung, von Quanten-hybridem DQN unter der Verwendung und mittels der Kombination von verfügbaren Methoden etablierter Softwareframeworks wurden bereits im zweiten Projektjahr umgesetzt. Dabei wurde außerdem auf die Effizienz der Softwareframeworks insbesondere hinsichtlich der Simulation von Quantenschaltkreisen geachtet. In ersten Tests hat

sich gezeigt, dass Pennylane mit Stablebaseline, mit den Standardeinstellungen, limitierte performance aufwies. Dies war ein zusätzlicher Anstoß für die Entwicklung eines neuen Frameworks unter Nutzung von JAX in AP3000.



In der Untergruppe der voll-quantengestützten Verfahren zeigt sich, dass die vielversprechendsten Methoden Quanten-zugängliche Umgebungen voraussetzen. Dies steht im Widerspruch zu den fundamental klassischen Umgebungen, welche im robotischen Kontext relevant sind. Allerdings versprechen die voll-quantengestützten Verfahren große Verbesserungspotenziale. Mit dem Ziel, auch klassische Umgebungen in voll-quantengestützten Verfahren nutzen zu können, haben wir in ersten Tests Übergangsfunktionen simpler klassischer Markov-Entscheidungsprozesse mittels eines parametrisierten Schaltkreises gelernt. Dieses so erlernte „Quanten-Weltmodell“ ist prinzipiell Quanten-zugänglich. Die Implementierung dieses Quanten-Weltmodells für eine simplifizierter Umgebung zeigt, dass es am kompliziertesten ist diese Untergruppe auf roboternähere Umgebung anzuwenden und bereits die Übertragung in „Quanten-Weltmodelle“ Gegenstand aktueller Forschung ist.

Bezüglich der Quanten-Weltmodelle wurde in Tests mit geringfügig erweiterten Markovschen Entscheidungsprozessen ein enormer Anstieg der benötigten Systemgröße und damit der Speicher- und Hardwareanforderungen festgestellt. Darüber hinaus wurde auch ein Anstieg der erforderlichen Lerndauer bis zur Konvergenz beobachtet. Ein nennenswerter Forschungserfolg diesbezüglich erscheint im Rahmen von QuBER-KI daher unwahrscheinlich. Daraus resultierend wurde der Fokus in AP5000 im Folgenden hauptsächlich auf die Tensornetzwerkmethoden gelegt.
Für die Tensornetzwerkmethoden konnte die Anwendung eines Matrix-Produkt-Operators (MPO) im Soft Actor-Critic (SAC) Framework realisiert werden. Diese demonstrierte eine bis zu 100-fache Reduktion der Parameteranzahl bei nur geringfügiger Verschlechterung der erzielten Performance. Diese Ergebnisse bestätigten, dass die Nutzung von MPOs zur Ersetzung eines MLP für eine verbesserte Speichereffizienz sorgt.
Darauf basierend wurden neue Experimente durchgeführt, in denen Tensornetzwerke im Zusammenhang mit Proximal Policy Optimization (PPO) angewendet wurden. Die Ergebnisse sind ähnlich zu denen der SAC-Experimente: Das Verwenden von Tensornetzwerk-Modellen erreicht eine signifikante Reduktion der Parameteranzahl (bis zu 100-fach) mit überwiegend geringen Leistungseinbußen. Allerdings zeigt die Verwendung von Tensornetzwerk-Modellen im Zusammenhang mit PPO eine langsamere Konvergenz und längere Trainingszeiten aufgrund der Rechenkosten von Tensor-Kontraktionen. Trotz dieser Nachteile zeigt die lineare Skalierung des Speicherbedarfs des MPO-basierten Ansatzes Potenzial für eine effiziente Modellentwicklung bei groß angelegten Aufgaben im Bereich des Reinforcement Learning.

