Experiment zur Verbesserung der Robustheit von Reinforcement Learning Policies anhand trainiertem Gegenspieler in Simulationen

Bachelorarbeit

vorgelegt am 6. März 2023

Fakultät Wirtschaft

Studiengang Wirtschaftsinformatik

Kurs WWI2020F

von

Leon Henne

Betreuerin in der Ausbildungsstätte:

IBM Deutschland GmbH
Sophie Lang
Senior Data Scientist

Unterschrift der Betreuerin

DHBW Stuttgart:

Prof. Dr. Kai Holzweißig
Studiendekan Wirtschaftsinformatik

Inhaltsverzeichnis

Abkürzungsverzeichnis

DHBW Duale Hochschule Baden-Württemberg

RL Reinforcement Learning

KPI Key Performance Indicator

MARL Multi-Agent Reinforcement Learning

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

1.1 Problemstellung

Reinforcement Learning (RL) findet heutzutage bereits Anwendung in vielerlei Forschungsprojekten wie Deepmind AlphaStar oder OpenAI Five, aber auch in Produkten und Dienstleistungen wie AWSDeepRacer oder Metas Horizon open-source RL-Plattform.[1] RL ist im Bereich des maschinellen Lernens eine Herangehensweise zur Lösung von Entscheidungsproblemen.[2] Ein Software-Agent leitet dabei durchzuführende Aktionen aus seiner Umgebung ab, mit dem Ziel die kumulierte erhaltene Belohnung zu maximieren, währenddessen sich seine Umgebung durch alle Aktionen verändert.[3] Die Umgebungen beinhalten in ihrer einfachsten Form eine simulierte Welt, welche zu jedem Zeitschritt eine Aktion entgegennimmt, und den eigenen nächsten Zustand sowie einen Belohnungswert zurückgibt.[4] Da ein Problem beim Einsatz von RL Algorithmen die Limitierungen sein können, Daten in der echten Welt zu sammeln und fürs Training zu verwenden, werden häufig hierfür Simulationsumgebungen eingesetzt.[5] Eine Limitierung können bspw. Sicherheitsaspekte sein, welche beim Training von Roboterarmen, oder sich autonom bewegenden Systemen auftreten, da die einzelnen physischen Bewegungen nicht vorhersehbar abschätzbar sind.[6] Simulationen nehmen damit als Testumgebung eine wichtige Rolle ein in der Entwicklung von Kontrollalgorithmen.[7] Insgesamt bedarf die erfolgreiche Anwendung von Reinforcement Learning demnach nicht nur effiziente Algorithmen, sondern auch geeignete Simulationsumgebungen.[8] Besonders schwierig, und daher sehr wichtig zu erforschen, ist es die Trainingsumgebung bestmöglich an die echte Welt anzupassen, sodass bspw. die Agenten für Roboter und autonome Fahrzeuge, nach dem Training mit generalisierten Policies in der Realität eingesetzt werden können.[9] In der Forschungsliteratur wird diese beschriebene Problematik als „Sim to real"-Transfer beschrieben.[10] Eine Domäne der echten Welt wird dabei eher selten ausschließlich von veränderten dynamischen Parametern und nur einer Person oder nur einer Organisation geprägt. Oftmals beeinflussen mehrere Parteien teilweise kooperierend aber auch teilweise konkurrierend den eigenen Erfolg, wie bspw. einen dem Wettbewerb unterliegenden Markt. Stellt man sich ein solches Szenario vor, ist es naheliegend, dass auch jene Einflüsse möglichst präzise in die Simulationsumgebung integriert sein müssen, um ein generalisierendes Modell erlernen zu können. Während bereits in Produkten wie Powertac nach Collins/Ketter 2022 die Simulation von Märkten entwickelt wurde, scheint der Einfluss des Gegenspielers in kompetitiven

[1] Vgl. Li 2019, S. 4
[2] Vgl. Schuderer/Bromuri/van Eekelen 2021, S. 3
[3] Vgl. Schuderer/Bromuri/van Eekelen 2021, S. 3
[4] Vgl. Reda/Tao/van de Panne 2020, S. 1
[5] Vgl. Zhao/Queralta/Westerlund 2020, S. 737
[6] Vgl. Zhao/Queralta/Westerlund 2020, S. 738
[7] Vgl. Cutler/Walsh/How 2014, S. 2
[8] Vgl. Reda/Tao/van de Panne 2020, S. 8
[9] Vgl. Slaoui u. a. 2019, S. 1
[10] Vgl. Zhao/Queralta/Westerlund 2020, S. 738

Simulationen auf die Robustheit von RL Algorithmen und demnach auf die Lösung des „Sim to real“-Transfers unerforscht.

