

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

Bachelorarbeit

vorgelegt am 28. Februar 2023

Fakultät Wirtschaft

Studiengang Wirtschaftsinformatik

Kurs WWI2020F

von

LEON HENNE

Betreuerin in der Ausbildungsstätte: DHBW Stuttgart:

IBM Deutschland GmbH  
Sophie Lang  
Senior Data Scientist

Prof. Dr. Kai Holzweißig  
Studiendekan Wirtschaftsinformatik

Unterschrift der Betreuerin

# Inhaltsverzeichnis

<b>Abkürzungsverzeichnis</b>	<b>III</b>
<b>Abbildungsverzeichnis</b>	<b>IV</b>
<b>Tabellenverzeichnis</b>	<b>V</b>
<b>1 Einleitung</b>	<b>1</b>
1.1 Problemstellung . . . . .	1
1.2 Zielsetzung . . . . .	2
1.3 Forschungsfrage . . . . .	2
1.4 Forschungsmethodik . . . . .	3
1.5 Aufbau der Arbeit . . . . .	3
<b>2 Diskussion des aktuellen Stands der Forschung und Praxis</b>	<b>4</b>
2.1 Aufbau der Literaturrecherche . . . . .	4
2.2 Verstärkendes Lernen . . . . .	4
<b>3 Durchführung des Laborexperiments</b>	<b>7</b>
<b>4 Ergebnisse des Laborexperiments</b>	<b>8</b>
<b>5 Reflexion und Forschungsausblick</b>	<b>9</b>
<b>Anhang</b>	<b>10</b>
<b>Literaturverzeichnis</b>	<b>12</b>

# Abkürzungsverzeichnis

**DHBW** Duale Hochschule Baden-Württemberg

**RL** Reinforcement Learning

**KPI** Key Performance Indicator

**MARL** Multi-Agent Reinforcement Learning

# Abbildungsverzeichnis

1	vereinfachte Darstellung der Interaktion zwischen dem Agenten und seiner Umgebung	5
---	---	---

# Tabellenverzeichnis

1	Konzept Matrix für Artikel zur Simulation und zur Robustheit RL Algorithmen nach Webster/Watson 2002. Legende: RL (Reinforcement Learning), MARL (Multi-Agent Reinforcement Learning), ES (Entwicklung von Simulationsumgebungen), DS (Dro- nensimulation), KS (kompetitive Simulationsumgebungen), DR (Domain Randomi- zation), RRLP (Robustheit von RL Policies), LE (Laborexperimente) . . . . .	4
---	---	---

# 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.<sup>1</sup> RL ist im Bereich des maschinellen Lernens eine Herangehensweise zur Lösung von Entscheidungsproblemen.<sup>2</sup> 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.<sup>3</sup> 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.<sup>4</sup> 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.<sup>5</sup> 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.<sup>6</sup> Simulationen nehmen damit als Testumgebung eine wichtige Rolle ein in der Entwicklung von Kontrollalgorithmen.<sup>7</sup> Insgesamt bedarf die erfolgreiche Anwendung von Reinforcement Learning demnach nicht nur effiziente Algorithmen, sondern auch geeignete Simulationsumgebungen.<sup>8</sup> 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.<sup>9</sup> In der Forschungsliteratur wird diese beschriebene Problematik als „Sim to real“-Transfer beschrieben.<sup>10</sup> 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

---

<sup>1</sup>Vgl. Yuxi 2019, S. 4

<sup>2</sup>Vgl. Schuderer/Bromuri/van Eekelen 2021, S. 3

<sup>3</sup>Vgl. Schuderer/Bromuri/van Eekelen 2021, S. 3

<sup>4</sup>Vgl. Reda/Tao/van de Panne 2020, S. 1

<sup>5</sup>Vgl. Zhao/Queralta/Westerlund 2020, S. 737

<sup>6</sup>Vgl. Zhao/Queralta/Westerlund 2020, S. 738

<sup>7</sup>Vgl. Cutler/Walsh/How 2014, S. 2

<sup>8</sup>Vgl. Reda/Tao/van de Panne 2020, S. 8

<sup>9</sup>Vgl. Slaoui u. a. 2019, S. 1

<sup>10</sup>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 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 durchschnittlich erzielte Belohnung ist unter 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 Belohnung ist unter 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 Anzahl von Abstürzen ist unter 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.<sup>11</sup> Eine Möglichkeit der Untersuchung der Ursache- und Wirkungsbeziehung stellt das Laborexperiment dar.<sup>12</sup> Dabei wird die kontrollierte Umgebung der Simulation erschaffen, in welcher dessen Aufbau die unabhängige Variable darstellt. Die Metriken anhand 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 durchgeführt, indem die Simulationsumgebung als Messinstrument entwickelt wird sowie verschiedene Messszenarien erläutert und entsprechende Daten gesammelt werden. Daraufhin sind im folgenden vierten Kapitel die gesammelten 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.

---

<sup>11</sup>Vgl. Recker 2021, S. S.89f.