Tabelle: Untersuchte Kombinationen von Algroithmus und Umgebung

	Cartpole	Pendulum	Double Pendulum	Reacher
DQN	✓	✗	✗	✗
SAC	✗	✓	✓	✓

PPO	✓	✓	✓	✓

Unter Verwendung des in AP3000 entwickelten hyperlax wurden im letzen Projektjahr umfassende numerische Experimente durchgeführt. Die drei gewählten Algorithmen (PPO, SAC, DQN) wurden mit den jeweils passenden Umgebungen (Cartpole, Pendulum, Double Pendulum, Reacher) kombiniert (siehe Tabelle). Aufgrund der integration in Zusätzlich wurde jeder Algorithmus jeweils mit klassischen multi layer perceptrons (MLP), data reupload PQCs (DRPQC) und Tensor-Netzwerk komprimierten multi layer perceptrons (TMLP) als Modell trainiert. Ergebnisse wurden jeweils über 128 agent-seitige random seeds und 8 environment-seitige random seed gemittelt. Darüber hinaus wurden jeweils 64 Hyperparametersets gesamplet und trainiert. Insgesamt läuft dies auf etwa 1,5 Millionen kombinationen aus Algorithmen Umgebungen, Modellen, Hyperparametern und Seeds hinaus. Zur Vergleichbarkeit wurden alle Trainings nach 5 Millionen trainingsschritten gestoppt.