1.2 Zielsetzung

Daher soll im Rahmen dieser Arbeit untersucht werden, ob die Integrierung eines RL basierten Gegenspielers in einer Simulation die Umgebung so beeinflussen kann, dass die erlernten Verhaltensmodelle, welche im Kontext von RL oftmals als Policies referenziert werden, robuster agieren unter den veränderten dynamischen Bedingungen und alternativen deterministischen Gegenspielern im Testszenario.

Dazu soll eine kompetitive Simulationsumgebung entwickelt werden, in welcher sich zwei konkurrierender Spieler in Form von Flugobjekten spielerisch gegenseitig bekämpfen. In der Simulation werden folgend Policies in drei verschiedenen Szenarien trainiert.

- Training mit regelbasiertem Gegenspieler unter gleichbleibenden Dynamikparametern
- Training mit RL basiertem Gegenspieler unter gleichbleibenden Dynamikparametern
- Training mit regelbasiertem Gegenspieler unter sich verändernden Dynamikparametern

Anschließend werden alle trainierten Policies in einer Reihe von Testszenarien untersucht. Jedes Testszenario verfügt dabei über festgelegte sich vom Training unterscheidende Dynamikparameter und jeweils leicht unterschiedliche Handlungspräferenzen des deterministischen Gegenspielers. Bei der Untersuchung werden jeweils die folgenden Variablen als Key Performance Indicator (KPI) betrachtet.

- durchschnittlich erzielte Belohnung
- Varianz der Belohnungen
- Anzahl an unbeabsichtigten Abstürzen

Aus der Auswertung der Testszenarien kann der Effekt des RL basierten Gegenspielers auf die Robustheit mittels des Vergleichs mit dem regelbasierten Gegenspieler und der Domain Randomization evaluiert werden.

1.3 Forschungsfrage

Aus der beschriebenen Problemstellung und der für den Rahmen dieser Arbeit festgelegten Zielsetzung ergibt sich folgende Forschungsfrage:

Kann durch den Einsatz eines mittels RL trainierten Gegenspielers die Robustheit der gelernten Policy verbessert werden?

Zur Beantwortung der Forschungsfrage werden folgende Hypothesen aufgestellt und im Rahmen der Arbeit untersucht:

Hypothese 1: *Die in den Testszenarien durchschnittlich erzielte Belohnung ist under Verwendung der Policy aus dem Training mit RL basiertem Gegenspieler signifikant und zuverlässig höher als die Policy aus dem Training mit regelbasiertem Gegenspieler.*

Hypothese 2: *Die Varianz der in den Testszenarien erzielten Belohnung ist under Verwendung der Policy aus dem Training mit RL basiertem Gegenspieler signifikant und zuverlässig geringer als die Policy aus dem Training mit regelbasiertem Gegenspieler.*

Hypothese 3: *Die in den Testszenarien erreichte Anzahl von unbeabsichtigten Abstürzen ist under Verwendung der Policy aus dem Training mit RL basiertem Gegenspieler signifikant und zuverlässig geringer als die Policy aus dem Training mit regelbasiertem Gegenspieler.*

1.4 Forschungsmethodik

Als Forschungsmethodik soll im Rahmen dieser Arbeit ein quantitatives Laborexperiment nach Recker 2021 durchgeführt werden. Hierbei wird häufig nach dem hypothetisch-deduktives Modell vorgegangen, in welchem Hypothesen formuliert, empirische Studien entwickelt, Daten gesammelt, Hypothesen anhand dessen evaluiert und gewonnene Erkenntnisse berichtet werden.[11] Eine Möglichkeit der Untersuchung der Ursache- und Wirkungsbeziehung stellt das Laborexperiment dar.[12] Dabei wird die kontrollierte Umgebung der Simulation erschaffen, deren Aufbau die unabhängige Variable darstellt. Die Metriken anhand welcher die Performance und die Robustheit der trainierten Policies gemessen werden, bilden im Experiment die abhängigen Variablen.

