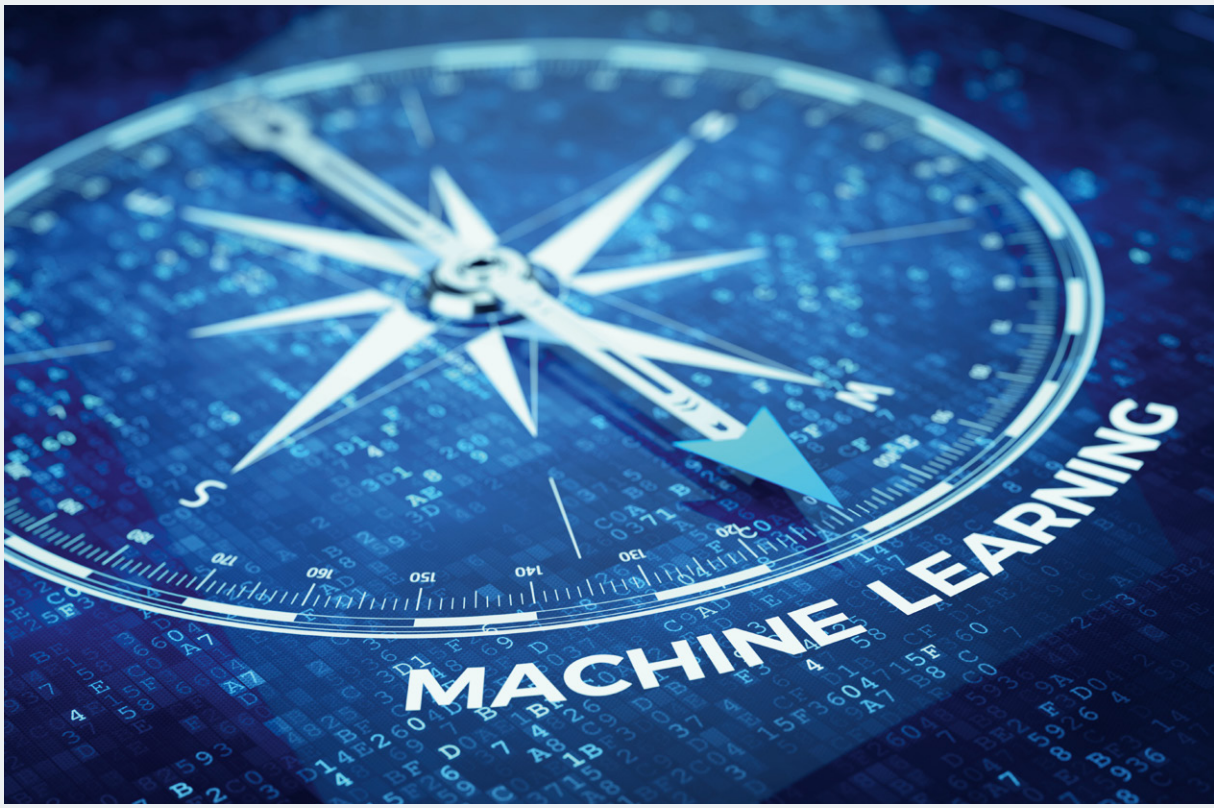
Leitfaden
Selbstlernende Produktionsprozesse

Einführungsstrategie für Reinforcement Learning in der industriellen Praxis



in Kooperation mit

IfU Institut für Unternehmenskybernetik

FKM Forschung im VDMA

VDMA

Forum Industrie 4.0

Editorial

Dietmar Goericke

Judith Binzer

Liebe Mitglieder,
sehr geehrte Damen und Herren,

Industrie 4.0 nimmt eine Schlüsselrolle ein, damit der Maschinen- und Anlagenbau langfristig erfolgreich bleibt. Digitalisierung, Vernetzung und die Integration von neuen Informations- und Internettechnologien in Produkte und Prozesse eröffnen dabei neue Geschäftspotenziale. Machine Learning ist eine wichtige Technologie zur Umsetzung der Zukunftsvision Industrie 4.0. Als Teilgebiet der künstlichen Intelligenz bringt Machine Learning für den Maschinen- und Anlagenbau spannende und neue Ansätze in der Optimierung von Produkten und Prozessen.

Vor diesem Hintergrund soll der vorliegende Leitfaden Orientierung und Unterstützung geben. Den Unternehmen wird ein Werkzeug zur Entwicklung einer eigenen Einführungsstrategie für die Methoden des Machine Learnings, insbesondere des Reinforcement Learning, zur Verfügung gestellt. Es werden dazu Grundlagen und Begrifflichkeiten erläutert sowie Leitfragen definiert, die den Unternehmen helfen, eine Einführungsstrategie zu finden. Dazugehörige Werkzeugkästen helfen bei der Beantwortung der Leitfragen.

Der Leitfaden ist das Ergebnis des Projekts „InPulS – Intelligente und selbstlernende Produktionsprozesse“. InPulS wurde als vorwettbewerbliches Forschungsprojekt des VDMA-Forum Industrie 4.0 in Kooperation mit dem Institut für Unternehmenskybernetik e. V. (IfU) und einem projektbegleitenden VDMA-Industriearbeitskreis durchgeführt. Gefördert wurde das Projekt vom Forschungskuratorium Maschinenbau (FKM) e.V. und dem VDMA in der Zeit vom 01. Oktober 2017 bis 30. September 2019.

Der VDMA realisiert mit diesem Leitfaden einen weiteren Umsetzungsbaustein für die Praxis und erweitert damit die VDMA-Leitfaden-Serie zu Industrie 4.0. Das VDMA-Forum Industrie 4.0 versteht sich als Wegbereiter in die Industrie-4.0-Welt für seine Mitgliedsunternehmen. Zugleich ist es Netzwerkplattform für den Dialog und den Erfahrungsaustausch. Wir hoffen, dass es uns gelungen ist, mit der vorliegenden Publikation einen Ratgeber anzubieten, der den Maschinenbauunternehmen bei dem Thema Künstlicher Intelligenz und Machine Learning als Orientierung dient und praxisnah unterstützen kann.

Unser Dank gebührt Prof. Dr. rer. nat. Sabina Jeschke vom Institut für Unternehmenskybernetik e. V. (IfU) sowie ihren Mitarbeiterinnen und Mitarbeitern für die wissenschaftliche Erarbeitung des Leitfadens. Zudem gilt es, den beteiligten VDMA-Mitgliedern für ihr Engagement im projektbegleitenden Arbeitskreis und insbesondere dem Vorsitzenden Herrn Dieter Herzig von der AZO GmbH + Co. KG zu danken.

Wir wünschen Ihnen eine spannende Lektüre.
Ihre

Dietmar Goericke
Geschäftsführer VDMA-Forum Industrie 4.0 und Forschungskuratorium Maschinenbau (FKM) e.V.

Judith Binzer
VDMA-Forum Industrie 4.0
Forschung & Innovation

Inhaltsverzeichnis

Leitfaden Selbstlernende Produktionsprozesse

Einführungsstrategie für Reinforcement Learning in der industriellen Praxis

Prof. Dr. Sabina Jeschke

Künstliche Intelligenz durchdringt alle Branchen und inspiriert neue Technologien in den verschiedensten Anwendungsfeldern. Auch vor der industriellen Produktion macht dieser Innovationstreiber keinen Halt - es ist unbestritten, dass KI die Produktionsstätten der Zukunft auf ein effizienteres Niveau heben wird. Machine Learning gilt als der Schlüssel zur wirtschaftlichen Realisierung kleiner Losgrößen bis hin zu „Losgröße 1“.

Derzeit werden zwei vielversprechende Ansätze verfolgt: das auf großen Datensätzen basierende Supervised und Unsupervised Learning und das auf Trial-and-Error-Verstärkungen basierende Reinforcement Learning (RL). Reinforcement Learning Algorithmen finden neue und unbekannte Lösungen, die über das menschliche Verständnis der Prozesse hinausgehen. Diese neue Herangehensweise eröffnet ein enormes Potential.

Zu welchen Leistungen Reinforcement Learning Methoden in der Lage sind zeigen die beeindruckenden Ergebnisse von z. B. Googles AlphaGo. Das Erlernen eines Spiels ist jedoch nicht zu vergleichen mit der Steuerung eines Produktionsprozesses. Ein Spiel ist eine streng überwachte, strukturierte Umgebung, Im industriellen Kontext greifen die Rahmenbedingungen und Unsicherheiten der realen Welt und es treten unvorhergesehene Ereignisse und Störungen auf. Reinforcement Learning erlaubt, eine an diese Umgebungsbedingungen angepasste Strategie zu erlernen. Genau diese Flexibilität und Adaptivität ist Kern und Stärke der Reinforcement Learning Methode.

Während große Unternehmen i.d.R. über F&E-Abteilungen verfügen, in denen solche Verfahren untersucht werden können, stellen entsprechende Eigenentwicklungen den Mittelstand vor außerordentliche Herausforderungen. Für diese Mehrheit der deutschen Unternehmen stellen sich im 4.0-Zeitalter folgende zentrale Fragestellungen:

- Wie bleibt der deutsche Mittelstand in Zeiten von 4.0 wettbewerbsfähig?
- Welche Produktionsverfahren sind besonders zugänglich für die Integration von KI im Sinne des Kosten-Nutzen-Faktors?
- Wie können fachliche Kompetenzen im Bereich der KI effizient und nachhaltig aufgebaut werden?
- Welche neuen Geschäftsmodelle oder -erweiterungen lassen sich durch den Einsatz von KI aufbauen?

Die Einführung von KI erfordert Veränderungen auf allen Ebenen: Mitarbeiter müssen qualifiziert werden, neue Berufsbilder entstehen, Prozesse verändern sich, neue Geschäftsmodelle verändern den Markt. Dieser Handlungsleitfaden bietet einen Einstieg in das Thema selbstlernende Prozesssteuerung und dient als Orientierungshilfe zur Einführung solcher Verfahren.

Das vorgestellte Anwendungsszenario zeigt, dass trotz der besonderen Anforderungen die Anwendung von Reinforcement Learning Methoden im industriellen Kontext möglich ist und zu einer deutlichen Effizienzsteigerung führt.

Deutschland investiert in die Erforschung Künstlicher Intelligenz wie nie zuvor. Der deutsche Mittelstand muss diese Möglichkeit zur Mitgestaltung unbedingt nutzen!

Prof. Dr. Sabina Jeschke

Berlin und Strömsund, im Sommer 2019

Management Summary

Das maschinelle Lernen als Bestandteil der Industrie 4.0 wird heutzutage als ein entscheidendes Instrument zur Effizienzsteigerung und eine Chance für die Entwicklung neuer Geschäftsmodelle gehandelt. Häufig liegt in diesem Zusammenhang jedoch der Fokus auf digitalen Anwendungsfeldern und es fehlt an Erfahrung wie die Methoden des maschinellen Lernens im industriellen Bereich Anwendung finden können. Insbesondere der Bereich des Reinforcement Learning zur autonomen Steuerung von Produktionsprozessen ist in der Industrie noch wenig bis gar nicht erschlossen.

Zielsetzung des Leitfadens Selbstlernende Produktionsprozesse ist es daher, mittelständischen Maschinen- und Anlagenbauern ein Werkzeug zur Entwicklung einer eigenen Einführungsstrategie für die Methoden des maschinellen Lernens, insbesondere des Reinforcement Learning, zur Verfügung zu stellen. Hierbei wird zum einen eine Einführung in die Begrifflichkeiten und Konzepte im Bereich Reinforcement Learning gegeben und zum anderen auf die konkreten Besonderheiten der industriellen Anwendung eingegangen. So stellt dieser Leitfaden keine vorgefertigte Lösung zur Umsetzung von industriellem Reinforcement Learning dar, sondern soll vielmehr Unterstützung zur Entwicklung einer individuellen Einführungsstrategie geben.

Dieser Leitfaden zielt darauf ab, ein Werkzeug zur Entwicklung einer eigenen Einführungsstrategie für die industrielle Anwendung von Reinforcement Learning zur Verfügung zu stellen.

Reinforcement Learning ist ein Teilgebiet des maschinellen Lernens, welches sich insbesondere dafür eignet eine intelligente Steuerungsstrategie zu erlernen. Es wird empfohlen eine solche Steuerungsstrategie zuerst in einem Pilotprojekt mit klar definierten Rahmenbedingungen zu erstellen, da viele Faktoren für ein erfolgreiches autonomes Lernen berücksichtigt werden müssen. Dieser Handlungsleitfaden soll dazu dienen ein solches Pilotprojekt auszuwählen und dieses Projekt anschließend als geeignete Problemstellung für Reinforcement Learning zu formulieren.

Der Handlungsleitfaden ist in acht Abschnitte unterteilt. Zunächst wird das Potential von industriellem Reinforcement Learning und der notwendige Perspektivwechsel von einer Anlagen- zu einer Prozesssteuerung erläutert. Dann werden die wichtigsten Begrifflichkeiten des Reinforcement Learning eingeführt. Den Kern des Leitfadens bildet eine Reihe von Leitfragen, welche im Unternehmen gestellt und beantwortet werden müssen, um einen geeigneten Anwendungsfall zu finden und eine Einführungsstrategie für diese Anwendung zu entwickeln. Des Weiteren wird ein Werkzeugkasten zur Verfügung gestellt, der bei der Beantwortung dieser Leitfragen unterstützen soll. Dieser Handlungsleitfaden ist im Zusammenhang mit dem vom VDMA initiierten Forschungsprojekt „InPulS – Intelligente und selbstlernende Produktionsprozesse“ entstanden. In diesem Projekt wurde Reinforcement Learning zum einen zum Erlernen eines autonomen Montageprozesses und zum anderen für einen selbstlernenden Prozess auf einem Schüttgutförderer angewandt. Die Erfahrungen, Ergebnisse und Erkenntnisse dieser Anwendungsbeispiele werden am Ende dieses Leitfadens zusammengefasst.

Einleitung und Zielsetzung

Ausgangslage

Heutige Automatisierungssysteme werden zunehmend mit einer Vielzahl an Sensoren ausgestattet, welche immer stärker miteinander vernetzt sind. Mithilfe dieser Sensoren kann der Zustand eines Systems erfasst werden. Die gewonnenen Daten stellen in der industriellen Automatisierung ein erhebliches Potential dar, indem sie neuartige Konzepte und Lösungen ermöglichen. Zur Ausschöpfung dieses Potentials wird den Methoden des maschinellen Lernens große Bedeutung beigemessen.

Maschinelles Lernen (Machine Learning) ist ein Teilgebiet der künstlichen Intelligenz und umfasst eine Vielzahl unterschiedlicher Konzepte und Methoden. Allen gemeinsam ist, dass sie einen Pool an gesammelten Daten ausnutzen, um ein Modell für eine gewünschte Aufgabe anzutrainieren. Es werden drei Klassen unterteilt: überwachtes Lernen (Supervised Learning), unüberwachtes Lernen (Unsupervised Learning) und bestärkendes Lernen (Reinforcement Learning). Das überwachte Lernen wird meist zur Klassifizierung und zur Regression verwendet. Mit den Methoden des unüberwachten Lernens können innerhalb der Daten bestehende Muster und Gruppen entdeckt werden und die einzelnen Datenpunkte diesen Gruppen zugeordnet werden. Das bestärkende Lernen basiert auf dem Prinzip der Belohnung bzw. Bestrafung. Abbildung 1 zeigt eine Übersicht über die drei Prinzipien. Dieser Handlungsleitfaden beschäftigt sich mit dem bestärkenden Lernen. Eine detaillierte Übersicht zum Thema Machine Learning findet sich im VDMA Quick Guide – Machine Learning im Maschinen- und Anlagenbau (VDMA Software und Digitalisierung).

Die genannten Methoden eignen sich für unterschiedlich komplexe Anwendungsfälle. Im industriellen Kontext lassen sich damit insbesondere Potentiale in den Bereichen der Prozessüberwachung, -optimierung und -steuerung ausschöpfen. Abbildung 2 gibt eine Übersicht über die drei Bereiche.