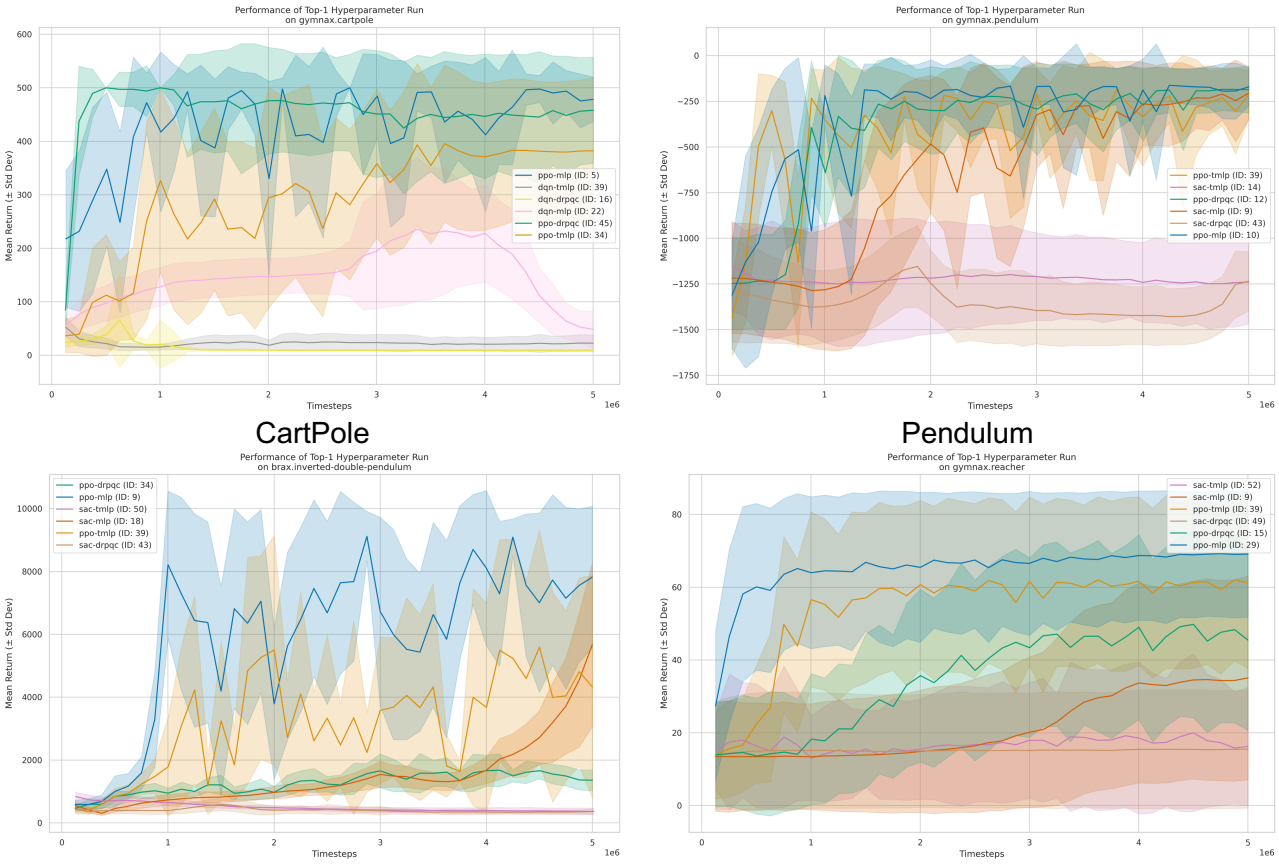
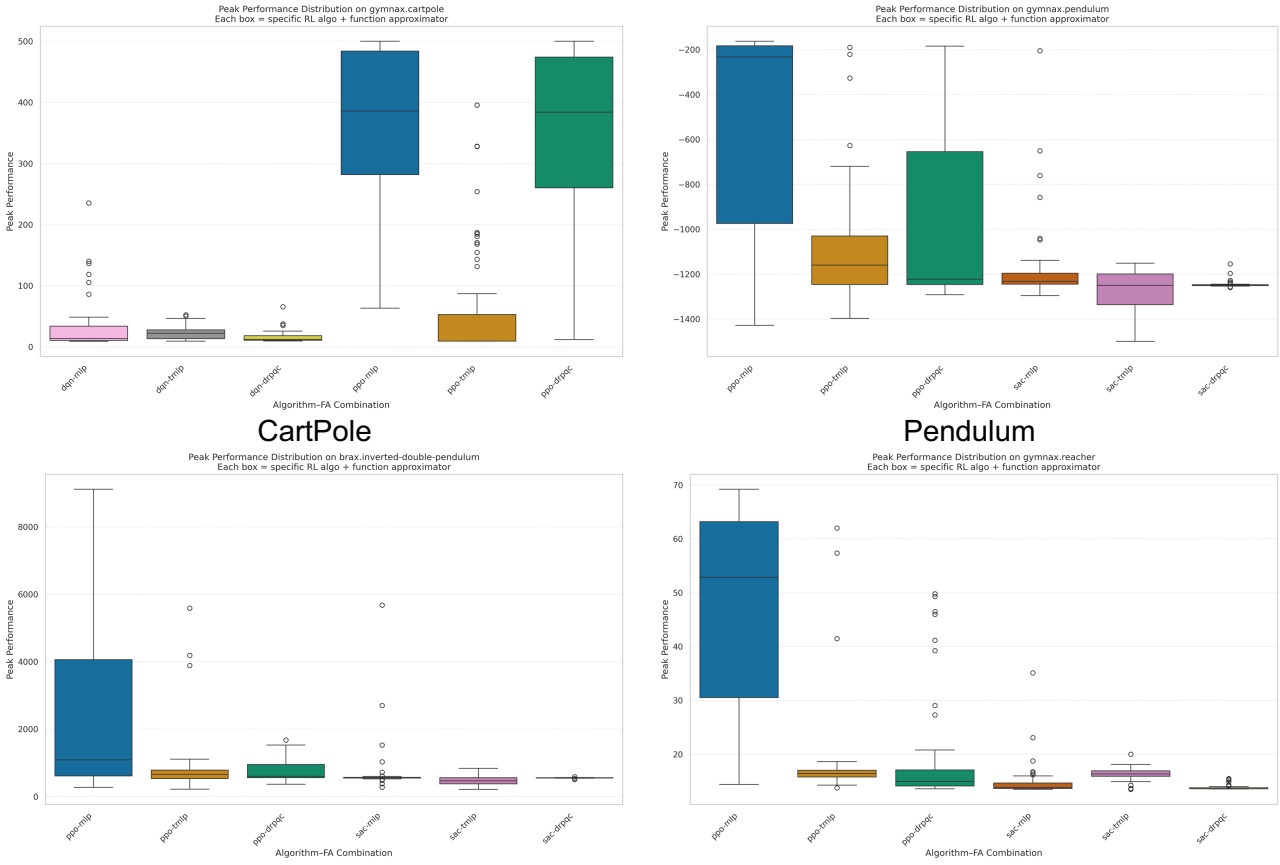