1.5 Aufbau der Arbeit

Insgesamt gliedert sich die Arbeit nach einem Schema von Holzweißig 2022 Die Arbeit beginnt mit einem einleitenden Kapitel in welchem Motivation, Problemstellung, Zielsetzung und Forschungsmethodik erläutert sind. Anschließend wird im zweiten Kapitel der aktuelle Stand der Forschung zu den relevanten Konzepten der Problemstellung wiedergegeben. Im dritten Kapitel wird die Forschungsmethodik dargestellt, indem die Simulationsumgebung als Messinstrument entwickelt wird sowie verschiedene Messszenarien erläutert und entsprechende Daten gesammelt werden. Daraufhin sind im folgenden vierten Kapitel die Messdaten auszuwerten und aufgestellte Hypothesen zu überprüfen. Im Zuge dessen kann ebenso die Forschungsfrage anhand der Annahme oder Ablehnung der Hypothesen beantwortet werden. Abschließend wird im letzten Kapitel ein Fazit zu den erzielten Forschungsergebnissen dargelegt und ein Ausblick auf weitere Forschung gegeben.

[11] Vgl. Recker 2021, S. S.89f.
[12] Vgl. Recker 2021, S. 106

2 Diskussion des aktuellen Stands der Forschung und Praxis

2.1 Aufbau der Literaturrecherche

In Anlehnung der Literaturrecherche nach Webster/Watson 2002 wurden alle voraussichtlich benötigten Konzepte für die Durchführung der beschriebenen Forschungsmethodik in Tabelle 1 festgehalten. Alle angeführten Konzepte wurden mittels verschiedener Suchbegriffe in Suchmaschinen, Datenbanken und Bibliotheken wie *Google Scholar* 2/28/2023, *IEEE Xplore* 2/28/2023 oder die digitale Bibliothek der Association for Computing Machinery (ACM) ACM Digital Library 2/28/2023 recherchiert. In der daraus gefunden Literatur wurden zitierte Werke ebenfalls nach den beschriebenen Konzepten durchsucht und insgesamt jede Literaturquelle in Tabelle 1 den in ihnen enthaltenen Konzepten zugeordnet.

Artikel	Konzepte							
	RL	MARL	ES	DS	KS	DR	RRLP	LE
Sutton/Barto 2018	X							
Li 2019	X							
Zhao/Queralta/Westerlund 2020	X		X			X		
Wang/Hong 2020	X							
Zhang/Wu/Pineau 2018	X		X				X	
Cutler/Walsh/How 2014	X							
Canese u. a. 2021	X	X						
Reda/Tao/van de Panne 2020	X		X			X		
Ningombam 2022	X							
Wong u. a. 2022	X	X						
Schuderer/Bromuri/van Eekelen 2021	X	X						

Tab. 1: Konzept Matrix für Artikel zu Simulationsumgebungen und zur Robustheit RL Algorithmen nach Webster/Watson 2002. Legende: RL (Reinforcement Learning), MARL (Multi-Agent Reinforcement Learning), ES (Entwicklung von Simulationsumgebungen), DS (Dronensimulation), KS (kompetitive Simulationsumgebungen), DR (Domain Randomization), RRLP (Robustheit von RL Policies), LE (Laborexperimente)

2.2 Verstärkendes Lernen

Verstärkendes Lernen oder als RL in der Fachsprache bezeichnet, definiert einen konzeptionellen Ansatz zielorientiertes Lernen von Entscheidungen zu verstehen und zu automatisieren.[13] Da-

[13] Vgl. Sutton/Barto 2018, S. 13

bei besteht der Fokus darauf, dass ein Agent aus der direkten Interaktion mit seiner Umgebung lernt, ohne das explizite Überwachung notwendig ist.[14] Der Agent lernt über die Zeit eine optimale Strategie zur Lösung des Entscheidungsproblems aus dem Ausprobieren und Scheitern mittels verschiedener Aktionen die gewünschte Veränderung in seiner Umwelt herzustellen.[15] Notwendig dabei ist es, dass der Agent den Zustand seiner Umgebung wahrnehmen, und auch durch entsprechende Aktionen beeinflussen kann, sodass die Erreichung des Zielzustandes möglich ist.[16] Zur Erreichung dieses Zielzustandes muss der Agent alle Aktionen entdecken, welche ihm die größtmögliche kumulierte Belohnung liefern, wobei Aktionen nicht nur die unmittelbare sondern auch zukünftige Belohnungen beeinflussen.[17] Zusammengefasst lässt sich die beschriebene Interaktion des Agenten mit seiner Umgebung wie folgt in Abbildung 1 darstellen.