<sup>12</sup>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

Tab. 1: Konzept Matrix für Artikel zur Simulation 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.<sup>13</sup> Dabei besteht der Fokus darauf das ein Agent aus der direkten Interaktion mit seiner Umgebung lernt, ohne das explizite Überwachung notwendig ist.<sup>14</sup> 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.<sup>15</sup> Notwendig

---

<sup>13</sup>Vgl. Sutton/Barto 2018, S. 13

<sup>14</sup>Vgl. Sutton/Barto 2018, S. 13

<sup>15</sup>Vgl. Yuxi 2019, S. 4

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.<sup>16</sup> 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.<sup>17</sup> Zusammengefasst lässt sich die beschriebene Interaktion des Agenten mit seiner Umgebung wie folgt in Abbildung 1 darstellen.

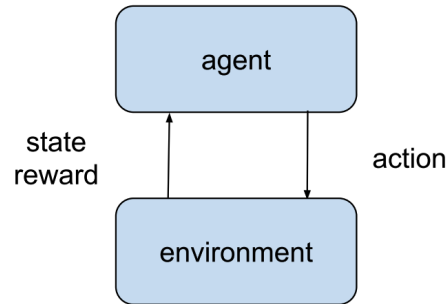


Abb. 1: vereinfachte Darstellung der Interaktion zwischen dem Agenten und seiner Umgebung<sup>18</sup>

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.<sup>19</sup> 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 sechs Elementen:<sup>20</sup>

- Alle Zustände  $S$
- Alle Aktionen  $A$
- initiale Zustandsverteilung  $p_0(S)$
- Übergangswahrscheinlichkeit  $T(S_{t+1}|S_t, A_t)$
- Belohnungswahrscheinlichkeit  $R(r_{t+1}|S_t, A_t)$
- Abzinsungsfaktor  $\gamma \in [0, 1)$

Der Agent sucht in diesem Kontext die optimale Strategie  $\pi_\phi$ , 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.<sup>21</sup>

---

<sup>16</sup>Vgl. Sutton/Barto 2018, S. 2

<sup>17</sup>Vgl. Sutton/Barto 2018, S. 1

<sup>18</sup>Enthalten in: Yuxi 2019, S. 5

<sup>19</sup>Vgl. Zhao/Queralta/Westerlund 2020, S. 2

<sup>20</sup>Vgl. Zhang/Wu/Pineau 2018, S. 2

<sup>21</sup>Vgl. Reda/Tao/van de Panne 2020, S. 2

Neben dieser kurzfristigen direkten Belohnung müssen auch die langfristigen zukünftigen Belohnungen aus den neuen Zuständen betrachtet werden, wofür das Konzept der Wertigkeit eingeführt wird.<sup>22</sup> Über eine Zustands- oder Aktionswertigkeitsfunktion, oftmals als Q-Funktion referenziert, wird eine Vorhersage über die zu erwartende kumulierte abgezinste zukünftige Belohnung berechnet.<sup>23</sup> Durch den Abzinsungsfaktor  $\gamma$  wird der Einfluss zukünftiger Belohnungen nach ihrer zeitlichen Reihenfolge priorisiert.<sup>24</sup> 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 aus dem zukünftigen Zustand eine hohe Belohnung zu erwarten ist.<sup>25</sup> 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.<sup>26</sup> Methoden basierend auf diesem Wertigkeitswert lernen eine Schätzfunktion der Wertigkeit für alle Zustände ( $V_\pi(s) \forall s$ ) und alle Zustandsaktions-Paare ( $Q_\pi(s, a) \forall s, a \in (S, A)$ ) der optimalen Strategie  $\pi$  durch aktualisieren der folgenden Funktionen:<sup>27</sup>

- 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.<sup>28</sup>

Neben den Methoden welche auf dem Konzept der Wertigkeit basieren können Strategie basierende Methoden die Strategie direkt durch bspw. ihren Gradienten optimieren.

---

<sup>22</sup>Vgl. Wang/Hong 2020, S. 3

<sup>23</sup>Vgl. Yuxi 2019, S. 5

<sup>24</sup>Vgl. Yuxi 2019, S. 5

<sup>25</sup>Vgl. Sutton/Barto 2018, S. 6

<sup>26</sup>Vgl. Sutton/Barto 2018, S. 6f.

<sup>27</sup>Vgl. Zhang/Wu/Pineau 2018, S. 2

<sup>28</sup>Vgl. Zhang/Wu/Pineau 2018, S. 2

### 3 Durchführung des Laborexperiments

## 4 Ergebnisse des Laborexperiments

## 5 Reflexion und Forschungsausblick

# Anhang

## Anhangverzeichnis

Anhang 1	Interview Transkripte . . . . .	11
Anhang 1/1	Interview Transkript: Mitarbeiter eines Unternehmens . . . . .	11

## **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/>.
- 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>.
- 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\_{\text{underscore}}39. URL: [https://link.springer.com/chapter/10.1007/978-3-030-85739-4\\_39](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.
- Yuxi, L. (2019):** Reinforcement Learning Applications. DOI: 10.48550/ARXIV.1908.06973. URL: <https://arxiv.org/abs/1908.06973>.
- 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)