Abbildung 7: Lernverläufe der besten Hyperparameter

In Abbildung 7 sind die Lernverläufe der jeweils besten Hyperparameterkonfiguration aufgezeichnet. Die X-Achse gibt den Umgebungsspezifischen Reward an. Die Fehlerbalken zeigen die Varianz über die Random seeds. In den beiden einfacheren Umgebungen CartPole und Pendulum zeigt PPO sowohl mit MLP als auch mit DRPQC gefolgt vom TMLP. In diesen Umgebungen können die beiden neuheitlichen Methoden also dem klassischen MLP ähnliche Ergebnisse liefern. In der Double Pendulum Umgebung weist der DRPQC kaum noch Lernerfolg auf. TMLP zeigt verbesserten Lernerfolg allerdings nicht auf dem gleichen Niveau wie das klassische MLP. Für die Umgebung Reacher sehen die Ergebnisse ähnlich aus. Der DRPQC zeigt hier deutlicheren Lernerfolg bleibt aber dennoch hinter TMLP und MLP zurück. Die abnehmende Leistung von DRPQC in den

komplexeren Umgebungen könnte darauf hinweisen, dass die gewählte Architektur zu simpel ist und eine größere Tiefe und Weite des Quantenschaltkreis nötig wäre. Die teilweise vergleichbare Performance von TMLP mit MLP legt nahe, dass TMLP eine adäquate Alternative zu herkömmlichen neuronalen Netzen sein kann in Anwendungsfällen, die wenig Speicher erlauben. Insgesamt zeigt sich PPO durchgängig als stärkster Algorithmus. PPO ist für Robustheit und kurze Trainingszeiten bekannt. Es ist also möglich, dass nach zusätzlichen Zeitschritten auch die anderen Algorithmen ähnliche Performance erreichen könnten.



CartPole Pendulum

Double Pendulum Reacher

Abbildung 8: Verteilung der Performance-Höchstwerte

Abbildung 8 legt nun den Blickpunkt mehr auf die Varianz in Bezug auf die gewählten Hyperparameter. Vor allem in den beiden komplexeren Umgebungen wird deutlich, dass MLP-PPO mit Abstand am robustesten auf die Wahl von nicht optimalen Hyperparemtern reagiert. Während bei allen anderen Varianten die besten erreichten Performances einzelne Ausreißer darstellen, zeigt MLP-PPO eine Durchweg mittlere bis hohe Performance im bereich der gesampleten Hyperparameter. In den beiden einfacheren Umgebungen zeigen vor allem die DRPQCs eine ähnliche bis nur leicht schlechtere Robustheit. Durch diese Abbildungen wird deutlich, dass die TMLPs deutlich empfindlicher auf eine Veränderung der Hyperparameter reagieren.

Eine abschließende Betrachtung legt die Bewertung nahe, dass PPO mit klassischem MLP deutlich die höchsten Lernerfolge aufweist. Hyperparameteranalyse macht noch deutlicher, wie viel unempfindlicher gegenüber suboptimaler Hyperparameterkonfiguration MLP-PPO ist im vergleich zu den anderen getesteten Kombinationen. Dies unterstreicht, dass die bestehenden Ansätze für QDRL noch in einer frühen entwicklungsphase stehen und unklar ist, ob diese Ansätze zukünftig in der Anwendung besser geeignet sein werden als die bestehenden klassischen Methoden.