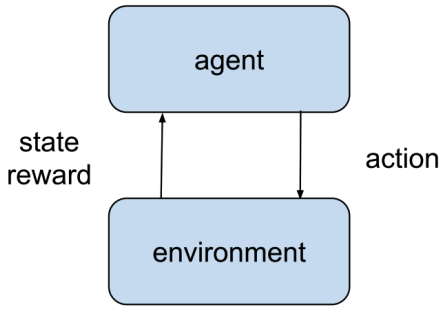

Abb. 1: vereinfachte Darstellung der Interaktion zwischen dem Agenten und seiner Umgebung[18]

Ein Standardaufbau einer Aufgabe für verstärkendes Lernen kann demnach verstanden werden, als sequentielles Entscheidungsproblem zu dessen Lösung ein Agent zu jedem diskreten Zeitschritt eine Aktion ausführt, welche den Zustand der Umgebung verändert.[19] Betrachtet man die technische Umsetzung einer solchen Interaktion zwischen dem Agenten und dessen Umgebung, wird häufig zur Modellierung ein Markov Entscheidungsprozess verwendet. Im Kontext von RL ist der Entscheidungsprozess definiert nach einem Tupel aus folgenden Elementen:[20]

- Alle Zustände S
- Alle Aktionen A
- initiale Zustandsverteilung $p0(S)$
- Übergangswahrscheinlichkeit $T(S_{t+1}|S_t, A_t)$
- Belohnungswahrscheinlichkeit $R(r_{t+1}|S_t, A_t)$

[14] Vgl. Sutton/Barto 2018, S. 13
[15] Vgl. Li 2019, S. 4
[16] Vgl. Sutton/Barto 2018, S. 2
[17] Vgl. Sutton/Barto 2018, S. 1
[18] Enthalten in: Li 2019, S. 5
[19] Vgl. Zhao/Queralta/Westerlund 2020, S. 2
[20] Vgl. Zhang/Wu/Pineau 2018, S. 2

Zum Finden der optimalen Strategie existieren modellbasierende und modellfreie Algorithmen des verstärkenden Lernens.[21] Bei modellbasierenden Algorithmen wird das Umgebungsverhalten, also die Übergangs- und Belohnungswahrscheinlichkeiten als bekannt vorausgesetzt.[22] Unter modellbasierenden Algorithmen wird dynamische Programmierung eingesetzt, um mittels Strategieevaluation und Strategieiteration die optimale Strategie zu finden.[23] Unter modellfreien Algorithmen werden die drei verschiedenen Ansätze Wertebasierend, Strategiebasierend und Akteur-Kritiker basierend unterschieden[24] Der Agent im Kontext von modellfreien RL Methoden kennt nur die Zustände S und die Aktionen A, jedoch nicht die Umgebungsverhalten T und die Belohnungswahrscheinlichkeit R.[25] Fasst man die Klassifizierung der Algorithmen und Methoden von RL zusammen, lässt sie sich wie folgt darstellen:

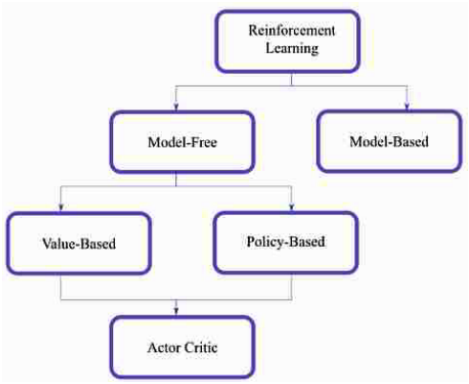

Abb. 2: Klassifizierung von Algorithmen im Bereich des RL[26]