Der Bereich der **Prozessüberwachung** profitiert direkt von der zunehmenden Sensorausstattung der Produktionsanlagen. Mithilfe dieser Sensoren kann der aktuelle Zustand der Anlage überwacht werden oder bereits für eine einfache Prädiktion des zukünftigen Zustands verwendet werden. Diese Technologien bieten eine erhöhte Prozessqualität durch bessere Überwachung, reduzierte Ausfallzeiten sowie eine erhöhte Prozesssicherheit. Diese Überwachung kann meist schon mit einfachen Analysemethoden realisiert werden.

Darauf aufbauend kann ein Prozess durch maschinelles Lernen optimiert werden. Diese Art der **Prozessoptimierung** bietet Unternehmen großes Potential in Form von Effizienzsteigerung und Kostenreduktion. In der Optimierung wird iterativ ein Optimum gefunden, z.B. eine optimale Task-Reihenfolge, und dieses Optimum angesteuert. Hier werden unter anderem Methoden aus dem Bereich des überwachten und unüberwachten Lernens verwendet, die beispielsweise zur Planung oder als Entscheidungshilfe eingesetzt werden.

Machine Learning		
Überwachtes Lernen	**Unüberwachtes Lernen**	**Bestärkendes Lernen**
Klassifizierung	Zugehörigkeitsbildung	Interaktion mit der Umgebung
Regression	Kategorisierung	Belohnungsprinzipien

Abbildung 1: Übersicht über die verschiedenen Verfahren des maschinellen Lernens.

	Prozessüberwachung	Prozessoptimierung	Prozesssteuerung
Stellt zur Verfügung	Situationserkennung und vorrauschauende Informationen	Planung und Entscheidungshilfe	Automatisierte Antwort auf Änderungen in der Umwelt
bietet	Erhöhte Qualität, reduzierte Ausfallzeiten, verringerte Fehlmengen	Erhöhte Effizienz, verbesserte Nutzung, größere Erträge, effektiveres Design	Erhöhte Produktion und Produktivität, geringere Arbeitskosten, weniger Ausschuss
benötigt	Datenquellen z.B. vernetzte Sensoren	**Prozessüberwachung** + ausgereifte Analysetools	**Prozessoptimierung** + Integration der physischen Systeme, z.B. Roboter
Methoden	Visualisierung und deskriptive Statistik	Supervised und unsupervised Methoden	Reinforcement Learning

Komplexität

Abbildung 2: Anwendung des maschinellen Lernens in den verschiedenen Bereichen der Automatisierung mit unterschiedlichem Komplexitätsniveau.

Die Komplexität nimmt weiter zu, wenn man einen Prozess über Machine Learning-Verfahren steuern möchte, da die Optimierung und die Ausführung auf dem physischen System sehr stark miteinander verzahnt sind. Diese Schritte werden in der Prozesssteuerung abwechselnd oder gleichzeitig ausgeführt. Bei der direkten Interaktion mit dem Prozess stößt man mit Methoden des überwachten und unüberwachten Lernens an die Grenzen des Möglichen. An dieser Stelle ist der Einsatz von Reinforcement Learning besonders vielversprechend, da diese Verfahren diese direkte Interaktion mit dem Prozess benötigen, um erfolgreich zu sein.

Für die Prozessüberwachung und -optimierung existieren bereits zahlreiche Beispielprojekte (vgl. Quellen) die eindrucksvoll die Leistungsfähigkeit von maschinellem Lernen demonstrieren. Die Steuerung von industriellen Prozessen ist bisher aufgrund seiner hohen Komplexität jedoch nur wenig untersucht worden. Dabei ermöglicht der Einsatz vom maschinellen Lernen zur Steuerung industrieller Prozesse eine hohe Adaptivität an unvorhersehbare und unmodellierbare Ereignisse, die unter anderem durch natürliche Rohstoffschwankungen, Wetterschwankungen und Verschleiß hervorgerufen werden können.

Prinzip des Reinforcement Learning

Reinforcement Learning bedeutet Lernen durch Ausprobieren. Die Art zu Lernen ähnelt der des Menschen in der frühen Kindheit, z.B. wenn ein Kind Laufen lernt. In diesem Falle weiß das Kind, wie der Zielzustand aussieht und probiert nach dem Prinzip Versuch-und-Irrtum solange aus, bis dieser Zielzustand erreicht ist. In jedem Versuchsschritt wird gelernt, ob ein Verhalten zielführend ist oder nicht. Am Anfang gleicht das Verhalten des Kindes eher wahllosem Ausprobieren und wird dann mit der Zeit zunehmend zielgerichteter. Überträgt man dieses Prinzip auf einen Produktionsprozess, werden innerhalb des vorgegebenen Handlungsspielraumes der Aktuatoren verschiedene Steuerungssignale ausprobiert und die darauffolgende Reaktion anhand geeigneter Kriterien bewertet. Die einzelnen Kriterien werden in einer Bewertungsfunktion zusammengefasst, mithilfe derer die Prozessgüte beschrieben ist. Mit der Zeit wird so eine intelligente Steuerungsstrategie gelernt.

Potential des Reinforcement Learning

Das selbstständige Lernen einer intelligenten Steuerungsstrategie birgt enormes Potential in sich, da es eine Prozessoptimierung ohne Modellierung und dadurch eine hohe autonome Flexibilität ermöglicht.

Manuell eingestellte Steuerungsstrategien beruhen auf über Jahre gesammeltem Expertenwissen. Fällt dieses Expertenwissen weg, haben Unternehmen häufig Schwierigkeiten geeigneten Ersatz zu finden. Reinforcement Learning kann helfen komplexe und von Expertenwissen unabhängige Steuerungsstrategien zu entwickeln.

Ein weiterer Vorteil von mit Reinforcement Learning gelernten Strategien ist, dass sie ausgetretene Pfade verlassen und ganz neue Lösungen für bekannte Steuerungsprobleme finden können, die oft effizienter als herkömmliche Strategien sind.

Die bekannteste Anwendung von Reinforcement Learning ist AlphaGo, das erste Computer Programm, welches den amtierenden Weltmeister im traditionellen chinesischen Spiel Go geschlagen hat. Neben der Überlegenheit der künstlichen Intelligenz ist hier besonders beeindruckend, dass es sich dabei um eine völlig neuartige Spielstrategie handelte. Der Lösungsraum der KI war damit größer als das von Menschen über Jahrhunderte gelernte und optimierte Wissen.

Aber auch im Bereich Robotik ist Reinforcement Learning mit großem Erfolg unter anderem zum Erlernen einer Fügeaufgabe (Schoettler et al. 2019) oder eines Drohnenfluges (Sadeghi und Levine 2016) angewandt worden.

Simulationen sind ein weiteres Tool um Parametereinstellungen auszuprobieren und diese so zu optimieren. Häufig stellt sich hier allerdings das Problem der Modellbildung. Reinforcement Learning kann bei Prozessen helfen, welche zu komplex sind, um ihr Verhalten in einer Simulation abzubilden. Mithilfe von Reinforcement Learning können Steuerungsstrategien sowohl für sehr komplexe Prozesse als auch für komplexe Umweltbedingungen gelernt werden, ohne diese explizit modellieren zu müssen.

Ein weiterer Vorteil von Reinforcement Learning ist die Möglichkeit eine Steuerungsstrategie in Echtzeit zu ermitteln, während dies mithilfe einer Simulation zu rechenintensiv wäre.

Abbildung 3: Zielsetzung des Leitfadens.

Durch die Komplexität des Verfahrens und die direkte Integration in das physische System bringt der Einstieg in die industrielle Anwendung von Reinforcement Learning zu Anfang relative viele Anforderungen mit sich. Sind diese erfüllt, zeigen zahlreiche Beispiele eine eindeutige Überlegenheit von selbstgelernten Steuerungsstrategien im Vergleich zu manuell erstellten Strategien.

Zielsetzung und Projekthintergrund

Dieser Handlungsleitfaden ist im Zusammenhang mit dem Projekt „InPulS – Intelligente und selbstlernende Produktionsprozesse“ entstanden. Im Rahmen dieses Projektes ist eine selbstlernende Prozessregelung am Beispiel eines pneumatischen Schüttgutförderers und eines kraftgeregelten Fügeprozesses mit einem Roboterarm entwickelt worden. InPulS wurde als vorwettbewerbliches Forschungsprojekt des VDMA-Forum Industrie 4.0 in Kooperation mit dem Institut für Unternehmenskybernetik e. V. (IfU) des Cybernetics Lab der RWTH Aachen und einem projektbegleitenden VDMA-Industriearbeitskreis durchgeführt. Gefördert wurde das Projekt vom Forschungskuratorium Maschinenbau (FKM) e.V. und VDMA in der Zeit vom 01. Oktober 2017 bis 30. September 2019.

Ziel dieses Handlungsleitfadens ist die Erarbeitung einer Einführungsstrategie für die Anwendung von Reinforcement Learning in der industriellen Automatisierung. Der Leser soll befähigt werden die Potentiale sowie die notwendigen Rahmenbedingungen für die industrielle Anwendung zu erkennen. Mithilfe von Leitfragen und einem zugehörigen Werkzeugkasten wird ein Tool bereitgestellt, um die Einführung von Reinforcement Learning zu erleichtern.

Zielgruppe

Adressat dieses Leitfadens sind Unternehmen, welche ihre Produktionssysteme mithilfe von Reinforcement Learning effizienter gestalten möchten und dabei neben einer Orientierung für die Risiken und Potentiale nach Hilfestellungen für eine Einführungsstrategie suchen.

Struktur des Leitfadens

Im Folgenden werden die Unterschiede zwischen einer konventionellen Steuerung und einer selbstlernenden Steuerung mithilfe von Reinforcement Learning erläutert. Weiterhin werden die notwendigen Begriffe und Prinzipien im Bereich Reinforcement Learning eingeführt. Anschließend werden Leitfragen und der passende Werkzeugkasten zur Beantwortung dieser Leitfragen bereitgestellt, um die Auswahl eines geeigneten Pilotprojektes zu erleichtern. Es folgt eine Skizze der aktuellen algorithmischen Ansätze im Bereich Reinforcement Learning, welche als Ausgangspunkt für eine tiefergehende Recherche dienen soll. Danach wird das Vorgehen zur Integration von Reinforcement Learning anhand eines Pilotprojektes erläutert. Der Fokus liegt hier insbesondere auf der Frage, welcher der verschiedenen Akteure für welchen Prozessschritt verantwortlich ist. Zuletzt werden die beschriebenen Handlungsempfehlungen anhand zwei konkreter Anwendungsbeispiele, einem autonomen Montageprozess und einer selbstlernenden Steuerung eines pneumatischen Schüttgutförderers illustriert.

Von der Anlagen- zur Prozesssteuerung

Wenn Reinforcement Learning eingesetzt werden soll, muss sich zunächst die Sicht auf die Steuerung grundlegend ändern. In der konventionellen Steuerung werden hauptsächlich Anlagenparameter in Form der vorhandenen Aktuatoren und ihren jeweiligen Stellbereichen betrachtet. Diese Parameter werden mithilfe von Erfahrungs- und Literaturkennwerten eingestellt und so lange verändert bis ein gutes Steuerergebnis zu beobachten ist.

Für eine Steuerung über Reinforcement Learning werden nicht mehr die Parameter einzelner Aktuatoren betrachtet. Stattdessen müssen Parameter gefunden werden, die den Prozess als Ganzes beschreiben. Das führt dazu, dass die Anlage zunächst einmal parameterfrei wird, übrig bleiben nur Prozessparameter. Es muss also beschrieben werden, was einen guten Prozess charakterisiert. Diese Prozessgüte wird dann quantisiert mithilfe einer Bewertungsfunktion, die im Reinforcement Learning Kontext Kostenfunktion genannt wird. Diese quantisierte Prozessgüte dient im Lernverfahren dazu, ausprobierte Steuerungsstrategien zu bewerten und gute Strategien zu bestärken, bzw. schlechte zu vermeiden. Auf diese Weise lernt das System selbstständig die Anlagenparameter so einzustellen, dass eine optimale Prozessgüte erreicht wird.

Die durch Reinforcement Learning gelernte Steuerungsstrategie kann für einen Experten zunächst ungewohnt sein, da häufig vollkommen neue Parameterbereiche entdeckt werden. Gleichzeitig liegt genau darin das Potential dieses Ansatzes. In diesem Kontext wird das Expertenwissen nicht mehr dazu eingesetzt, um die Anlagenparameter einzustellen. Die neue Aufgabe des Experten ist es hingegen, eine gute Kostenfunktion zu erstellen. Diese muss die Prozessgüte abbilden und sicherstellen, dass ein sicheres und effizientes Verhalten gelernt wird. Diese Aufgabe ist von essentieller Bedeutung für die erfolgreiche Implementierung der selbstlernenden Steuerung und muss für jeden Prozess individuell angepasst werden.

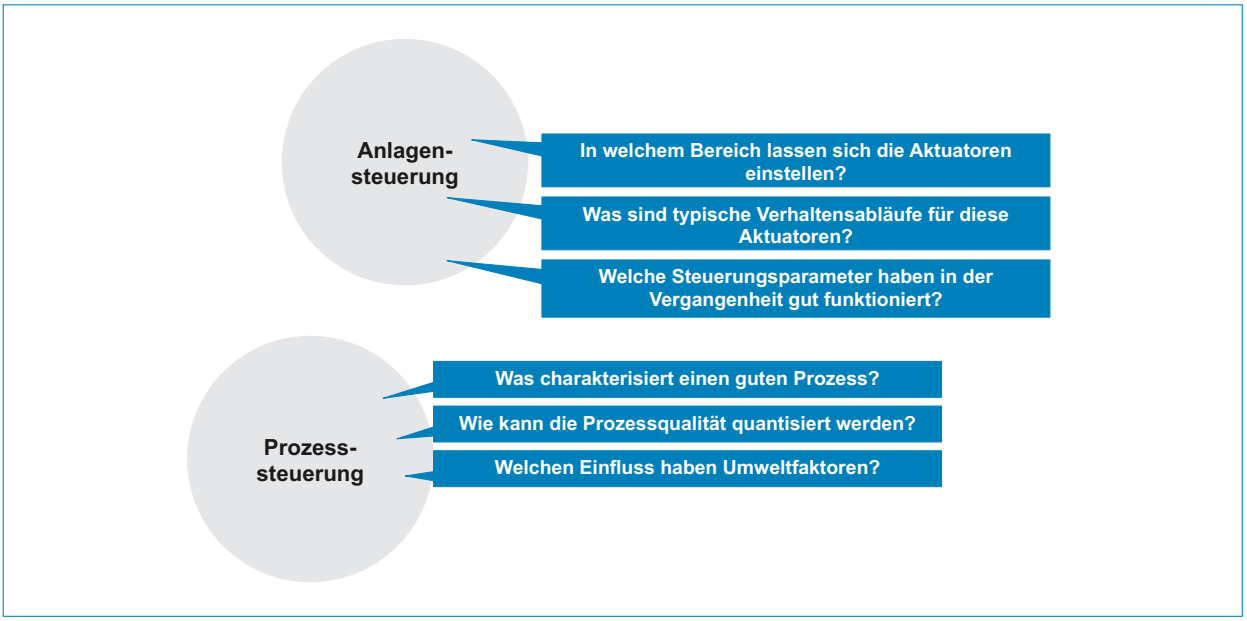


Abbildung 4: Reinforcement Learning bedeutet Perspektivwechsel: von der Anlagen- zur Prozesssteuerung.

Reinforcement Learning für die industrielle Anwendung

Reinforcement Learning ermöglicht es einer Maschine bzw. einer Anlagensteuerung einen komplexen Zusammenhang selbstständig zu erlernen. Dafür muss nicht der gesamte Prozess bekannt sein. Stattdessen wird der Lösungsweg Schritt für Schritt durch Ausprobieren gefunden und optimiert. Im Folgenden werden das Prinzip und die notwendigen Begriffe definiert:

Die formale Darstellung des Prinzips ist in Abbildung 5 dargestellt. Ein Agent wirkt über einen oder mehrere Aktuatoren auf seine Umgebung ein. Anhand der Kostenfunktion wird diese Aktion bewertet. Der Agent erhält als Feedback den neuen Folgezustand und als Bewertung einen Kostenwert auf Basis der Kostenfunktion. Auf dieser Grundlage führt er im nächsten Iterationsschritt erneut eine Aktion aus. Dieses Vorgehen wird solange iteriert, bis ein hinreichend gutes Ergebnis erzielt wurde.