04 Verwertbarkeit der Ergebnisse

Die im Vorhaben *QuBER-KI* erzielten Ergebnisse lassen sich in mehrfacher Hinsicht – wissenschaftlich, technologisch und ökonomisch – verwerten.

Da es sich bei der Universität Bremen und dem DFKI um eine öffentliche Hochschule beziehungsweise eine gemeinnützige Forschungseinrichtung handelt, erfolgt keine unmittelbare wirtschaftliche Verwertung der Projektergebnisse. Gleichwohl kommt beiden Einrichtungen eine zentrale Rolle bei der Überführung der erzielten Forschungsergebnisse zu, insbesondere durch Kooperationen mit der Industrie, die Teilnahme an nationalen und europäischen Forschungsinitiativen sowie gegebenenfalls durch die Ausgründung technologiegetriebener Spin-offs.

Die im Rahmen von *QuBER-KI* entwickelten Technologien und Methoden – insbesondere die quanten-hybriden bestärkenden Lernverfahren sowie das Benchmarking-Framework für robotische Anwendungen – stellen einen grundlegenden Schritt zur Integration von Quantentechnologien in die Robotik dar. Die erfolgreiche Anwendung quantenunterstützter Lernverfahren auf Aufgabenstellungen der robotischen Navigation und Manipulation hat zu einer deutlichen Steigerung der wissenschaftlichen wie auch technologischen Konkurrenzfähigkeit im Bereich der Quantentechnologien, insbesondere im Hinblick auf robotische Anwendungskontexte beigetragen.

Die im Projekt geschaffenen Simulationsumgebungen, modularen algorithmischen Frameworks und Validierungsketten bilden eine reproduzierbare Grundlage für die weiterführende Evaluierung und Fortentwicklung quantenunterstützter bestärkender Lernverfahren. Diese Werkzeuge dienen künftig als standardisierte Testplattformen für Forschungsvorhaben im Bereich der quantenunterstützen Robotik und ermöglichen eine methodisch fundierte Bewertung neuer Quantenalgorithmen in realistischeren Anwendungsszenarien. Die öffentlich bereitgestellten Benchmarks und Vergleichsstudien aus *QuBER-KI* können als Referenzrahmen zwischen Quantencomputing und der Robotikforschung eingesetzt werden. Durch die systematische Durchführung von Experimenten und Benchmark-Analysen erfolgte eine präzise Kartierung der **aktuellen Grenzen** quantenunterstützter bestärkender Lernverfahren sowie eine Identifikation der Bereiche, in denen künftig Verbesserungen möglich erscheinen, und jener, in denen gegenwärtig klassische Verfahren noch eine höhere Leistungsfähigkeit aufweisen. Diese umfassende Analyse ermöglicht eine **gezieltere Nutzung wissenschaftlicher Ressourcen** und trägt zur **Optimierung der Ressourcenallokation innerhalb der Forschungsgemeinschaft** bei, wodurch zukünftige Forschungsaktivitäten auf besonders aussichtsreiche quantenunterstützte Strategien ausgerichtet werden können.