2.2.1 Wertebasierende Methoden

Der Agent sucht in diesem Kontext die optimale Strategie π^*, welche allen Zuständen S die jeweilige Aktion $A(S)$ zuordnet, sodass die kummulierte Belohnungswahrscheinlichkeit $R(r_{t+1}|S_t, A_t)$ über alle Zeitschritte t maximal ist.[27] Neben dieser kurzfristigen direkten Belohnung müssen auch die langfristigen zukünftigen Belohnungen aus den neuen Zuständen betrachtet werden, wofür das Konzept der Wertigkeit eingeführt wird.[28] Über eine Zustands- oder Aktionswertigkeitsfunktion, oftmals als Q-Funktion referenziert, wird eine Vorhersage über die zu erwartende kumulierte abgezinste zukünftige Belohnung berechnet.[29] Durch den Abzinsungsfaktor $\gamma \in [0, 1)$ wird der Einfluss zukünftiger Belohnungen nach ihrer zeitlichen Reihenfolge priorisiert.[30] Mit der Wertigkeitsfunktion kann evaluiert werden, welche Strategie langfristig am erfolgreichsten ist, da bspw. manche Aktionen trotz geringer sofortiger Belohnung einen hohen Wert aufweisen können, wenn

[21] Vgl. Wang/Hong 2020, S. 3
[22] Vgl. Wang/Hong 2020, S. 3
[23] Vgl. Li 2019, S. 5
[24] Vgl. Li 2019, S. 5
[25] Vgl. Cutler/Walsh/How 2014, S. 2
[26] Enthalten in: Canese u. a. 2021, S. 6
[27] Vgl. Reda/Tao/van de Panne 2020, S. 2
[28] Vgl. Wang/Hong 2020, S. 3
[29] Vgl. Li 2019, S. 5
[30] Vgl Li 2019, S. 5

aus dem zukünftigen Zustand eine hohe Belohnung zu erwarten ist.[31] Die Wertigkeitsfunktion und die daraus berechneten Wertigkeiten von Aktionen oder Zuständen werden über alle Zeitschritte neu geschätzt und stellen mit die wichtigste Komponenten in Algorithmen des verstärkenden Lernens dar.[32] Methoden basierend auf diesem Wertigkeitswert lernen eine Schätzfunktion der Wertigkeit für alle Zustände ($V_\pi(s) \forall S$) und alle Zustandsaktions-Paare ($Q_\pi(s_t, a_t) \forall s, a \in (S, A)$) der optimalen Strategie π^* durch aktualisieren der folgenden Funktionen eins und zwei:[33]

$$(1)\ Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \right]$$

$$(2)\ V(s_t) = \max_a Q(s_t, a|\omega)$$

Aus den geschätzten Wertigkeit jedes Zustandsaktions Paares kann die optimale Strategie $\pi^*(s)$ durch $\arg\max_a Q(s, a)$ bestimmt werden.[34]

2.2.2 Strategiebasierende Methoden

Methoden, welche die Strategie durch deren direkte Parametrisierung anstelle einer Bewertung aller Handlungsalternativen mittels Wertigkeitsfunktion optimieren, werden als strategiebasierend bezeichnet.[35] Diese Methodik kann beim Trainieren deterministischer Strategien zu unerwarteten Aktionen führen, weshalb häufig das Optimieren einer Wahrscheinlichkeitsverteilung für alle Aktionen bevorzugt wird.[36] Als Subklasse der RL Methoden wird der statistische Gradientenabstieg verwendet um die parametrisierte Strategie π_θ hinsichtlich der maximalen langfristigen kumulierten Belohnung zu optimieren.[37] Die Strategie π_θ oder auch $\pi(a|s, \theta)$ beschreibt dabei die Wahrscheinlichkeit Aktion a im Zustand s auszuwählen unter dem Parametervektor θ.[38] Zur Optimierung der Strategie wird die Funktion der kumulierten Belohnungen J nach dem Parameter der Gewichte θ wie folgt in Formel drei abgeleitet und der optimierte Parametervektor anhand Formel vier aktualisiert.[39]

$$(3)\ \nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{i=1}^{N} \left[\left(\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) \right) \left(\sum_{t=1}^{T} r(s_t, a_t) \right) \right]$$

$$(4)\ \theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$$

Zusammengefasst kann Formel vier und fünf dabei so interpretiert werden, dass die logarithmierte Wahrscheinlichkeit Aktion a_t im Zustand s_t auszuwählen erhöht werden soll, wenn a_t in einer höheren kumulierten Belohnung resultiert.[40]