Agent

Der Agent ist ein autonomes Softwareprogramm, das im Reinforcement Learning die Rolle des Entscheiders übernimmt. Er bekommt zu jedem Zeitschritt Informationen über den aktuellen Zustand der Umwelt bzw. des Systems und eine Belohnung für die Ausführung der letzten Aktion. Mithilfe dieses Zustands und der aktuellen Kostenfunktion bestimmt der Agent die Aktion für den nächsten Zeitschritt.

Umgebung

Die Umwelt ist durch das zu steuernde System gegeben. Dies kann z.B. eine Produktionsstraße sein. Dieses System wird charakterisiert durch einen aktuellen Zustand und kann durch Aktionen des Agenten direkt beeinflusst werden.

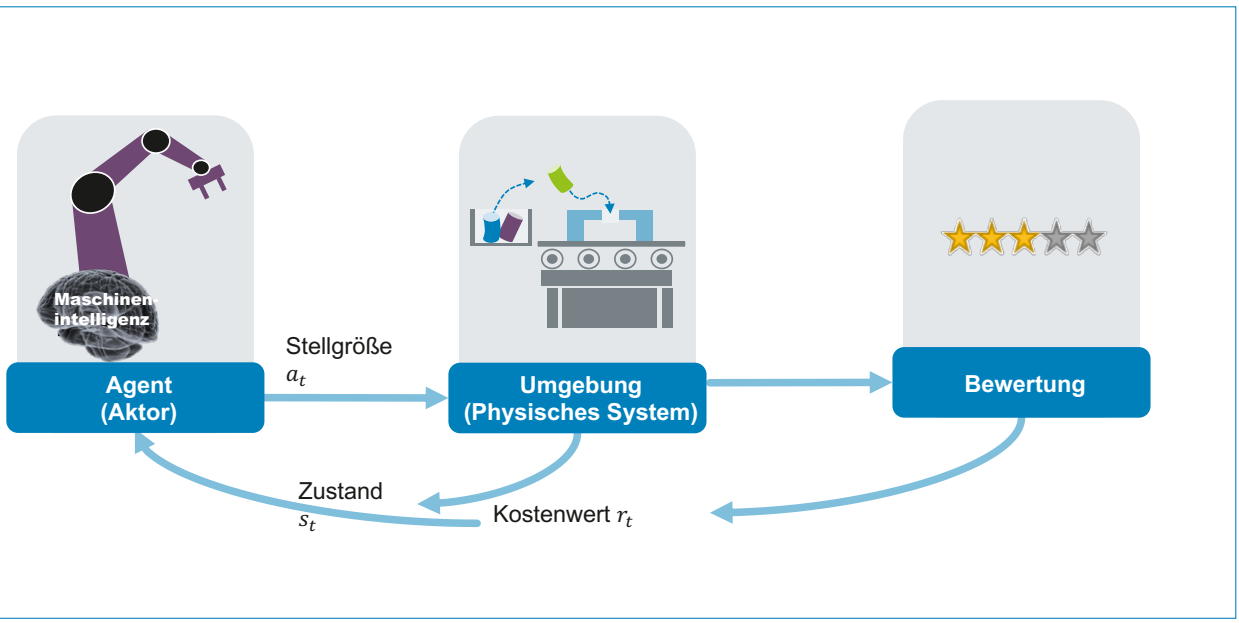


Abbildung 5: Der Reinforcement Learning Zyklus: Der Agent wählt eine Aktion, bzw. bestimmt seine Stellgrößen und wirkt so auf die Umgebung ein. Als Rückmeldung bekommt er einen neuen Zustand und eine Bewertung der durchgeführten Aktion zurück.

Zustand und Zustandsraum

Der Zustand des Systems wird über die verschiedenen Sensorsignale beschrieben. Abhängig vom Prozess kann dies eine Kamera am Endeffektor eines Roboterarms, Temperatur-, Druck-, Strahlungs- oder beliebige andere Sensoren beinhalten.

Der Zustandsraum beschreibt alle Zustände die das System einnehmen kann. Damit ist er unter anderem vom Messbereich der Sensoren abhängig.

Aktion/Aktionsraum

Eine Aktion beinhaltet die Signale für alle einstellbaren Aktuatoren des Systems.

Der Aktionsraum beschreibt dementsprechend den möglichen Einstellungsraum der Aktuatoren.

Kostenfunktion

Durch die Kostenfunktion wird die aktuelle Prozessgüte beschrieben. Mit Kosten sind hier keineswegs wirtschaftliche Kosten gemeint, sondern eine Belohnung oder Bestrafung als Bewertung für eine Aktion des Agenten. Die Kostenfunktion wird in jedem Schritt ausgewertet. Anschließend werden die Aktionen des nächsten Schrittes mithilfe der aktuellen Kostenfunktion ermittelt. Daraus resultiert der Wandel von einer Steuerung der Anlagenparameter zu einer direkten Steuerung des Prozesses. Die Kostenfunktion muss den Prozess direkt betrachten, um eine zielgerichtete Optimierung zu ermöglichen. Die Qualität der Kostenfunktion ist essentiell für den Erfolg der selbstlernenden Steuerung. Über diese Funktion kann eine Loslösung von alten Steuerungsmustern erreicht werden.

Policy

Die Policy beschreibt im Reinforcement Learning die Strategie des Agenten. Diese ist abhängig vom aktuellen Zustand des Systems. So wird eine Strategie gelernt, welche auf verschiedene Zustände optimal reagieren kann. Der Begriff Policy beschreibt im Reinforcement Learning Kontext also eine Art intelligente Steuerungsstrategie.

Trainings- und Wertschöpfungsphase

Beim Reinforcement Learning werden zwei Phasen unterschieden, die Trainingsphase und die Wertschöpfungsphase.

Während des Trainings müssen möglichst viele Parameter des Prozesses bewusst gewählt und kontrolliert werden können. Die Trainingsumgebung ist also im besten Falle steuerbar und das Verhalten des Systems in dieser Umgebung möglichst durch einen Prozessexperten erklärbar. Zusätzlich muss der Prozessexperte möglichst diverse Trainingsdaten generieren. Hierzu sollte er Fragen beantworten wie: Welchen Einfluss haben Umwelteinflüsse wie Temperatur, Luftfeuchtigkeit etc. auf den Prozess? Welche verschiedenen Stoffe werden eventuell produziert/gefördert/verarbeitet? Sind Menschen/andere Maschinen im Prozess mit involviert und haben diese möglicherweise Verhaltensunterschiede? All diese unterschiedlichen Szenarien müssen in den Trainingsdaten abgebildet sein. Ist die Varianz in den Trainingsdaten genügend hoch, lernt der Reinforcement Learning Agent die verschiedenen Szenarien in seiner Strategie abzudecken.

Nachdem Abschluss des Trainings folgt die Wertschöpfungsphase. Man geht in dieser Phase davon aus, dass eine sinnvolle, möglichst optimale Strategie gefunden worden ist, welche nun angewandt werden kann. In der Produktionsumgebung müssen die genauen Bedingungen des Prozesses dann nicht mehr streng kontrolliert werden. Es wird davon ausgegangen, dass die Strategie des Agenten an die unterschiedlichen Prozessbedingungen angepasst ist und diese durch den Zustand des Systems erkannt werden.

In der Wertschöpfungsphase kann eine Reinforcement Learning Strategie auf verschiedene Maschinen, Standorte etc. übertragen werden. Weicht das neue Verhalten allerdings zu sehr vom gelernten Verhalten ab, ist ein eventuelles

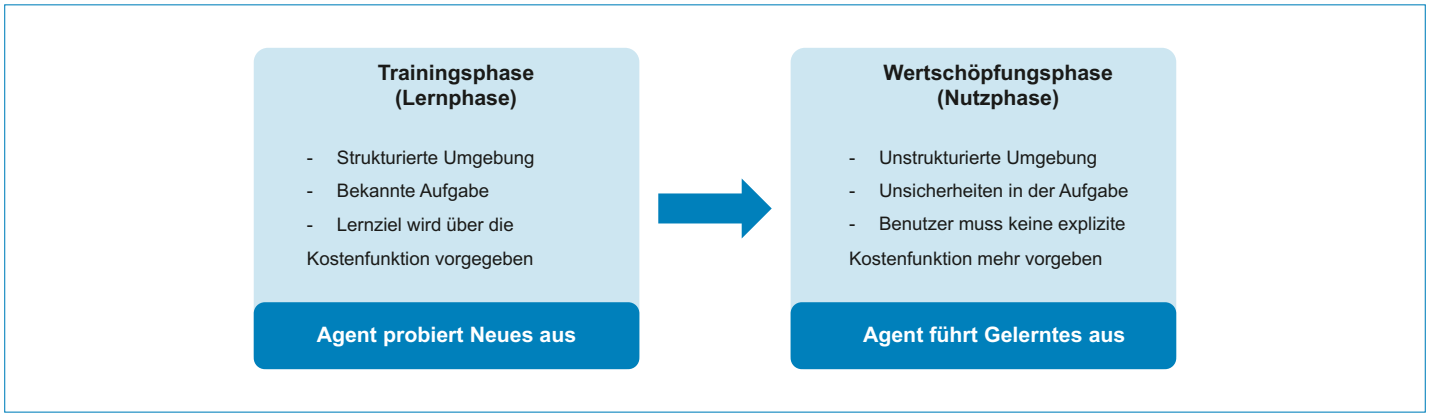


Abbildung 6: Die Anwendung von Reinforcement Learning teilt sich in die Trainings- und die Wertschöpfungsphase.

Nachtrainieren (und damit der Wechsel in eine erneute Trainingsphase) des Prozesses notwendig. Abbildung 6 zeigt die beiden Phasen und ihre Eigenschaften.

Was kann mit Reinforcement Learning gelernt werden? Was nicht?

Reinforcement Learning bedeutet Lernen durch Ausprobieren. Dies hat insbesondere in der industriellen Anwendung weitreichende Konsequenzen auf die möglichen Projekte und die notwendigen Rahmenbedingungen. So ist, wie in Abbildung 7 illustriert, das industrielle Reinforcement Learning durch die besonderen Anforderungen hinsichtlich Robustheit, Sicherheit und Dateneffizienz der Algorithmen gekennzeichnet.

Das Trainieren eines Reinforcement Learning Modells benötigt eine gewisse Menge an Ressourcen. Diese Ressourcen, oft Trainingskosten genannt, sollen möglichst gering gehalten werden. Der wichtigste Kostenfaktor ist der zeitliche Aufwand. Für das Trainieren einer Reinforcement Learning Strategie muss eine große Menge an Daten mithilfe von Testdurchläufen des realen Prozesses generiert werden. Dies erfordert, dass der reale Prozess kurz ist und häufig wiederholt durchgeführt werden kann.

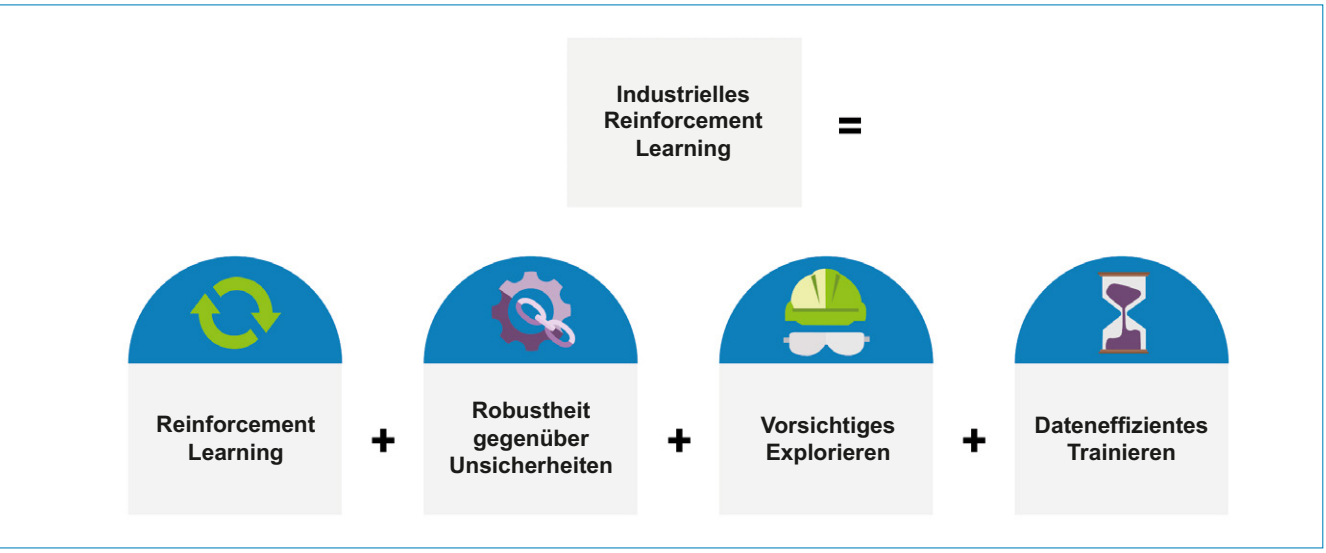


Abbildung 7: Industrielles Reinforcement Learning stellt besondere Anforderungen an Robustheit, Sicherheit und Dateneffizienz.

Zu den Trainingskosten zählen auch die materiellen Ressourcen, also beispielsweise Rohstoffe die ein Prozess verbraucht. Auch diese Kosten sollten so gering wie möglich gehalten werden. Insgesamt stellen diese Trainingskosten also besondere Anforderung an die Dateneffizienz des gewählten Algorithmus.

Alternativ zum realen Prozess können Trainingsdaten auch mithilfe einer Simulation generiert werden. Ist eine aussagekräftige Simulation des Prozesses möglich, erleichtert dies die Generierung der Trainingsdaten. In einer Simulation spielt der zeitliche Aufwand eine untergeordnete Rolle, da Simulationen nicht zwingend in Echtzeit laufen müssen und parallel ausgeführt werden können. Fehlzustände des Prozesses stellen in einer Simulation keine Gefahr dar und Rohstoffkosten können gänzlich vernachlässigt werden.

Außerdem werden während des Trainings auch neue und potentiell instabile Parametereinstellungen ausprobiert. In einer Simulation können diese Parametereinstellungen gefahrlos getestet werden. Die Anwendung in einem realen System erfordert allerdings ein vorsichtiges Ausprobieren der unbekannten Parametereinstellungen um die Sicherheit zu jedem Zeitpunkt zu gewährleisten. Hierfür müssen die realen Systeme möglichst fehlertolerant sein bzw. die Fehler müssen rechtzeitig erkannt und behoben werden können. Außerdem muss schon bei der Wahl des Algorithmus die vorsichtige Exploration der Umgebung beachtet werden.

Die Fehlertoleranz des Systems spielt auch bei der Robustheit gegenüber Unsicherheiten eine große Rolle. Hier ist es allerdings auch wichtig den Reinforcement Learning Ansatz so zu wählen oder zu designen, dass dieser robust auf Unsicherheiten in der Umwelt reagiert.

Leitfragen

Reinforcement Learning ist eine äußerst vielversprechende Möglichkeit zum selbstständigen Erlernen von komplexen Steuerungsstrategien. Die Einführung einer solchen Strategie ist jedoch relativ komplex. Daher wird dazu geraten den Einstieg in diese Methodik anhand eines geeigneten Pilotprojektes zu gestalten. So können die notwendigen personellen und materiellen Kompetenzen langsam aufgebaut und erste Erfolge sichtbar gemacht werden. Mithilfe der folgenden Leitfragen werden wichtige Aspekte zur Findung und Durchführung eines geeigneten Pilotprojektes aufgezeigt. Der im nächsten Kapitel folgende Werkzeugkasten gibt Hilfestellung zur Beantwortung der vorgestellten Leitfragen.