Für das DFKI RIC führt die Integration der *QuBER-KI*-Ergebnisse zu einer signifikanten Stärkung der wissenschaftlichen und technologischen Wettbewerbsfähigkeit im Bereich der Quantentechnologien für autonome Systeme. Die entwickelten Frameworks bilden eine wesentliche Grundlage für nachgelagerte BMBF-, BMWK- und DLR-QCI-Projekten – insbesondere mit Fokus auf quantenbasierte Steuerungs-, Lern- und Simulationsansätzen. Teile der *QuBER-KI*-Ergebnisse werden im Rahmen des Projekts Robotics Institute Germany (RIG), das über ein weitreichendes industrielles Netzwerk verfügt, präsentiert, wodurch die entscheidende Rolle des DFKI RICs, als Akteur in Deutschland, der Quantentechnologien im Kontext der Robotik auf modernstem Niveau erforscht, sichtbar wird. Diese Sichtbarkeit innerhalb des RIG trägt maßgeblich zur strategischen Positionierung des DFKI RIC im nationalen Robotikökosystem bei und stärkt zugleich das Profil Deutschlands als zukunftsorientierten Standort an der Schnittstelle zwischen Quantencomputing und der Robotikforschung.

Für die Universität Bremen stellen, die im Rahmen von *QuBER-KI* erzielten, wissenschaftlichen und methodischen Fortschritte eine wesentliche Grundlage für den nachhaltigen Aufbau von Forschungskompetenzen im Bereich des Quantum Machine Learning für die Robotik dar. Die neu entwickelten Algorithmen, Open-Source-Frameworks und wissenschaftlichen Publikationen bilden

eine tragfähige Basis für zukünftige Forschungsaktivitäten, fördern interdisziplinäre Ausbildungskonzepte und erhöhen zugleich die Attraktivität des Wissenschaftsstandorts Bremen für Nachwuchsforscherinnen und -forscher.

Darüber hinaus haben die Ergebnisse von *QuBER-KI* bereits zur Initiierung neuer Folgevorhaben und zu verstärkten Kooperationen mit Partnern aus Wissenschaft und Industrie im Rahmen der DLR Quantum Computing Initiative sowie weiterer nationaler Programme im Bereich Quanten-KI geführt. Somit hat *QuBER-KI* entscheidend zur Stärkung wissenschaftlicher Kompetenzen, zur Erweiterung bestehender Forschungsnetzwerke und zur Erhöhung der Sichtbarkeit und Attraktivität sowohl des DFKI RIC als auch der Universität Bremen beigetragen.

Insgesamt leistet das Projekt einen substanziellen Beitrag zur langfristigen Entwicklung quantenunterstützter KI-Methoden für die Robotik und verankert deutsche Forschungseinrichtungen nachhaltig an der Spitze dieses sich dynamisch entwickelnden technologischen Feldes.

05 Fortschritt bei anderen Stellen

Seit 2023 ist eine Zunahme konzeptioneller Forschungsarbeiten im Bereich Quantum Reinforcement Learning zu beobachten, wobei die meisten Beiträge weiterhin theoretischer oder simulationsbasierter Natur sind. Zunächst gab es eine Ausweitung der Forschungsaktivitäten im Bereich des Multi-Agent Quantum Reinforcement Learning (MAQRL), mit einem besonderen Fokus auf Skalierbarkeits- und Optimierungsproblematiken, die den Anwendungsbereich des quanten-hybriden bestärkenden Lernverfahren (QRL) auf einzelne Agenten und kleine Beispiele beschränkt hatten. Die Arbeiten von Kölle et al. (https://doi.org/10.5220/0012382800003636) schlagen die Einführung eines gradientenfreien evolutionären MAQRL-Frameworks auf Basis von Variational Quantum Circuits (VQCs), zur Abschwächung der Barren-Plateau-Problematik. Erste Simulationen zeigen in einfachen Testumgebungen vergleichbare Ergebnisse zu klassischen Netzen, allerdings unter stark vereinfachten Bedingungen und ohne Nachweis der Übertragbarkeit auf komplexere Szenarien. Dies liefert einen Indikator für das Potenzial parameter-effizienter quantenbasierter Kooperationsagenten. Parallel dazu entwickelten Yun et al. (https://doi.org/10.48550/arXiv.2208.11510) ein Quantum Meta Multi-Agent Reinforcement Learning (QM²ARL)-Schema, unter Ausnutzung der dualen Struktur von Quantum Neural Networks – einer Trennung von Winkel- und Pol-Parametern – zur Realisierung metabasierter Lernprozesse und schneller Anpassung zwischen Agenten. Der Einsatz von Angle-to-Pole-Regularisierung sowie einer Pole-Memory-Struktur verbesserte die Konvergenz und Anpassungsfähigkeit in nicht-stationären Umgebungen. Ergänzt wird dies durch die Arbeit von Park et al. (10.1109/MCOM.020.2300199). Mit einer Anwendung quantenbasierter MARL-Modelle im Kontext autonomer Mobilitätskooperation mittels der Einführung eines projection value measure Mechanismus, der die Kompression der Dimension des effektiven Aktionsraumes von linear auf nahezu logarithmisch ermöglicht. Außerdem deutet die quantenbasierte Variante auf eine mögliche Reduktion der Modellkomplexität hin. Insgesamt liefern die Arbeiten erste konzeptionelle Ansätze für skalierbare MAQRL-Architekturen, deren praktische Umsetzbarkeit auf NISQ-Hardware bislang jedoch nur in stark vereinfachten Experimenten untersucht wurde.