[31] Vgl. Sutton/Barto 2018, S. 6
[32] Vgl. Sutton/Barto 2018, S. 6f.
[33] Vgl. Zhang/Wu/Pineau 2018, S. 2
[34] Vgl. Zhang/Wu/Pineau 2018, S. 2
[35] Vgl. Zhang/Wu/Pineau 2018, S. 2
[36] Vgl. Ningombam 2022, S. 3
[37] Vgl. Ningombam 2022, S. 3
[38] Vgl. Sutton/Barto 2018, S. 321
[39] Vgl. Wang/Hong 2020, S. 6
[40] Vgl. Wang/Hong 2020, S. 6

2.2.3 Akteur-Kritiker Methoden

Unter Akteur-Kritiker Methoden werden hybride wertebasierende und strategiebasierende Methoden verstanden, welche zugleich die Strategie optimieren und eine Wertefunktion approximieren.[41] Die strategiebasierende Methodik mit der lernenden Strategie agiert dabei als Akteur, wohingegen die Wertefunktion, welche jeder Aktion und jedem Zustand einen Behlohnungswert zuweist, als Kritiker handelt.[42] Der Akteur wählt somit aus seiner Wahrscheinlichkeitsverteilung die auszuführende Aktionen aus, während der Kritiker diese anhand seiner Wertigkeit bewertet.[43] Betrachtet man den Trainingsprozess von Akteur-Kritiker basierten Methoden ist dieser wie folgt aufgebaut:[44]

- Aktueller Zustand der Umgebung als Eingabe dem Akteur und Kritiker übergeben
- Akteur liefert eine auszuführende Aktion basierend auf dem Umgebungszustand
- Der Kritiker bekommt die Aktion als Eingabe und berechnet dessen Wertigkeit mittels Q-Funktion
- Durch die Wertigkeit seiner Aktion kann der Akteur seine Strategie anpassen
- Mit der neuen Strategie führt der Akteur die nächste Aktion im folgenden Zustand aus
- Die Q-Funktion des Kritikers wird mit den neuen Informationen aus der erhaltenen Belohnung angepasst

2.2.4 Abgrenzung zu Multi-Agent Reinforcement Learning (MARL) Algorithmen

Innerhalb dieses Unterkapitels soll der beschriebene Aufbau von RL Algorithmen und deren Optimierungsproblem zu den von MARL Systemen abgegrenzt werden. Bei MARL Systemen wird anstatt einem Agenten eine Menge von Agenten eingesetzt welche alle mit ihrer Umgebung interagieren um den Weg der Zielerreichung zu lernen.[45] Dieser Ansatz dient dazu Problemstellungen welche nicht vollständig durch einen Agenten lösbar sind zu bearbeiten.[46] Einsatzgebiete von MARL sind dabei unter anderem das Routing von Netzwerkpaketen, Wirtschaftsmodellierung oder zusammenhängende Robotersysteme.[47] Je nach Ziel und der demnach definierter Belohnungsfunktion können die Agenten auf die drei unterschiedlichen Arten vollständig kooperativ, vollständig kompetitiv und der Mischung aus beiden miteinander interagieren.[48] Aus der unterschiedlichen Interaktion jedes Agenten mit der selben Umgebung ergibt sich der Unterschied, dass die Umgebungsdynamik aus der Kombination aller Aktionen der Agenten beeinflusst wird

[41] Vgl. Zhang/Wu/Pineau 2018, S. 2f.
[42] Vgl. Sutton/Barto 2018, S. 321
[43] Vgl. Ningombam 2022, S. 3
[44] Vgl. Ningombam 2022, S. 4
[45] Vgl. Wong u. a. 2022, S. 6
[46] Vgl. Canese u. a. 2021, S. 1
[47] Vgl. Canese u. a. 2021, S. 1
[48] Vgl. Canese u. a. 2021, S. 8f.