Abbildung 8: Übersicht über die für eine Einführungsstrategie wichtigen Leitfragen bezüglich der Prozessanalyse (grün), der personellen Ressourcen (grau) und der materiellen Ressourcen (lila).

Prozessanalyse – Wie kann das vorliegende System charakterisiert werden?

In dieser Leitfrage wird herausgearbeitet, ob ein Prozess für eine Optimierung mit Reinforcement Learning Ansätzen geeignet ist. Es werden verschiedene Systemkategorien vorgestellt, so zum Beispiel der Unterschied zwischen diskreten und kontinuierlichen Systemen, partiell und vollständig beobachtbaren Systemen, sowie häufigen und seltenen Rückführung von Qualitätsparametern. Die Einordnung eines Prozesses in diese Klassen hilft dabei, den Optimierungsprozess durch Reinforcement Learning besser nachvollziehen zu können.

Welche Zielgrößen sollen in der Kostenfunktion optimiert werden?

Wenn ein möglicher Prozess zur Anwendung gefunden wurde, muss dieser im Anschluss weitergehend untersucht werden. Wichtig ist es hier die zu optimierende Zielgröße zu definieren. Die Zielgröße ist Teil der Kostenfunktion und von entscheidender Bedeutung, da sie die Prozessgüte charakterisiert. Diese Größe muss über Qualitätsparameter messbar und über die vorhandenen Aktuatoren beinflussbar sein. Ein Beispiel für eine solche Zielgröße könnte der Durchfluss in einem Rohr sein.

Was sind die Zustands- und Aktionsräume meines Prozesses?

Für ein Reinforcement Learning Verfahren werden Eingangswerte in Form von Messsignalen und Ausgangswerte in Form von Stellparametern benötigt. Die eingehenden Sensoren beschreiben den Zustandsraum des Systems. Es muss untersucht werden, ob die Sensoren den Zustand des Systems hinreichend genau beschreiben, um das Systemverhalten zu opti-

mieren. Die Ausgangssignale für die vorhandenen Regler beschreiben den Aktionsraum. Auch hier müssen die vorhandenen Aktuatoren genug Handlungsspielraum besitzen, damit ein Optimum gefunden werden kann, das nicht durch die Grenzwerte der Aktuatoren beschränkt wird. Hierzu sollten folgende Fragen beantwortet werden: kann das System kontinuierlich geregelt werden? Wie hoch ist die Genauigkeit des Reglers?

Personelle Ressourcen – Welche unterschiedlichen Kompetenzen sind notwendig?

Nach der Prozessbeschreibung erfolgt nun die Untersuchung der personellen Anforderungen. Welche personellen Ressourcen müssen im Unternehmen vorhanden sein um ein Pilotprojekt zur Anwendung von Reinforcement Learning im jeweiligen Unternehmen durchführen zu können? Hier wird vor allen Dingen auf drei verschiedene Personengruppen eingegangen. Zum einen wird Expertenwissen im Bereich Reinforcement Learning benötigt. Des Weiteren werden interne Experten benötigt, die den zu optimierenden Prozess gut kennen. Hier eignen sich unter anderem langjährige Mitarbeiter, die selbst Erfahrung damit haben, das System händisch zu optimieren. Zuletzt ist ein erfahrener Softwaretechniker unerlässlich. Dieser stellt insbesondere für die Anwendung in der Wertschöpfungsphase eine effiziente und robuste Implementierung zur Verfügung.

Personelle Ressourcen – Wo sollten die Kompetenzen liegen?

Für die drei Personengruppen, Reinforcement Learning Experten, Prozessexperten und Softwaretechniker muss entschieden werden, wo diese Kompetenzen liegen sollen.

Diese Kompetenzen können entweder im Unternehmen selbst liegen, durch externe Dienstleister eingeholt oder durch die Kooperation mit Universitäten akquiriert werden. Es ist auch denkbar, dass sich mehrere KMUs zusammenschließen um dieses Wissen gemeinsam aufzubauen. Hier sollten die folgenden Fragestellungen betrachtet werden: Wie kann Reinforcement Learning langfristig als Kompetenz im Unternehmen aufgebaut werden? Gibt es weitere mögliche Anwendungsfälle im Unternehmen auf die die akquirierten Kompetenzen angewendet werden können? Falls externe Experten hinzugezogen werden, wie kann sichergestellt werden, dass die Systeme auch später betrieben, gewartet und erweitert werden können?

Materielle Ressourcen – Wie kann die Anlage erweitert werden?

In den meisten Anwendungsfällen ist eine besondere Hardware zum Trainieren der Reinforcement Learning Methoden notwendig. Die Anforderungen für diese Hardware müssen spezifiziert werden. Hier geht es um Leistungsanforderungen, die abhängig von der Art und Menge der zu verarbeitenden Daten sind. Außerdem müssen eventuelle Echtzeitanforderungen beachtet werden. Zusätzlich muss die Kommunikation zwischen den schon vorhandenen Schnittstellen und dem neuen Reinforcement Learning Modul unter Beachtung von Echtzeitanforderungen und Ausfallsicherheit betrachtet werden.

Werkzeugkästen zum Klären der Leitfragen

Im Folgenden wird der Hintergrund der voran gestellten Leitfragen erläutert. Mit diesem Wissen soll eingeschätzt werden können, ob ein Prozess für Reinforcement Learning geeignet ist, bzw. welche Anforderungen ergänzt werden müssen, damit ein solches Projekt erfolgreich durchgeführt werden kann.

Prozessanalyse

Vor der Auswahl des Reinforcement Learning Ansatzes und der weiteren Planung des Vorgehens muss der zu optimierende Prozess genau untersucht werden. Hierbei geht es darum, sich über Prozessmerkmale und deren Einfluss auf die spätere Wahl des passenden Reinforcement Learning Konzeptes, bewusst zu werden. An dieser Stelle muss das zuvor beschriebene Umdenken von der Anlagen- zur Prozesssteuerung stattfinden. Mit diesem Verständnis kann der Prozess in der „Reinforcement Learning Denkweise“ beschrieben werden. Abbildung 10 gibt einen Überblick über die wichtigsten Begriffe mit denen sich der zu analysierende Prozess beschreiben lässt.

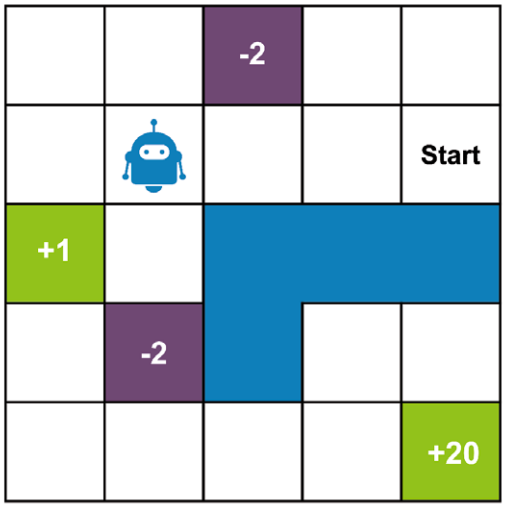


Abbildung 9: diskrete Zustands- und Aktionsräume am Beispiel der so genannten Grid-World.

1. Wie kann das vorliegende System charakterisiert werden?
Im Reinforcement Learning Kontext ist es wichtig, zwischen kontinuierlichen und diskreten Prozessen, bzw. Zustands- und Aktionsräumen zu unterscheiden. Diskrete Zustands- und Aktionsräume lassen sich am Beispiel der sogenannten Grid-World (vgl. Abbildung 9) verdeutlichen.

Hierbei steht ein Agent auf einem Feld eines gitterförmigen Spielfeldes. Der Agent kann sich nicht frei auf dem Spielfeld platzieren, sondern es bleiben ihm (5x5) diskrete Möglichkeiten. In diesem Kontext besteht der Zustand des Agenten also aus seinem aktuellen Ort. Da der aktuelle Ort nur diskrete Zustände annehmen kann, ist auch der Zustandsraum des Agenten diskret. Der Aktionsraum ist ebenfalls diskret, er besteht aus den Aktionen: gehe einen Schritt nach oben, nach rechts, nach links oder nach unten.

Im Gegensatz dazu kann der Zustand eines Roboterarms über seine aktuelle Position, die Geschwindigkeit und die Beschleunigung in allen Gelenken beschrieben werden. Da all diese Werte kontinuierlich gesetzt werden können, handelt es sich hier um einen kontinuierlichen Zustandsraum. Der Aktionsraum eines solchen Roboterarms besteht häufig aus einem Drehmomentsignal für jedes Gelenk und ist dementsprechend auch kontinuierlich.

Eine weitere wichtige Eigenschaft zur Charakterisierung von Prozessen ist die Beobachtbarkeit. Hier unterscheidet man zwischen partiell beobachtbaren Systemen und vollständig beobachtbaren Systemen. In einem partiell beobachtbaren System können einige interne Systemzustände des Prozesses nicht direkt gemessen werden. Ein Beispiel hierfür ist eine Rohrleitung, die punktuell mit Sensoren ausgestattet ist. Dies bedeutet das System ist zwar beobachtbar, aber das Fließverhalten ist nicht an allen Stellen im Rohr bekannt. Als Beispiel für ein vollständig beobachtbares System eignet sich wieder das Beispiel eines Roboterarms. Hier kann mithilfe der

angebrachten Sensoren und der Roboterkinematik der aktuelle Ort und die aktuelle Geschwindigkeit jedes Bauteils genau bestimmt werden.

Um den Prozess weiter zu analysieren sollte als nächstes die Rückführung von Qualitätsparametern untersucht werden. Ist eventuell eine Rückführung nur zu bestimmten Zeiten oder erst nach Abschluss des Prozesses möglich? Ein Beispiel hierfür ist eine Gießmaschine, bei der die Qualität erst nach dem ersten Guss und anschließender Abkühlphase bestimmt werden kann. Geht es im Gegensatz dazu zum Beispiel darum, den Durchfluss durch einen Schüttgutförderer zu überwachen kann das geförderte Produkt mithilfe einer Waage dauerhaft überprüft werden. Hier gibt es natürlich auch Beispiele, in denen eine Rückführung der Qualitätsparameter zwar regelmäßig aber nicht kontinuierlich möglich ist.

Ein weiterer, zu beachtender Punkt ist eine eventuelle Totzeit des Systems. Diese kann abhängig vom jeweiligen Prozess von einigen Sekunden zu einigen Stunden variieren. Diese Zeit hat einen wichtigen Einfluss auf die zeitlichen Trainingskosten des Systems und sollte daher gering sein. Bei einer großen Totzeit ist es umso wichtiger besonders dateneffiziente Reinforcement Learning Methoden zu wählen.

2. Welche Zielgrößen sollen optimiert werden?

Nach einer genaueren Analyse der Prozessmerkmale müssen in einem nächsten Schritt die zu optimierende Zielgrößen bestimmt werden. Alle Zielgrößen zusammen beschreiben die Prozessgüte. Anhand dieser Größen kann bestimmt werden, ob ein Prozess für die aktuelle Aufgabe gut parametrisiert ist. Diese Zielgrößen werden dann in der Kostenfunktion zusammengeführt und mithilfe eines Reinforcement Learning Ansatzes optimiert.

Hierbei ist es wichtig, den genauen Zusammenhang zwischen einer Zielgröße und dem Prozess zu kennen. Ein Reinforcement Learning Ansatz fördert das, was er über die Zielgröße vorgegeben bekommen hat und nicht unbedingt das, was vom Prozessexperten beabsichtigt worden ist. Soll beispielsweise besonders viel Masse mit einem Schüttgutförderer gefördert werden, kann man zunächst annehmen, dass eine hohe Fördergeschwindigkeit des Produktes

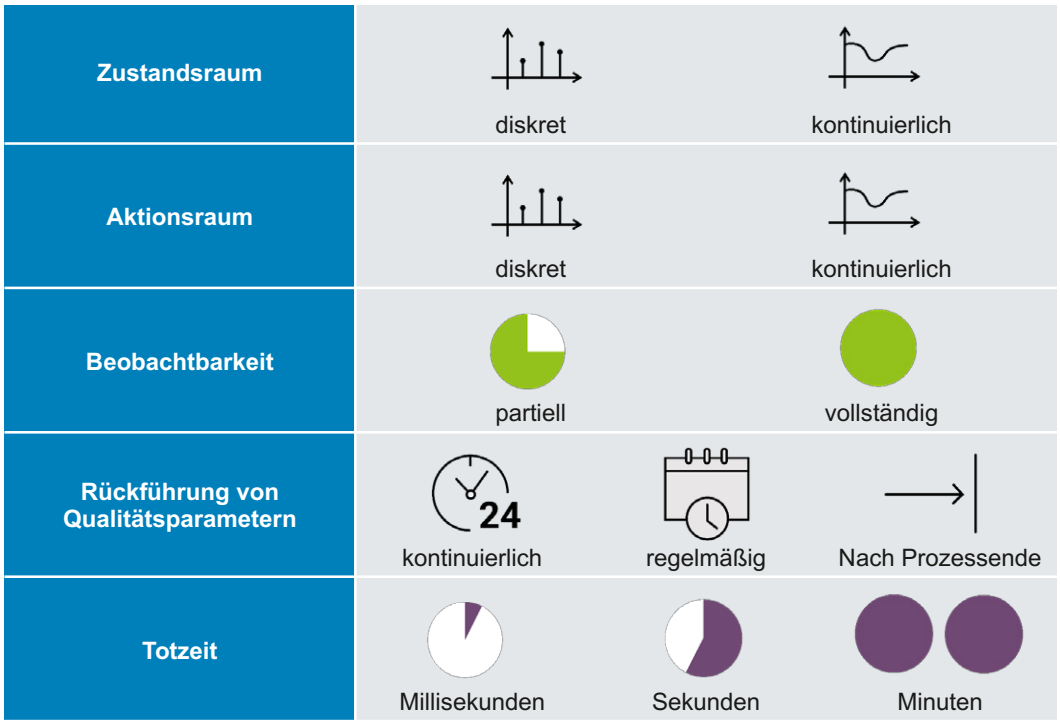


Abbildung 10: Werkzeugkasten zur Analyse des Prozesses.

zu einer großen Fördermasse und dementsprechend zu einem guten Prozess führt. Gibt man allerdings die Fördergeschwindigkeit als Zielgröße vor, kann es dazu kommen, dass nur wenige Partikel des Materials gefördert werden, wenn diese auch eine sehr hohe Geschwindigkeit erreichen. Daher ist es sinnvoller das Gewicht der geförderten Masse am Ausgang des Schüttgutförderers als zu optimierende Zielgröße auszuwählen.

3. Was sind die Zustands- und Aktionsräume meines Prozesses?
Nach der Bestimmung der Zielgröße kann der Zustands- und Aktionsraum festgelegt werden. Hier werden zunächst die vorhandenen Sensoren betrachtetet. Falls diese den Zustandsraum noch nicht vollständig beschreiben, müssen weitere Sensoren angebracht werden. Zusätzlich muss die Art der Sensoren genauer untersucht werden. Hier kann man zwischen Sensoren, die nur einen Datenpunkt aufnehmen, und Sensoren, die einen Vektor von Datenpunkten aufnehmen unterscheiden. Zur ersten Kategorie gehören zum Beispiel Temperaturmesssensoren, die einen einzelnen Temperaturmesswert aufzeichnen. Zur zweiten Kategorie gehören unter anderem Kameras, die gleich eine große Anzahl an Pixeln aufzeichnen. Die Unterscheidung ist wichtig, da diese Eigenschaften die aufgezeichnete Datenmenge und damit die benötigte Rechenkapazität maßgeblich beeinflussen.

Nicht alle Reinforcement Learning Verfahren eignen sich für große Zustandsräume. Von einem großen Zustandsraum spricht man ab einer Größe von circa 12 Dimensionen, also 12 einzelnen Sensorwerten.

Personelle Ressourcen

1. Welche unterschiedlichen Kompetenzen sind notwendig?
Für die erfolgreiche Durchführung eines Pilotprojektes mit Reinforcement Learning in der industriellen Anwendung ist ein breites Spektrum an unterschiedlichen Kompetenzen notwendig. Abbildung 11 gibt einen Überblick über die notwendigen personellen Ressourcen und ihre Kompetenzen.