Im Jahr 2024 ergab sich eine weitere Konsolidierung des Forschungsfeldes der quanten-hybriden bestärkenden Lernverfahren durch die Ausweitung auf Multi-Agent-, hierarchische und domänenspezifische Methoden sowie durch die Bearbeitung kombinatorischer Optimierungsprobleme mittels quantenmechanischer Ansätze. Park & Kim (https://doi.org/10.4218/etrij.2024-0153) verbinden in ihrer Arbeit theoretische und ingenieurtechnische Grundlagen quantenneuronaler RL-Methoden zu einer Methode mit verbesserten Trainingsgeschwindigkeit, Skalierbarkeit und reduzierten Parametern. Außerdem werden Multiagenten Erweiterungen – etwa des zentralisierten critic mit multiplen actors – im Kontext von gemeinsamen und autonomen Lernens erörtert. In der Publikation von Zhu et al. (https://doi.org/10.1007/s40747-024-01381-8) ermöglicht die Anwendung von QRL auf das Gebiet der Sprachverarbeitung, mit einem hierarchischen quantenbasierten RL-Framework, das Relationserkennung und Entitätsextraktion in getrennten Ebenen modelliert, eine effiziente Repräsentation überlagerter Relationen sowie eine signifikante Leistungssteigerung gegenüber klassischen Verfahren. Kruse et al. (https://doi.org/10.1109/QCE60285.2024.00189) stellen ein neuartiges Hamiltonian-basiertes QRL-Konzept zur Lösung kombinatorischer Optimierungsprobleme vor. Durch eine direkte Implementierung von VQC Modellen auf Hamiltonian Ebene für kombinatorische Optimierungsprobleme wie z.B. Max-Cut und Knapsack, wird eine erhöhte Trainierbarkeit und Generalisierbarkeit über Graphen Probleme hinaus erreicht. Schließlich veröffentlichten Liu et al. (https://doi.org/10.48550/arXiv.2407.06103) ein hybrides Quantum-Train-Konzept, bei dem klassische policy Netzwerke durch eingebettete Quantenschaltkreise trainiert werden, um eine polylogarithmische Reduktion der Parameteranzahl zu erzielen, wobei die Inferenz

weiterhin klassisch bleibt. Dies bietet eine Lösungsstrategie für Datenkodierungs- und Latenzprobleme in Echtzeit-RL-Szenarien auf NISQ-Hardware. Zusammengenommen sind die Arbeiten des Jahres 2024 ein Indikator für eine Verschiebung von proof-of-concept Agenten hin zu anwendungsgetriebenen, ressourceneffizienten und strukturell angepassten QRL-Architekturen über Agenten, Hierarchien und Domänen hinweg.