anstatt aus der Aktion des einzelnen Agenten.[49] Da dieser Effekt auch die Annahme der Stationarität von Markov Entscheidungsprozessen verletzt, bedarf dieUmgebung auch einer anderen Representation.[50] Ein Konzept was dafür häufig verwendet ist das Markov Spiel, welches sich anders als der Entscheidungsprozess durch einen mehrdimensionalen Aktions- und Belohnungsraum aus der Kombination aller N Agenten auszeichnet.[51] Betrachtet man die Limitierungen von MARL erkennt man aus den beschriebenen Punkten die Herausforderungen der nicht vorhandenen Stationarität und der Skalierbarkeit, welcher sich die Herausforderung der teilweisen Beobachtbarkeit der Umgebung anschließt.[52]

2.2.5 Limitierungen und Herausforderungen von RL

Trotz signifikanter Errungenschaften birgt der Einsatz von den besprochenen RL Algorithmen weiterhin Limitierungen und Risiken für ungewolltes Verhalten.[53]

Eine der Herausforderungen zeigt sich bei der Representation der Agentenumwelt, da RL stark auf diesem Konzept basiert.[54] Daraus ergibt sich die Aufgabe, die Umwelt und dessen Verhalten sowie die Wahrnehmung durch den Agenten realitätsgetreu und präzise zu gestalten.[55] Neben der Definition und Wahrnehmung des Umweltverhaltens ist die Spezifikation des Ziels des Agenten ein ebenso kritischer Teil, da unerwartete Intentionen aus der Zielstellung abgeleitet werden könnten.[56] Zusätzlich teilen RL Algorithmen auch Herausforderungen aus anderen Gebieten des maschinellen Lernens wie Genauigkeit, Interpretierbarkeit und die im Rahmen dieser Arbeit untersuchte Robustheit von Modellen.[57]

Eine weitere Limitierung stellt der große Suchraum an Aktionen und das unbekannte Verhalten der Umgebung dar. Dies sorgt dafür, dass häufig die Effizienz einzelner Daten sehr gering ist und die Abwägung zwischen Exploration neuer Strategie und der Optimierung bekannter Verhaltensmuster ein wichtiger Bestandteil ist.[58] Aufgrund der geringen Effizienz der Daten aber des dennoch hohen Bedarfs an bewerteter Agentenerfahrung wird häufig auf simulierte Daten zurückgegriffen.[59] Simulierte Daten werden dabei häufig von möglichst hoch qualitativen Simulationsumgebungen bereitgestellt, da zu dem hohen Bedarf der Methodik häufig Limitierungen in der Sammlung von Daten in der echten Welt bestehen.[60]

Aufgrund dieser Bedeutung der Simulationsumgebung für RL Algorithmen und deren Transfer in die echte Welt wird im nachfolgenden Kapitel dessen Entwicklung genauer betrachtet.

[49] Vgl. Wong u. a. 2022, S. 2
[50] Vgl. Wong u. a. 2022, S. 6
[51] Vgl. Canese u. a. 2021, S. 4
[52] Vgl. Canese u. a. 2021, S. 9ff.
[53] Vgl. Li 2019, S. 7
[54] Vgl. Sutton/Barto 2018, S. 8
[55] Vgl. Sutton/Barto 2018, S. 7
[56] Vgl. Li 2019, S. 7
[57] Vgl. Li 2019, S. 7
[58] Vgl. Li 2019, S. 7
[59] Vgl. Zhao/Queralta/Westerlund 2020, S. 7
[60] Vgl. Li 2019, S. 8

3 Durchführung des Laborexperiments

4 Ergebnisse des Laborexperiments

5 Reflexion und Forschungsausblick

Anhang

Anhangverzeichnis

Anhang 1: Interview Transkripte

Anhang 1/1: Interview Transkript: Mitarbeiter eines Unternehmens

Literaturverzeichnis

ACM Digital Library (2/28/2023): ACM Digital Library. URL: https://dl.acm.org/.

Canese, L./Cardarilli, G. C./Di Nunzio, L./Fazzolari, R./Giardino, D./Re, M./Spanò, S. (2021): Multi-Agent Reinforcement Learning: A Review of Challenges and Applications. In: *Applied Sciences* 11.11, S. 4948. DOI: 10.3390/app11114948. URL: https://www.mdpi.com/2076-3417/11/11/4948.