Zunächst ist ein **Reinforcement Learning Experte** notwendig. Dieser sollte Grundlagen in Mathematik, Regelungstechnik, Optimierung und Statistik besitzen. Außerdem ist ein grundlegendes technisches Verständnis, zugeschnitten auf den jeweiligen Anwendungsfall notwendig. Im Bereich des maschinellen Lernens muss dieser Experte fundierte Kenntnisse, insbesondere in den Bereichen neuronale Netze und Reinforcement Learning, besitzen. Dazu gehören auch Kenntnisse in der Datenvisualisierung und -interpretation. Reinforcement Learning Methoden sind häufig stark abhängig von ihren Hyperparametern. Diese Konfigurationsvariablen definieren die genaue Architektur des neuronalen Netzes und den Trainingsprozess. So gehören beispielsweise die Anzahl der Schichten und Neuronen des neuronalen Netzes aber auch die Anzahl der Trainingsiterationen zu den Hyperparametern. Ein geeigneter Reinforcement Learning Experte sollte bereits Erfahrung im Einstellen dieser Hyperparameter haben. Neben den Kenntnissen des maschinellen Lernens ist für den Reinforcement Learning Experten auch ein mittleres Niveau im Bereich der Programmierung unerlässlich. Auch wenn ein Softwaretechniker hauptsächlich für das Design und die Umsetzung einer guten Softwarearchitektur verantwortlich ist, muss der Reinforcement Learning

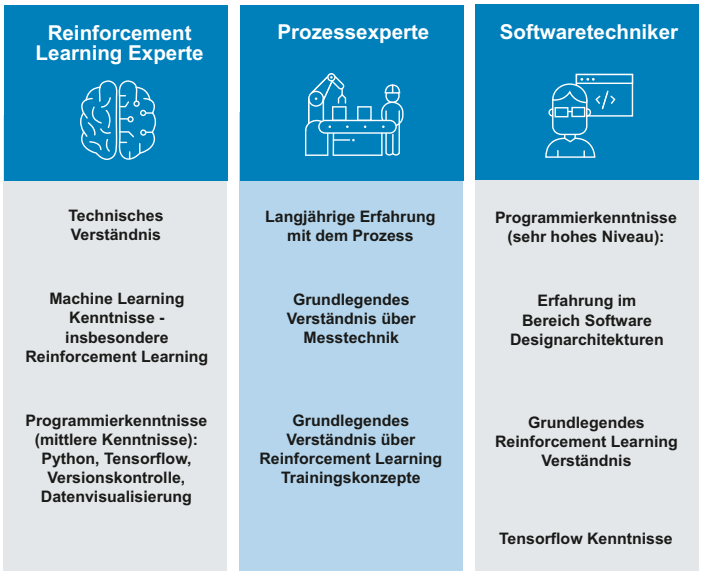


Abbildung 11: Werkzeugkasten zur Beschreibung der personellen Ressourcen.

Experte gute Kenntnisse in diesem Bereich mitbringen, da maschinelles Lernen und die Umsetzung in Programmcode Hand in Hand gehen. Hierzu gehören Kenntnisse in der Versions-Kontrolle in Softwareprojekten. Auch über gängige Software-Frameworks für maschinelles Lernen, wie Tensorflow oder PyTorch, sollte der Reinforcement Learning Experte Bescheid wissen.

Für wartbare und einsetzbare Software ist ein **Softwaretechniker** unerlässlich. Dieser ist für das Design der Softwarearchitektur und die Umsetzung in eine einsetzbare Software zuständig. Daher sind für den Softwaretechniker ein sehr hohes Programmierniveau und fundierte Kenntnisse im Design von Softwarearchitekturen notwendig. Außerdem sollte er gute Kenntnisse im Bereich der Algorithmen und Datenstrukturen besitzen, um effiziente Reinforcement Learning Algorithmen implementieren zu können.

Der **Prozessexperte** besitzt das notwendige Wissen über die Anlage und den Prozess. Prozessexperten haben oft jahrelange Erfahrung mit dem Prozess und kennen sich insbesondere mit den Anlagenparametern gut aus. Dieses Wissen ist notwendig für die Bestimmung des Zustands- und Aktionsraumes, der Findung einer sinnvollen Kostenfunktion und dem Festlegen einer initialen Policy. Hierfür muss sich der Prozessexperte ein grundlegendes Verständnis über das Reinforcement Learning aneignen, um den bekannten Prozess in der neuen Reinforcement Learning Denkweise betrachten zu können. Des Weiteren ist der Prozessexperte auch für die Auswahl geeigneter Sensoren verantwortlich, er benötigt hierfür also Kompetenzen in der Messtechnik.

2. Wo sollten diese Kompetenzen liegen?
Nicht alle Kompetenzen müssen direkt im Unternehmen liegen. Abhängig von der unternehmensinternen Strategie kann es sinnvoll sein einige Kompetenzen an Kooperationspartner abzugeben.

Der **Reinforcement Learning Experte** ist in dem meisten Fällen noch nicht im Unternehmen angesiedelt. Hier ist es sinnvoll sich externe Hilfe in Form einer Kooperation mit einer Universität oder einem externen Dienstleister einzuholen. Sind in Folge des Pilotprojektes weitere

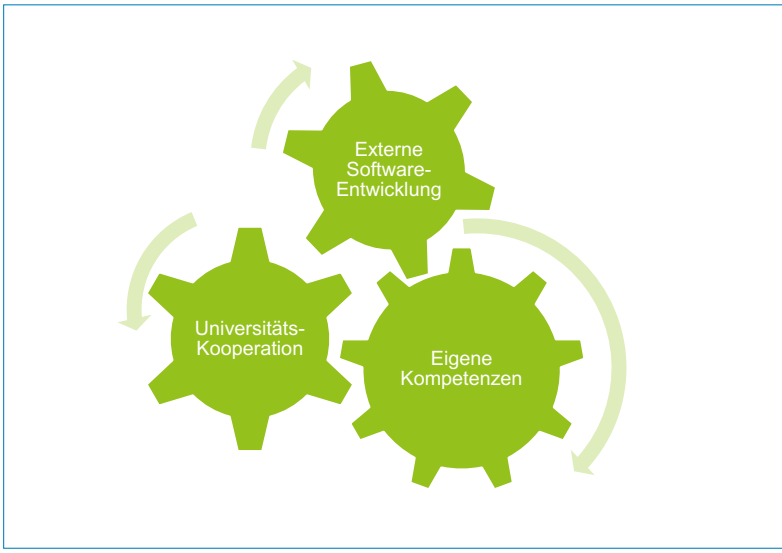


Abbildung 12: Die verschiedenen Kompetenzen können auch durch eine Kooperation mit Universitäten oder Einbindung externer Softwareentwicklung in das Unternehmen geholt werden.

Reinforcement Learning Projekte geplant, kann es sinnvoll sein die Kompetenzen im Unternehmen selbst aufzubauen.

Ist der **Softwaretechniker** schon im Unternehmen vorhanden, sollten möglichst seine Kompetenzen genutzt werden. Ansonsten gibt es mehrere Möglichkeiten diese Kompetenzen einzuholen. Im Falle einer bestehenden Universitätskooperation können sowohl die Kompetenzen im Reinforcement Learning als auch in der Softwaretechnik akquiriert werden. Alternativ gibt es in der Softwaretechnik viele Möglichkeiten externe Dienstleister zu beauftragen.

Die Kompetenzen des **Prozessexperten** liegen im Unternehmen. Nur dort ist das notwendige Wissen über die Anlage, ihre Parameter und ihre Anforderungen zu finden.

Der zeitliche Aufwand der drei oben genannten Personengruppen (Reinforcement Learning Experten, Prozessexperten und Softwaretechniker) zur Umsetzung des Projekts muss geschätzt werden. Hier ist es eventuell von Vorteil externes Expertenwissen hinzuzuziehen. Es wird empfohlen stets die Machbarkeit und den Aufwand von externen Experten abschätzen zu lassen, auch wenn alle drei Expertenkompetenzen im eigenen Unternehmen vorhanden sind.

Abbildung 13: Für die Anwendung von Reinforcement Learning ist spezielle Hardware notwendig. Hier müssen insbesondere die Schnittstellen zwischen dieser Hardware und der Anlagensteuerung definiert werden.

Materielle Ressourcen

Das Trainieren von neuronalen Netzen fordert häufig eine spezielle Hardwarearchitektur. Im Folgenden werden Hilfestellungen zur Wahl dieser Hardware gegeben. Nach einer Auswahl dieser Hardware müssen auch die Schnittstellen zwischen der Hardware und der Anlagensteuerung definiert werden. Diese Schnittstelle muss individuell von Fall zu Fall betrachtet, definiert und implementiert werden.

1. Welche Hardware benötige ich für einen selbstlernenden Prozess?
Bei Heimcomputer werden Rechenoperationen auf dem zentralen Prozessor, der CPU (Central Processing Unit), ausgeführt. Das Trainieren von neuronalen Netzen wird heutzutage immer mehr auf der Grafikkarte, der GPU (Graphics Processing Unit), durchgeführt. Der Unterschied zwischen CPUs und GPUs lässt sich einfach erklären. CPUs sind in der Lage einige wenige, dafür sehr komplexe Berechnungen auszuführen. Die Stärke von GPUs liegt dagegen darin, eine Vielzahl an einfachen Berechnungen parallel auszuführen. Dies bringt einen enormen Geschwindigkeitsvorteil für das Trainieren von neuronalen Netzen und macht sie hierfür besonders gut geeignet. Es gibt noch weitere Prozessormöglichkeiten für das Lernen von neuronalen Netzen. So gibt es zum Beispiel von Google Tensor-Prozessoren, sogenannte TPUs, welche speziell für Anwendungen im Bereich des maschinellen Lernens entwickelt worden sind. Hier ist abhängig vom Anwendungsfall, der Reinforcement Learning Methode und der Datenmenge eine geeignete Architektur auszuwählen. Im Allgemeinen sind GPUs für das Lernen von neuronalen Netzen insbesondere im Kontext von Reinforcement Learning gut geeignet.

2. Welche Anforderungen stellen sich an diese Hardware?
Die Leistungsanforderungen sind abhängig vom jeweiligen Anwendungsfall und lassen sich nicht pauschal aufstellen.

Da die GPU normalerweise den größten Teil der Berechnungen ausführt, sollte insbesondere darauf geachtet werden, eine leistungsstarke GPU zu verwenden. Bei sehr komplexen Berechnungen kann auf die auf das Training von neuronalen Netzen optimierten TPUs (Tensor Processing Unit) zurückgegriffen werden. Es ist aktuell möglich Rechenzeit auf TPUs zu mieten. Außerdem gibt es Architekturen der aktuellsten Generation, welche auch eine kleine Anzahl an TPUs besitzen.

Neben der eigentlichen Leistung von GPUs, ist der auf der Grafikkarte vorhandene Grafikspeicher ein wichtiger Faktor. Dieser beschleunigt zwar nicht direkt die Berechnungen, jedoch gilt je größer der Grafikspeicher, desto mehr Daten können gleichzeitig auf der GPU berücksichtigt werden. Zusätzlich sollte darauf geachtet werden, dass die GPU eine hohe Speicherbandbreite und Taktfrequenz besitzen, da diese maßgeblich zur Datentransparenz zwischen den Speichermodulen beitragen.

Auch wenn die CPU normalerweise nicht die Hauptberechnungen ausführt, übernimmt sie dennoch wichtige Aufgaben im Hintergrund. Sie lädt zum Beispiel die Daten in den Arbeits- und Grafikspeicher. Damit hier kein Bottleneck entsteht, muss eine gewisse Rechenleistung vorhanden sein. Als grobe Richtlinie gilt, dass aktuelle mittel- bis hochklassige Endbenutzer CPUs ausreichend sind.

Algorithmische Ansätze für selbstlernende Produktionsprozesse

Reinforcement Learning Methoden beginnen in der Regel mit einem Datensatz. Dieser Datensatz enthält für jeden Zeitschritt einer Trainingsepisode den aktuellen Zustand, die durchgeführte Aktion und die zugehörigen Kosten. Dieses Datenset kann dann auf verschiedene Weisen genutzt werden, um die vom Algorithmus gelernte intelligente Steuerungsstrategie, die Policy, zu optimieren. Standard Reinforcement Learning Methoden berechnen mindestens eine der folgenden Größen: eine direkte Schätzung der aktuellen Policy, eine Schätzung der sogenannten Value Funktion oder eine Schätzung der Systemdynamik. Im Folgenden werden diese Begriffe und Konzepte eingeführt und im Anschluss einige, auf diesen Konzepten basierende Methoden, erklärt.

Policy Suche

Die Policy Suche, oder auch direkte Policy Optimierung, versucht ausgehend von den gesammelten Daten iterativ eine parametrisierte Policy zu lernen. Hierfür werden die Parameter der Policy iterativ so verändert, dass die Kostenfunktion minimal wird. Diese Vorgehensweise kann als ein numerisches Optimierungsproblem betrachtet werden. Abbildung 14 zeigt schematisch eine direkte Policy Suche. Ein entscheidender Nachteil der direkten Policy Suche ist die Beschränkung der Anzahl der Parameter. Aktuell liefert die direkte Policy Suche nur für Policies mit weniger als 100 Parametern zufriedenstellende Ergebnisse (Deisenroth et al. 2013). Diese limitierte Komplexität führt zu einer beschränkten Komplexität in der zu erlernenden Aufgabe.

Im Bereich der Policy Suche unterscheidet man zwischen Verfahren, die eine Ableitung der Policy verwenden, und solchen, die keine Ableitung benötigen. Erstere Verfahren nennt man im englischen Sprachgebrauch, **Policy Gradient** Methoden und zweitere **Derivative Free Optimization** Methoden.