Im Jahr 2025 steht im Forschungsfeld des QRL vor allem die Systematisierung der Evaluierungen im Fokus, die die früheren Arbeiten um Benchmarks, Überblicksstudien und anwendungsorientierten Frameworks ergänzen. In der Überblickstudie von Alomari & Kumar (10.1109/CAI64502.2025.00283) schlagen die Autoren die Erstellung einer umfassenden Taxonomie von QRL-Methoden vor, in der zwischen klassischen und quantenmechanischen Umgebungen sowie Trade-offs zwischen Repräsentations-, Policy- und Value basierten Methoden, unterschieden wird. Des Weiteren werden zukünftige Forschungsrichtungen, einschließlich der Optimierung von Schaltkreise für Quantensonsoren. Kruse et al. (https://doi.org/10.5220/0013393200003890) entwickelten ein vereinheitlichtes Benchmarking-Framework für fünf zentrale QRL-Klassen (basierend auf PQCs, policy gradients, free-energy, Amplitudenenkodierung, und Q-learning) mit Evaluation auf Gitterweltszenarien und der systematischen Identifizierung von Bedingungen unter denen Quanteneffekte tatsächlich lernrelevante Beiträge liefern. Das Anwendungsspektrum wurde von Liu et al. (https://doi.org/10.48550/arXiv.2507.12835) erweitert durch die Entwicklung eines quanten-hybriden RL-Frameworks für Finanzvorhersage und algorithmischen Handel, basierend auf der Kopplung eines klassischen LSTM für makroökonomische Vorhersagen mit einer asynchronen Advantage-Actor-Critic-(A3C)-Quanten Policy. Die Ergebnisse demonstrieren eine verbesserte Stabilität bzgl. der erzielten Rewards unter volatilen Marktbedingungen. Zusammengenommen sind die Beiträge des Jahres 2025 eine zunehmende methodische Konsolidierung des Forschungsfeldes, mit ersten Schritten hin zu systematischer Evaluation und Reproduzierbarkeit. Eine direkte industrielle Relevanz bleibt aber ein langfristiges Ziel.

06 Veröffentlichungen

Neben der Kommunikation von Forschungsergebnissen in internen und externen Workshops konnten im Rahmen des am DFKI RIC durchgeführten Projekts Q3UP (s. https://robotik.dfki-bremen.de/de/forschung/projekte/q3up) sowie in anderweitig durchgeführten Vorträgen für Teilnehmende aus der Industrie wesentliche Ergebnisse den möglichen Nutzer*innen präsentiert werden.

Weiterhin wurden Teile der Resultate der Forschung aus dem vorliegenden Verbundvorhaben im Rahmen der folgenden Konferenzposter und Journalartikel veröffentlicht:

- *Tensor Networks for Efficient Reinforcement Learning with Continuous Action Spaces,*
  Bolat et al., QTML 2023
- *Quantum World Models: Training PQCs on MDP Data,*
  Gross et al., QIP 2023
- *Pulse sequences with iLQR for superconducting qubits,*
  Heimann et al., QOC Workshop Berlin, 2024
- *Quantum Optimal Control Algorithms of Superconducting Qubits. A Study [...],*
  Mellentin, Master's Thesis, 2025
- *Adaptive Model-Based Control of Quadrupeds via Online System Identification using Kalman Filter,*
  Haack et al., IROS 2025
- *Quantum Deep Reinforcement Learning for Robot Navigation Tasks*
  Hohenfeld et al., IEEE Access, 2025
- *Iterative Linear Quadratic Regulator for Quantum Optimal Control,*
  Heimann et al., IEEE QCE 2025
- *A Case Study on The Effects of Hyperparameters in Quantum Deep Reinforcement Learning,*
  Bolat et al., im Review-Prozess
- *Hyperlax a Framework for Vectorized Hyperparameter Analysis,*
  Bolat et al., Software on GitHub (https://github.com/dfki-ric-quantum/hyperlax-quantum)