Collins, J./Ketter, W. (2022): Power TAC: Software architecture for a competitive simulation of sustainable smart energy markets. In: *SoftwareX* 20, S. 101217. ISSN: 2352-7110. DOI: https://doi.org/10.1016/j.softx.2022.101217. URL: https://www.sciencedirect.com/science/article/pii/S2352711022001352.

Cutler, M./Walsh, T. J./How, J. P. (2014): Reinforcement learning with multi-fidelity simulators. In: *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. DOI: 10.1109/icra.2014.6907423.

Google Scholar (2/28/2023). URL: https://scholar.google.de/.

Holzweißig, K. (2022): Wissenschaftliches Arbeiten. In: 6.6.

IEEE Xplore (2/28/2023). URL: https://ieeexplore.ieee.org/Xplore/home.jsp.

Li, Y. (2019): Reinforcement Learning Applications. DOI: 10.48550/ARXIV.1908.06973. URL: https://arxiv.org/abs/1908.06973.

Ningombam, D. D. (2022): Deep Reinforcement Learning Algorithms for Machine-to-Machine Communications: A Review. In: *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE. DOI: 10.1109/icccnt54827.2022.9984457.

Recker, J. (2021): Scientific research in information systems: A beginner's guide. Second Edition. Progress in IS. Cham: Springer International Publishing. ISBN: 9783030854362. URL: https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=6789173.

Reda, D./Tao, T./van de Panne, M. (2020): Learning to Locomote: Understanding How Environment Design Matters for Deep Reinforcement Learning. In: *Motion, Interaction and Games*. Hrsg. von Daniele Reda/Tianxin Tao/Michiel van de Panne. New York, NY, USA: ACM, S. 1–10. DOI: 10.1145/3424636.3426907.

Schuderer, A./Bromuri, S./van Eekelen, M. (2021): Sim-Env: Decoupling OpenAI Gym Environments from Simulation Models. In: *International Conference on Practical Applications of Agents and Multi-Agent Systems*. Springer, Cham, S. 390–393. DOI: 10.1007/978-3-030-85739-4{\textunderscore}39. URL: https://link.springer.com/chapter/10.1007/978-3-030-85739-4_39.

Slaoui, R. B./Clements, W. R./Foerster, J. N./Toth, S. (2019): Robust Domain Randomization for Reinforcement Learning. In: *CoRR* abs/1910.10537. arXiv: 1910.10537. URL: http://arxiv.org/abs/1910.10537.

Sutton, R. S./Barto, A. G. (2018): Reinforcement Learning, second edition: An Introduction. MIT Press. ISBN: 9780262352703.

Wang, Z./Hong, T. (2020): Reinforcement learning for building controls: The opportunities and challenges. In: *Applied Energy* 269, S. 115036. ISSN: 0306-2619. DOI: 10.1016/j.apenergy.2020.115036.

Webster, J./Watson, R. T. (2002): Analyzing the Past to Prepare for the Future: Writing a Literature Review. In: *MIS Q.* 26.2, S. xiii–xxiii. ISSN: 0276-7783.

Wong, A./Bäck, T./Kononova, A. V./Plaat, A. (2022): Deep multiagent reinforcement learning: challenges and directions. In: *Artificial Intelligence Review*. ISSN: 0269-2821. DOI: 10.1007/s10462-022-10299-x.

Zhang, A./Wu, Y./Pineau, J. (2018): Natural Environment Benchmarks for Reinforcement Learning. DOI: 10.48550/ARXIV.1811.06032. URL: https://arxiv.org/abs/1811.06032.

Zhao, W./Queralta, J. P./Westerlund, T. (2020): Sim-to-Real Transfer in Deep Reinforcement Learning for Robotics: a Survey. In: *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE. DOI: 10.1109/ssci47803.2020.9308468.

Erklärung

Ich versichere hiermit, dass ich meine Bachelorarbeit mit dem Thema: *Experiment zur Verbesserung der Robustheit von Reinforcement Learning Policies anhand trainiertem Gegenspieler in Simulationen* selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Ich versichere zudem, dass die eingereichte elektronische Fassung mit der gedruckten Fassung übereinstimmt.

(Ort, Datum) (Unterschrift)