Value Funktion

Ein zweites Konzept auf dem einige Reinforcement Learning Methoden beruhen ist die sogenannte Value Funktion. Diese Value Funktion ist zu unterscheiden von der Kostenfunktion. Sie gibt zu jedem Zeitschritt eine Schätzung der noch zu erwartenden Kosten bis zum Abschluss der Trainingsepisode an. Diese minimalen, zukünftig zu erwartenden Kosten sind unbekannt, und werden daher mithilfe eines neuronalen Netzes geschätzt. Die geschätzte Value Funktion kann benutzt werden, um für jeden Zustand die optimale Aktion zu bestimmen, d.h. die Aktion, welche die erwarteten Kosten minimiert. Damit die Value Funktion schneller konvergiert, können mathematische Annäherungen verwendet werden,

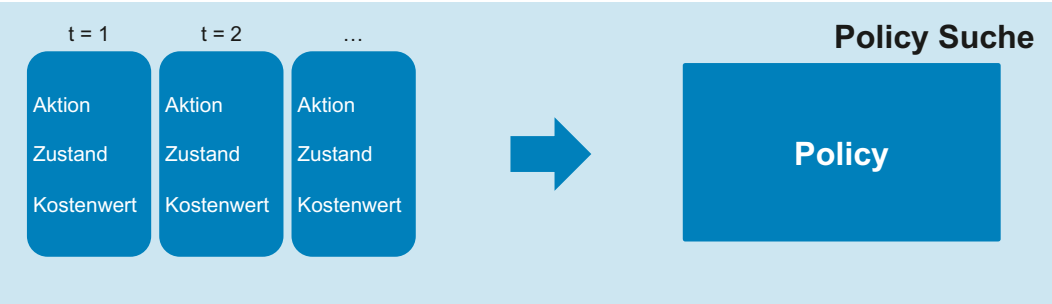


Abbildung 14: Direkte Policy Suche

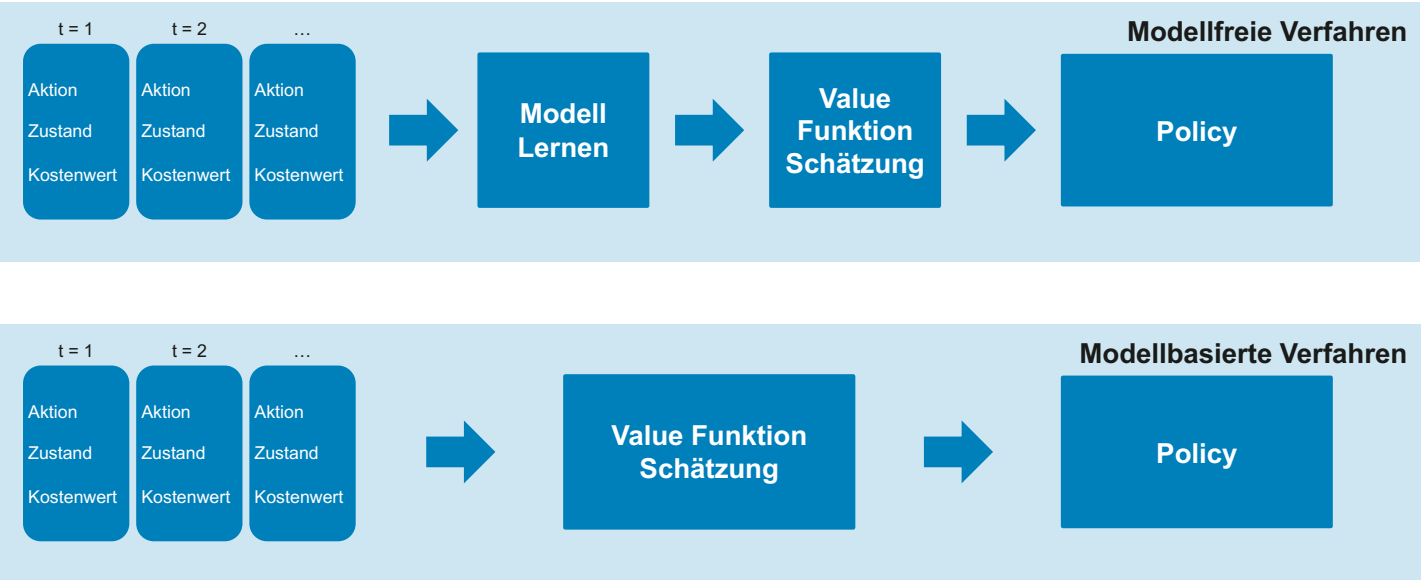


Abbildung 15: Unterschied zwischen modellfreien und modellbasierten Reinforcement Learning Verfahren.

die es erlauben die geschätzten Kosten nach jedem Schritt anzupassen, um eine genauere Value Funktion zu erhalten. Diese Methodik basiert auf dem Prinzip der dynamischen Programmierung.

Modellfreie / Modellbasierte Verfahren

Im Kontext von Reinforcement Learning unterscheidet man zwischen sogenannten modellbasierten und modellfreien Methoden. In Abbildung 15 werden modellbasierte und modellfreie Verfahren schematisch gegenübergestellt.

Das Modell beschreibt die Dynamik des Systems, also wie sich in jedem Zustand des Agenten die gewählte Aktion auf das System auswirkt. In einem kontinuierlichen System kann man dies als Übergangswahrscheinlichkeit von einem Zustand A in einen Zustand B interpretieren, wenn eine bestimmte Aktion ausgeführt wurde. Dieses Übergangsverhalten wird in Abbildung 16 dargestellt. Ein diskretes Anschauungsbeispiel ist Schach. Ein geübter Schachspieler hat Wissen über die mögliche Reaktion seines Gegners gesammelt und kann so den zu erwartenden Spielverlauf vorher im Kopf durchspielen und das für ihn beste Szenario auswählen. Ein modellbasiertes Verfahren wird also zunächst einmal die Systemdynamik lernen. Mithilfe dieses Wissens kann dann die Reaktion des Systems antizipiert werden.

Modellfreie Verfahren kommen ohne eine solche Beschreibung der Dynamik aus. Sie basieren meist auf einer Schätzung der Value Funktion. Dies bedeutet aber auch, dass sie eventuelles vorhandenes Wissen über die Dynamik nicht nutzen können.

Modellbasierte Verfahren sind im Allgemeinen deutlich dateneffizienter als modellfreie Reinforcement Learning Verfahren. Da Dateneffizienz ein entscheidendes Kriterium für den industriellen Einsatz von Reinforcement Learning ist, sind in diesem Kontext modellbasierte Verfahren häufig zu bevorzugen.

Methoden

Die meisten Reinforcement Learning Verfahren basieren auf diesen Prinzipien. Nicht alle lassen sich aber eindeutig einer Kategorie zuordnen, Komplexere Methoden vereinen häufig mehrere dieser Prinzipien. Zwei hochentwickelte Algorithmus-Gruppen sind Aktor-Kritiker Verfahren und Guided Policy Search Verfahren. Beide dieser Ansätze vereinen das Konzept der Policy Suche mit einer Value Funktion.

Aktor-Kritiker Verfahren

Aktor-Kritiker Verfahren gehören zu den modellfreien Verfahren. Sie vereinen das Prinzip von direkter Policy Suche mittels Policy Gradient und einer Value Funktion. Sie bestehen aus zwei Teilen, welche jeweils durch ein neuronales Netz repräsentiert werden. Abbildung 17 zeigt den Aufbau eines solchen Aktor-Kritiker Verfahren.

Die Aufgabe des ersten neuronalen Netzes, dem so genannten Kritiker Netz, ist es, eine Schätzung der Value Funktion zu lernen. Diese kann für jeden Zustand und Zeitpunkt eine Schätzung der noch mindestens notwendigen Kosten bis zum Ende der Episode geben.

Das zweite Netzwerk, auch Aktor Netzwerk genannt, nutzt die aktuelle Schätzung der Value Funktion, um die bestehende Policy durch ein Gradienten Verfahren in Richtung der kleinsten Kosten zu minimieren. In einer Trainingsepisode wird wie üblich zunächst ein Datensatz bestehend aus Aktionen, Zuständen und Kosten für die Trainingsepisode gesammelt. Ausgehend auf diesen Daten wird dann zunächst der Kritiker angepasst. Nachdem der Kritiker eine aktuelle Schätzung der Value Funktion gefunden hat, verbessert der Aktor die Policy, welche er dann an die Umwelt weitergibt.

Guided Policy Search

Guided Policy Search Verfahren gehören zu den modellbasierten Verfahren. In diesen Verfahren wird wie oben beschrieben zunächst eine Beschreibung des dynamischen Übergangsverhaltens gelernt. Hierfür wird eine stochastische Dynamik vorausgesetzt. Das heißt eine in einem bestimmten Zustand ausgeführte Aktion führt nicht immer in den gleichen nächsten Zustand sondern kann mit einer gewissen Wahrscheinlichkeit in verschiedene Zustände führen.

Die Besonderheit bei einem Guided Policy Search Verfahren ist, dass mehrere lokale Lösungen für verschiedene Trainingskonditionen gelernt werden. Solche Trainingskonditionen können beispielsweise unterschiedliche Startpunkte bei einer gelernten Bewegung des Roboterarms oder aber auch unterschiedliche

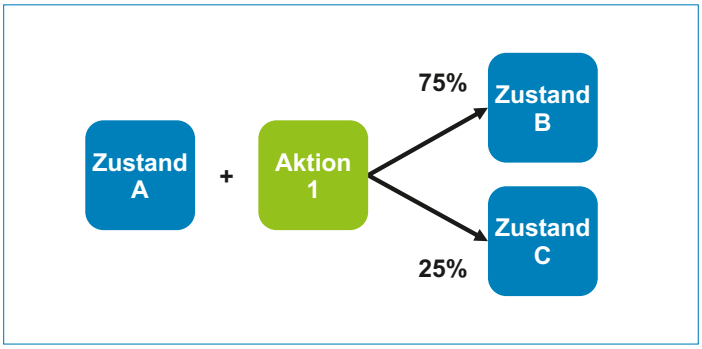


Abbildung 16: Das Dynamikmodell gibt die Übergangswahrscheinlichkeiten von einem Zustand in einen anderen Zustand wieder. Befindet der Agent sich in Zustand A und führt Aktion 1 aus, so landet er zu 75% in Zustand B und zu 25% in Zustand C.

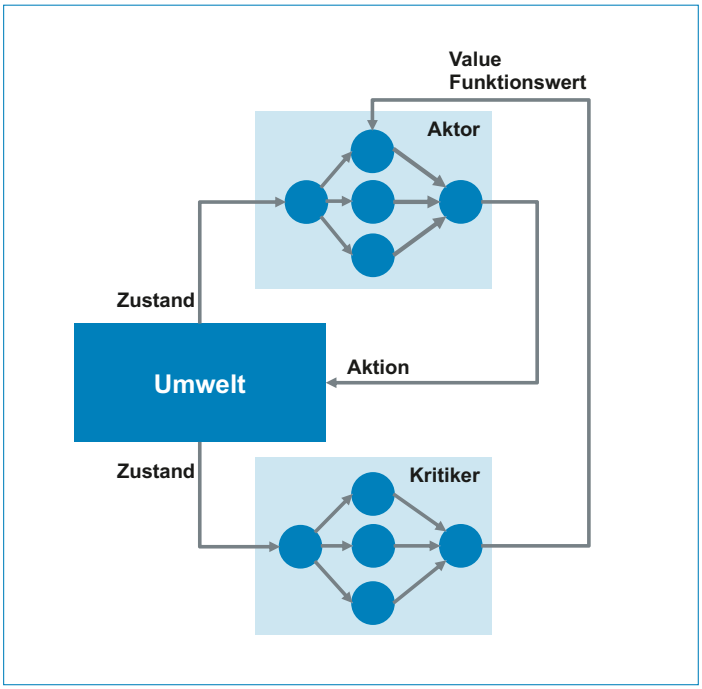


Abbildung 17: Ein Aktor-Kritik Verfahren besteht aus zwei neuronalen Netzen, welche immer abwechselnd trainiert werden.

Klimabedingungen bei Betrieb einer Produktionsanlage sein. Eine Lösung für eine spezifische Trainingskondition wird dann lokale Lösung oder lokale Policy genannt. Das Lernen der Policy für eine Kondition stellt eine deutlich einfachere Aufgabe dar als direkt eine für alle verschiedenen Settings geltende Policy zu trainieren. Nach dem Antrainieren der lokalen Lösungen wird eine globale Lösung mithilfe einer überwachten Trainingsmethode gelernt. Diese globale Policy generalisiert, das stellt auch eine Steuerungsstrategie für die Zustände außerhalb der Trainingskonditionen dar.

Vorgehen zur Integration einer Reinforcement Learning Methodik

Die Anwendungsfälle für Reinforcement Learning in der Industrie sind vielfältig. Das Vorgehen, um eine solche Methodik zu integrieren, folgt jedoch in den meisten Fällen einem klar definierten Schema. Dieses Vorgehen kann in zwei Phasen, die Planungs- und die Realisierungsphase, unterteilt werden. Insgesamt kann die Integration in acht zeitlich aufeinander aufbauende Schritte unterteilt werden. Abbildung 18 zeigt einen Ablaufplan des Integrationsprozesses.

1. Identifikation eines Pilotprojektes
Zunächst einmal muss ein geeignetes Pilotprojekt gefunden werden. Hierbei sind insbesondere die besonderen Anforderungen für die industrielle Anwendung zu betrachten. Dies beinhaltet insbesondere die Robustheit der Algorithmen und des Prozesses, die Sicherheit während des Trainings und die Dateneffizienz des Reinforcement Learning Ansatzes. Hier muss auch das Potential einer Reinforcement Learning Methodik für das Pilotprojekt evaluiert werden.

2. Prozessanalyse
In diesem Schritt wird der Prozess detailliert betrachtet. Dieser Schritt liegt hauptsächlich im Aufgabenbereich des Prozessexperten, da fundiertes Wissen über die Anlage und die Sensoren notwendig ist. Hier können die Leitfragen und der entwickelte Werkzeugkasten zu Hilfe gezogen werden. Es gilt insbesondere den Prozess in der neuen „Reinforcement Learning Denkweise“ zu betrachten. Dafür müssen der Agent und die Umgebung und die zugehörigen Zustands- und Aktionsräume definiert werden. Danach muss das Optimierungsziel definiert werden und die für die Kostenfunktion relevanten Komponenten identifiziert werden. Als nächstes müssen die aktuell vorhandene Sensorik und ihre Qualität betrachtet werden.

3. Wahl des Reinforcement Learning Ansatzes
Nach einer Analyse des Prozesses kann dieses Wissen genutzt werden, um einen geeigneten Reinforcement Learning Ansatz zu wählen. Dies ist die Aufgabe des Reinforcement Learning Experten. Dieser muss die in diesem Leitfaden beschriebenen besonderen Anforderungen der industriellen Anwendung berücksichtigen. Die oben beschriebenen algorithmischen Ansätze können hier einen ersten Ansatzpunkt geben.

4. Abstimmung
Die Prozessanalyse wird hauptsächlich durch den Prozessexperten durchgeführt und die Wahl des Reinforcement Learning Ansatzes liegt im Aufgabenbereich des Reinforcement Learning Experten. Nach diesen beiden Aufgaben müssen sich die beiden Experten erneut abstimmen. Hier müssen insbesondere die für den Reinforcement Learning Ansatz notwendigen Daten und die vorhandene Sensorik und Aktorik betrachtet werden. Die Anforderungen und Voraussetzungen müssen in diesem Schritt solange iterativ angepasst werden, bis eine Übereinstimmung gefunden wurde. Wenn diese gefunden wurde, ist die Planung abgeschlossen und es kann in die Realisierungsphase übergegangen werden.

5. Implementierung
Der erste Schritt in der Realisierungsphase ist die Implementierung. Hierfür müssen der Reinforcement Learning Experte und der Softwaretechniker eingebunden werden. An dieser Stelle kann es sinnvoll sein, noch einen vierten Experten, einen Steuerungstechniker, der sich mit der aktuellen Anlage auskennt, einzubeziehen. Dieser kann zusammen mit dem Softwaretechniker die Schnittstelle zwischen der vorhandenen Anlage und der Reinforcement Learning Hardware definieren und implementieren. Der Reinforcement Learning Experte und der Softwaretechniker setzen dann gemeinsam einen ersten Prototyp der Reinforcement Learning Methodik auf.

6. Prototypische Lernzyklen
Wenn ein erster Softwareentwurf steht, können die ersten prototypischen Lernzyklen gefahren werden. Die Daten aus diesen Zyklen müssen anschließend visualisiert und interpretiert werden. Auf Grundlage der Visualisierung können im weiteren Verlauf die Hyperparameter des neuronalen Netzes angepasst werden. Außerdem können in diesem Schritt eine erste Evaluation und eventuelle Anpassung der Kostenfunktion durchgeführt werden. Nach den prototypischen Lernzyklen sollten erste Erfolge die Wahl des Reinforcement Learning Algorithmus, des Zustands- und Aktionsraumes sowie der Kostenfunktion bestätigen.

7. Codebereinigung
Am Ende der Realisierungsphase steht die Codebereinigung. Hier ist es die Aufgabe des Softwaretechnikers, den im Laufe der Entwicklung entstandenen Programmcode zu vereinfachen und nach Effizienzkriterien zu optimieren. Hier ist es besonders wichtig, auf eine gute Dokumentation zu achten. Diese Dokumentation erleichtert die Wartung und spätere Adaptionen des Programmcodes.

8. Lernzyklen
Nach der Codebereinigung können erneut Lernzyklen mit dem entstandenen Softwaretool gefahren werden. Mit den so generierten Daten können der Prozess- und der Reinforcement Learning Experte gemeinsam die Performanz der gelernten intelligenten Steuerungsstrategie evaluieren. Hier sollte die Effizienzsteigerung im Vergleich zum ursprünglichen Verhalten als Kriterium herangezogen werden. Zusätzlich sollte der Reinforcement Learning Experte das Konvergenzverhalten des Trainings analysieren.

Nach dem eine ausreichende Performanz der Steuerungsstrategie validiert wurde, kann von der Trainingsphase in die Wertschöpfungsphase übergegangen werden. Hierfür muss die Steuerungsstrategie zunächst in ein geeignetes Format gebracht werden. Danach kann diese Strategie – statt in der stark kontrollierten Trainingsumgebung – in der weniger kontrollierten Wertschöpfungsumgebung eingesetzt werden. Auch in dieser Umgebung sollte die Leistung der Steuerungsstrategie erneut evaluiert werden.

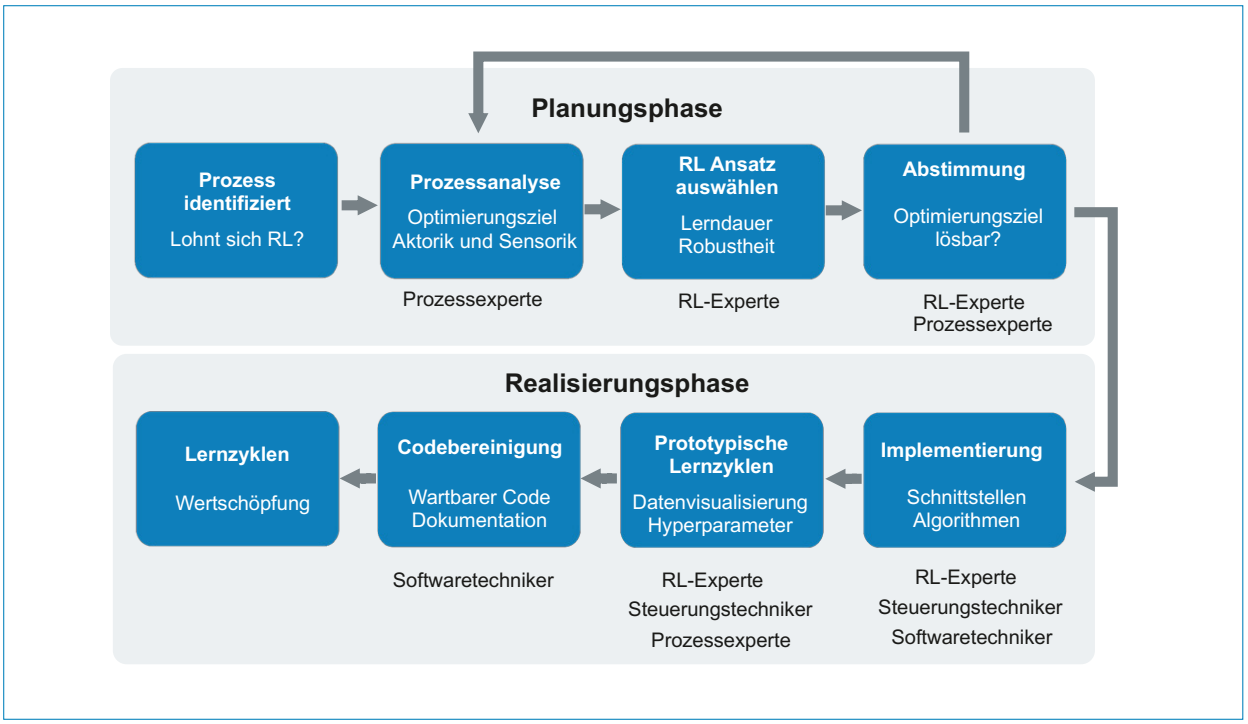


Abbildung 18: Der Prozess zur Integrations- von Reinforcement Learning besteht aus 8 Schritten und trennt sich insbesondere in die Planungs- und in die Realisierungsphase.

Anwendungsbeispiel

Autonomer Montageprozess

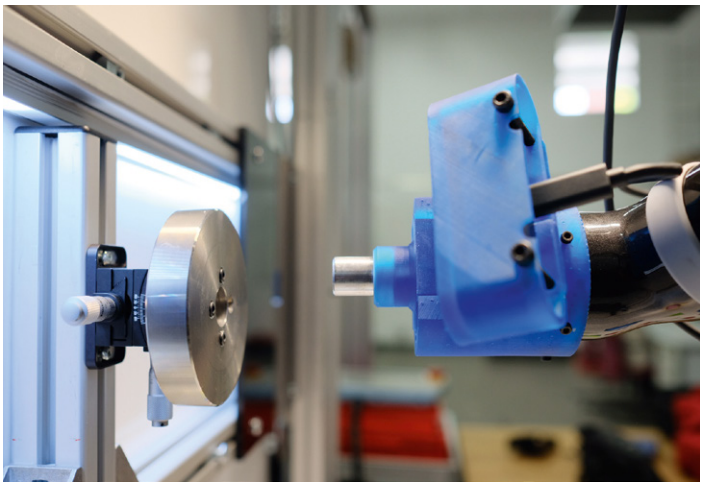


Abbildung 19: Autonomer Montageprozess anhand einer Stift-in-Loch Aufgabe.

Im Rahmen des Projektes InPulS wurde am Institut für Unternehmenskybernetik e.V. ein wissenschaftlicher Demonstrator aufgebaut, welcher mithilfe von Reinforcement Learning einen autonomen, kraftgeregelten Montageprozess erlernt. Hierbei ist das langfristige Ziel eine autonome Montagezelle für unplanbare Montagesituationen zu entwickeln. In dieser Montagezelle erlernt ein Roboter eigenständig eine Montagebewegung, ohne dass eine exakte kinematische und dynamische Beschreibung des Greifsystems oder des Bauteiles notwendig ist. Als Beispiel für diesen Montageprozess wurde eine Fügeaufgabe betrachtet, welche zu den klassischen Evaluationsszenarien in der Robotik gehört. Fügeaufgaben sind sehr kontaktreich und stellen daher eine komplexe Lernaufgabe dar, welche oft nicht in Simulationen erlernbar ist. Diese so genannten Stift-in-Loch Aufgaben erfordern in der Industrie häufig eine höhere Positioniergenauigkeit als mit aktuellen Robotern möglich ist.

Reinforcement Learning Setting

In dem wissenschaftlichen Demonstrator wird mithilfe eines 6-achsigen Roboterarms gelernt, einen zylindrischen Stift in ein passendes Loch zu fügen. Der Zustandsraum des Roboters besteht aus sechs Gelenkwinkelstellungen und sechs zugehörigen Gelenkwinkelgeschwindigkeiten. Der Aktionsraum enthält dementsprechend sechs Drehmomente, eines für jedes Gelenk. Die Aufgabe ist es, den Endeffektor des Roboterarms zu einer vorher definierten Zielkoordinate zu bewegen. Die Kostenfunktion ist dementsprechend über den Abstand zwischen dem Endeffektor und dem Zielpunkt definiert.

Lernprozess

Das Erlernen des Montageprozesses folgt der Analogie eines Kindes, welches lernt, ein Bauklötzchen in ein entsprechendes Loch zu stecken. Zum Erlernen der Bewegung wird eine

Ø Loch (mm)	20	20	20	20	20	20	20	20	20
Ø Stift (mm)	19,9	19,8	19,7	19,6	19,5	19,4	19,3	19,2	19,1
Erfolgsrate	**3/10**	**8/10**	**8/10**	**8/10**	**10/10**	**10/10**	**10/10**	**10/10**	**10/10**

Tabelle 1: Erfolgsrate des autonomen Montageprozesses mit verschiedenen Stiftgrößen.

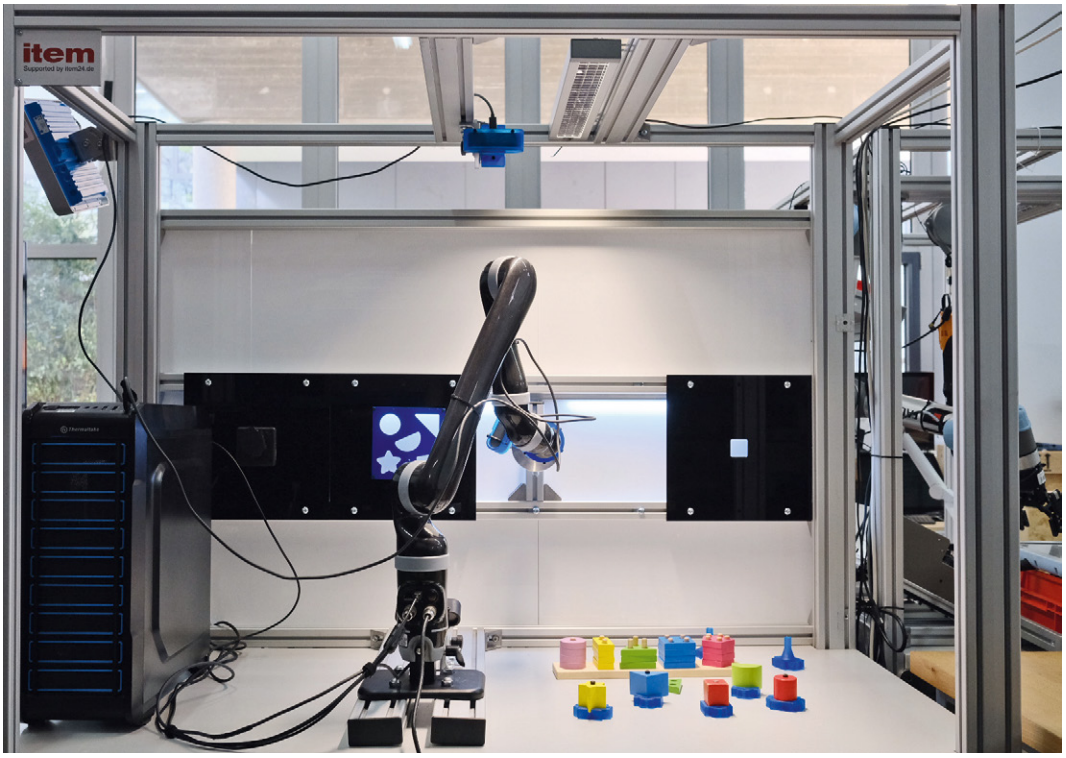


Abbildung 20: Autonomer Fügeprozess mit verschieden geformten Gegenständen.

Methodik basierend auf dem Konzept von Guided Policy Search verwendet. Zunächst werden fünf zufällige Bewegungstrajektorien ausprobiert. Der Begriff Trajektorie beschreibt hier die vom Roboter durchgeführte Bewegung anhand einer Bahnkurve ausgehend von einem festgelegten Startpunkt bis zu einem Zielpunkt. Danach wird anhand der Kostenfunktion die aktuell beste Trajektorie und eine dazu passende Policy bestimmt. Anschließend werden auf Basis der aktuell besten Trajektorie in der nächsten Iteration neue Bewegungen ausprobiert.

Ergebnisse

Das Szenario wurde mit einer Lochgröße von 20mm Durchmesser und verschiedenen Stiftgrößen durchgeführt um die Präzision der Montage zu testen. Die Ergebnisse der Versuchsdurchläufe mit Stiftgrößen zwischen 19,1 und 19,9 mm Durchmessern sind in Tabelle 1 dargestellt. Bei einer Stiftgröße von 19,9 mm Durchmesser fällt die Erfolgsquote deutlich ab. Mit dem aktuellen Verfahren kann also eine Präzision von etwa 2 mm erreicht werden. Es ist zu beachten, dass für dieses Szenario ein Roboterarm mit einer hohen Positionsabweichung benutzt wurde. Ein Roboter mit einer geringeren Ausgangsabweichung ist möglicherweise in der Lage eine noch höhere Präzision zu erreichen.

Anwendungsbeispiel

Selbstlernender Prozess auf einem Schüttgutförderer

Abbildung 21: Pneumatischer Schüttgutförderer im Technikum der AZO GmbH + Co. KG

Das zweite Anwendungsszenario wurde ebenfalls im Rahmen des Projektes InPulS im Technikum der AZO GmbH + Co. KG realisiert. In diesem Szenario wird eine verfahrenstechnische Problemstellung am Beispiel eines pneumatischen Schüttgutförderers adressiert.

Anwendungsszenario

Pneumatische Förderer werden in der Industrie immer da angewendet, wo ein Produkt, genauer gesagt ein Schüttgut, von einem Produktionsort zu einem nächsten transportiert werden muss. Schüttgüter sind beispielsweise Mehl, Sand aber auch Kunststoffpulver. Die Schüttgüter werden mittels eines Gases, typischerweise Luft, gefördert. Hierbei unterscheidet man zwischen der sogenannten Druckförderung und der Saugförderung. Bei der Druckförderung wird der Luftfluss über die Erzeugung eines Drucks am Eingang und bei der Saugförderung durch die Erzeugung eines Vakuums am Ausgang erzeugt (Hilgraf 2019). Im betrachteten Anwendungsfall wurde der Luftfluss über ein Gebläse am Ausgang erzeugt. Die Gebläsedrehzahl und damit der Luftvolumenstrom waren über einen Frequenzumformer regelbar. Der Materialeinfluss am Eingang erfolgt über eine Dosierschnecke, deren Dosierleistung über die Rotationsgeschwindigkeit gesteuert werden kann. Die Anlagengröße von solchen Schüttgutförderern in der Industrie kann von einigen Metern bis zu mehreren tausend Metern reichen (Hilgraf 2019). Die in diesem Anwendungsfall betrachtete Testanlage umfasst eine Förderstrecke von 40 m.

Unsicherheitsbehafteter Prozess

Die Förderung von Schüttgütern ist ein mit einer Vielzahl von Unsicherheiten behafteter Prozess. Zum einen sind die verschiedenen förderbaren Produkte in ihren Eigenschaften stark unterschiedlich. So reichen sie beispielsweise in ihrer Größe von staubfein bis grobkörnig, bzw. von Durchmessern von wenigen Mikrometern bis zu einigen Zentimetern. Außerdem sind bei Naturprodukten, wie beispielsweise Nüssen, häufig geometrische Unterschiede von einer Produktcharge zu nächsten vorhanden. Des Weiteren sind die Material- und Fließeigenschaften vieler Schüttgüter abhängig von Wetterbedingungen wie der Temperatur und der Luftfeuchtigkeit. Das dynamische Prozessverhalten bringt weitere Unsicherheiten mit. So gibt es während der Förderung das Risiko einer abrupten Verstopfung des Rohrs. Dies passiert immer dann, wenn mehr Material eingeleitet wird als über den aktuellen Luftstrom abtransportiert werden kann. Eine solche Verstopfung passiert häufig unerwartet und ist meist nicht automatisch zu beheben. Oftmals kann das Rohr nur durch einen manuellen Eingriff wieder befreit werden.

Aufgrund der Vielzahl der Unsicherheiten und dem hohen Wartungsaufwand im Falle einer Verstopfung, erfolgte bisher meist eine sehr konservative Auslegung des Förderprozesses. So konnte auch bei ungünstigen Bedingungen noch ein robusten Fließprozess gewährleistet werden. Mit dem Einsatz der künstlichen Intelligenz wird nun angestrebt, den Fließprozess

stets im optimalen Betriebspunkt zu halten und somit eine deutliche Effizienzsteigerung hinsichtlich der Förderungsmenge zu erreichen.

Die Ziele einer selbstlernenden Steuerung für den Schüttgutförderer können daher wie folgt zusammengefasst werden:

- Die Steuerungsstrategie muss in der Lage sein sich selbstständig an neue Materialien und Umweltbedingungen anzupassen.
- Die Steuerungsstrategie muss möglichst nah am optimalen Betriebspunkt liegen.
- Die Steuerungsstrategie muss in einer unsicherheitsbehafteten Umgebung ein robustes Verhalten haben.

Reinforcement Learning Setting

Wie in den Leitfragen beschrieben, wurden zunächst einmal der Zustands- und der Aktionsraum definiert.

Der kontinuierliche **Zustandsraum** besteht aus insgesamt 18 Sensoren. Dazu gehören

- 8 Drucksensoren an verschiedenen Stellen im Rohr
- 1 Temperatursensor
- 1 Luftfeuchtigkeitssensor
- 4 Sensoren um die Luft- bzw. Produktgeschwindigkeit an verschiedenen Stellen im Rohr zu messen
- 1 virtuellen Blockagesensor.

Der virtuelle Blockagesensor nimmt eine Verstopfung an, wenn die Luftgeschwindigkeit innerhalb des Rohres null ist.

Der **Aktionsraum** ist ebenfalls kontinuierlich und besteht aus dem Gebläse und der Dosierschnecke am Anlageneingang. Für beide Aktuatoren sind technische Beschränkungen gegeben, welche eine minimale und eine maximale Drehzahl und einen Sicherheitsmechanismus umfassen.

Zur Definition der **Kostenfunktion** musste im nächsten Schritt ein gutes Prozessverhalten beschrieben werden. Dies kann über drei Kriterien definiert werden. Zum einen sollte ein möglichst ruhiges Aktuator-Verhalten angestrebt werden, um unnötigen Verschleiß zu vermeiden. Des Weiteren soll ein möglichst großer Massestrom gefördert werden und Verstopfungen sollen vermieden werden.

Als Methodik wurde ein **modellbasiertes Reinforcement Learning** Verfahren verwendet. Der modellbasierte Ansatz wurde insbesondere aufgrund seiner Dateneffizienz gewählt. Speziell wurde ein Guided Policy Search Verfahren verwendet.

Lernprozess

Der Lernprozess ist ähnlich wie beim autonomen Montageprozess episodisch aufgebaut. Dies bedeutet, es werden drei verschiedene Trainingsläufe gefahren und Zustands-, Aktions- und Kostenwerte aufgezeichnet. Danach wird mithilfe dieser Trainingsdaten eine aktuelle Policy gelernt. Diese Policy wird dann für den nächsten Trainingszyklus verwendet und mit einem Rauschen überlagert ausgeführt. Das Verrauschen der Policy ist insbesondere wichtig, um auch bisher noch unbekannte Zustände zu erreichen. Der Förderprozess des Schüttgutförderers kann in drei Phasen unterteilt werden, der Anlaufphase, der Förderungsphase und der Auslaufphase. Der gesamte Prozess ist 90 Sekunden lang. Da die

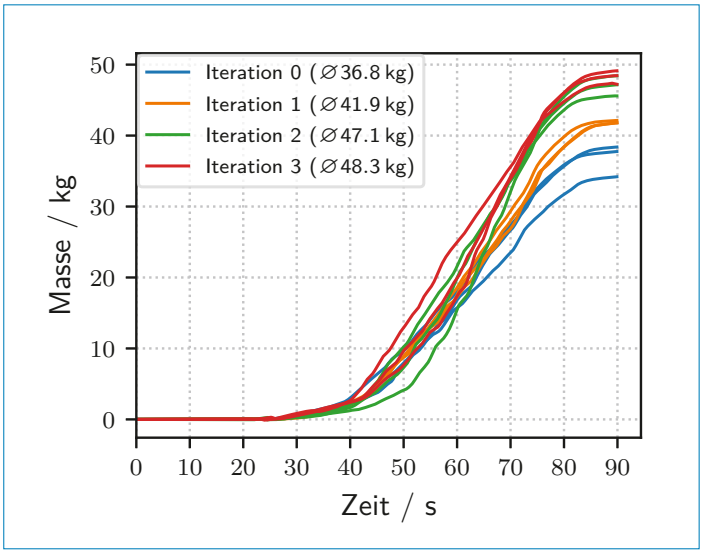


Abbildung 22: Förderung eines Kunststoffgranulats. Die blauen Trajektorien zeigen das Förderverhalten der initial gewählten Policy. Im Laufe des Trainings konnte eine Effizienzsteigerung von 31 % erreicht werden.

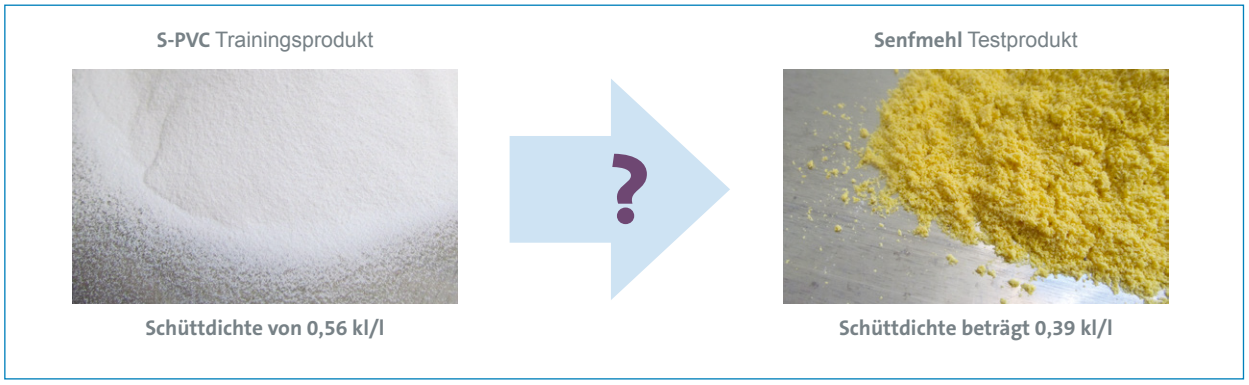


Abbildung 23: Die auf dem Trainingsprodukt erlernte Policy wird nun auf das Testprodukt Senfmehl übertragen.

einzelnen Trainingsläufe immer mit dem gleichen Anfangszustand beginnen müssen, wird nach einem Trainingslauf zunächst das Rohr geleert. Zu Trainingsbeginn wird eine initiale Policy als Ausgangspunkt gewählt. Bei der Wahl dieser Policy steht die Robustheit über der Performanz der Anlage. Dennoch soll ein annehmbares Fließverhalten erreicht worden sein.

Ergebnisse

Zunächst wurde ein Szenario zur Förderung eines Kunststoffpulvers (S-PVC) getestet. Zu Trainingsbeginn wird eine initiale Policy gewählt. Diese ist in Abbildung 22 in blau dargestellt. Danach werden vier Trainingsiterationen durchlaufen. Nach diesem Training wurde ein durchschnittlicher prozentualer Zugewinn des Massedurchflusses von 31 % erreicht. Dieses Ergebnis verdeutlicht das enorme Potential des Reinforcement Learning in der Industrie.

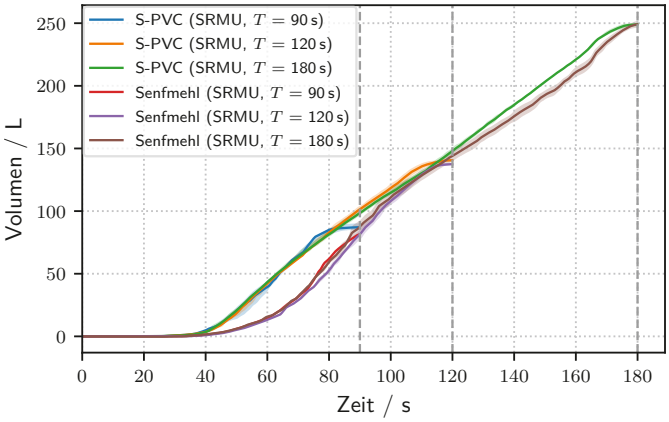


Abbildung 24: Übertragung der Policy für Kunststoffpulver auf Senfmehl. Die blaue Trajektorie zeigt die ursprünglich gelernte Policy für das Kunststoffpulver. Mithilfe dieser Policy wurde dann zunächst über einen längeren Zeitraum von 120 bzw. 180 Sekunden gefördert

Nach dem Erlernen des Prozesses für das das S-PVC Pulver wurde die gelernte Policy auf den Rohstoff Senfmehl übertragen. Senfmehl hat eine andere Schüttdichte und ist im Unterschied zum Kunststoffpulver ein öliges Produkt. Dies führt zu deutlich anderem Fließverhalten. Ein wichtiger Faktor für die pneumatische Förderung und damit insbesondere auch für den maximal möglichen Massendurchfluss, ist die Schüttdichte eines Produktes. Eine geringere Schüttdichte bedeutet einen geringeren möglichen Massendurchfluss. Daher muss der Faktor der Schüttdichte aus der Performanz herausgerechnet werden. Eliminiert man diesen Faktor, kann die Policy für das S-PVC Pulver, auf die Förderung von Senfmehl angewendet, ein ähnlich gutes Förderungsverhalten wie im ursprünglichen Trainingsprozess erlangen. So werden, wie in Abbildung 24 dargestellt, zwischen 75 und 100 % der ursprünglichen Förderleistung erreicht. Des Weiteren wurde für die Übertragung die Förderzeit von den ursprünglichen 90 Sekunden auf 180 Sekunden erweitert. Auch hier bleibt die Effizienz weiter erhalten. Dies zeigt, dass die gelernte Policy auch außerhalb des Trainingsbereiches zu einer robusten und effizienten Steuerung des Schüttgutförderers führt.

Insgesamt konnte mit der Anwendung auf dem Schüttgutförderer das Potential des Reinforcement Learning komplexe und effiziente Steuerungsstrategien zu erlernen gezeigt werden. Diese Steuerungsstrategie generalisiert auf neue unbekannte Schüttgüter und ist in der Lage robust auf Temperatur- und Luftfeuchtigkeitsschwankungen zu reagieren.

Fazit und Ausblick

Die Anwendung auf dem Schüttgutförderer hat das enorme Potential des industriellen Reinforcement Learning bestätigt. Mithilfe von Reinforcement Learning können Prozesse gesteuert werden, für die eine Modellierung mit konventionellen Methoden zu komplex wäre. Hier lässt sich zum einen ein Effizienzgewinn bezüglich konventioneller Steuerungen erreichen und zum anderen entfällt oft aufwendiges manuelles Einstellen der Anlagenparameter. Allerdings zeigen die Erfahrungen auch, dass eine solche selbstlernende Steuerung oft entweder einen großen Effizienzzuwachs bietet oder vollständig fehlschlägt. So ist Reinforcement Learning kein simpler Einstieg in das Themenfeld der künstlichen Intelligenz.

Das Potential des Reinforcement Learning als ein Teil des Machine Learning wird heutzutage nur langsam entdeckt. So wird dieses Themenfeld an den Universitäten noch nicht ausreichend gelehrt und es besteht noch weiterer Forschungsbedarf. Für die Anwendung auf dem Schüttgutförderer ist eine besonders robuste Variante eines Guided Policy Search Algorithmus entwickelt worden. Es bleibt die Frage, wie sich ein solcher Algorithmus auf weitere Szenarien übertragen lässt. Hier ist weitergehende anwendungsorientierte Grundlagenforschung notwendig um neue Anwendungsgebiete und entsprechende Algorithmen für Reinforcement Learning zu entwickeln. Bei der Entwicklung dieser Algorithmen müssen insbesondere die besonderen Anforderungen der Industrie hinsichtlich Sicherheit und Robustheit erfüllt werden.

In der Industrie wurde in den letzten Jahren große Erwartungen an Big Data gesetzt. Heutzutage gibt es einen Trend weg von Big Data hin zu Smart Data. So gibt es entgegen der weitverbreiteten Einschätzung, dass in der Industrie große Datenmengen zur Verfügung stehen, immer noch viele Szenarien, beispielsweise aus dem Bereich des Sondermaschinenbaus, in denen keine umfassende Datengewinnung möglich ist. Für diese Anwendungen sind besonders dateneffiziente Verfahren notwendig. Reinforcement Learning passt in diesen Trend, da ausschließlich ausgewählte, auf den spezifischen Anwendungsfall zugeschnittene Daten erfasst werden müssen.

Ein weiterer Trend geht von der zentralen Datenverarbeitung in der Cloud zu einer zunehmend dezentralen Verarbeitung. Dies bringt zum einen Vorteile im Bereich Sicherheit und Datenschutz mit sich. Zum anderen bringt die dezentrale Datenverarbeitung auch Vorteile um vorhandenen Echtzeitanforderungen gerecht zu werden. Heutzutage gibt es erste Industrie PCs, die einen integrierten Machine Learning Chip mit sich bringen, auf dem bereits trainierte Modelle berechnet werden können. Hier ist der Fokus meist auf Bilderkennungsmodellen mit direkter Kameraschnittstelle, in Zukunft sind aber auch spezielle Reinforcement Learning Module denkbar.

Projektpartner / Impressum

Herausgeber
VDMA
Forum Industrie 4.0
Lyoner Straße 18
60528 Frankfurt am Main
Telefon +49 69 6603-1810
E-Mail industrie40@vdma.org
Internet industrie40.vdma.org

FKM Forschungskuratorium
Maschinenbau e.V.
Lyoner Straße 18
60528 Frankfurt am Main
Telefon +49 69 66 03-16 81
E-Mail info@fkm-net.de
Internet www.fkm-net.de

Institut für Unternehmenskybernetik e.V.
Dennewartstraße 27
52068 Aachen
Internet www.ifu.rwth-aachen.de/de

Projektleitung
VDMA- Forum Industrie 4.0, Judith Binzer

Inhaltliche Beiträge
Institut für Unternehmenskybernetik e.V.
Philipp Ennen
Emma Pabich
Robin Kupper
Dr. Pia Benmoussa
Dr. René Vossen
Kontakt pia.benmoussa@ifu.rwth-aachen.de

Beteiligte VDMA-Mitglieder aus der InPuIS-Arbeitsgruppe:
AZO GmbH + Co. KG
Festo AG & Co. KG
FIBRO GmbH
Hans Weber Maschinenfabrik GmbH
Karl Mayer Textilmaschinenfabrik GmbH
Geschäftsbereich Komponentenfertigung
Lenze SE
MAHLE Behr GmbH & Co. KG
Oskar Frech GmbH + Co. KG
Schaeffler Technologies AG & Co. KG
SchuF-Armaturen und Apparatebau GmbH
SMC Deutschland
TE Connectivity Germany GmbH a
TE Connectivity Ltd. Company
THEEGARTEN-PACTEC GmbH & Co.KG
Voith GmbH & Co. KGaA
Volkswagen AG
Weidmüller Interface GmbH & Co. KG
ZIMMER GmbH

Design und Layout
VDMA DesignStudio / VDMA Verlag GmbH

Erscheinungsjahr
2019

Druck
h. reuffurth gmbh, Mühlheim am Main

Bildnachweise
Titelbild: iStock / Olivier Le Moal
Seite 1: VDMA
Seite 3: Institut für Unternehmenskybernetik e.V.
Seite 26, 27: Institut für Unternehmenskybernetik e.V.
Seite 28, 30: AZO GmbH + Co. KG

Grafiken
Institut für Unternehmenskybernetik e.V.

Literaturverzeichnis

Deisenroth, Marc Peter; Neumann, Gerhard; Peters, Jan (2013): A survey on policy search for robotics. In: Foundations and Trends® in Robotics 2 (1–2), S. 1–142.

Hilgraf, Peter (2019): Grundlagen der pneumatischen Förderung. In: Peter Hilgraf (Hg.): Pneumatische Förderung. Grundlagen, Auslegung und Betrieb von Anlagen, Bd. 40. Berlin, Heidelberg: Springer Berlin Heidelberg, S. 109–232.

Sadeghi, Fereshteh; Levine, Sergey (2016): CAD2RL: Real Single-Image Flight without a Single Real Image. Online verfügbar unter http://arxiv.org/pdf/1611.04201v4.

Schoettler, Gerrit; Nair, Ashvin; Luo, Jianlan; Bahl, Shikhar; Ojea, Juan Aparicio; Solowjow, Eugen; Levine, Sergey (2019): Deep Reinforcement Learning for Industrial Insertion Tasks with Visual Inputs and Natural Rewards. Online verfügbar unter http://arxiv.org/pdf/1906.05841v1.

VDMA Software und Digitalisierung: Quick Guide. Machine Learning im Maschinen- und Anlagenbau 2018.

VDMA
Forum Industrie 4.0
Lyoner Straße 18
60528 Frankfurt am Main
Telefon +49 69 6603-1810
E-Mail industrie40@vdma.org
Internet industrie40.vdma.org

FKM Forschungskuratorium Maschinenbau e.V.
Lyoner Straße 18
60528 Frankfurt am Main
Telefon +49 69 6603-1681
E-Mail info@fkm-net.de
Internet www.fkm-net.de

Institut für Unternehmenskybernetik e.V.
Dennewartstr. 27
52068 Aachen
Internet www.ifu.rwth-aachen.de/de

industrie40.vdma.org