Achtung! Hunde auf der Fahrbahn
Reinforcement Learning und die Modellierung autonomer Agenten

Dawid Kasprowicz

1. Einleitung

Das Machine Learning (ML) hat für die Programmierung autonomer Autos einen zentralen Stellenwert. Um den etwaigen Unvorhersehbarkeiten des Verkehrs durch die zeitkritische Exekution von Fahrmanövern begegnen zu können, werden neuronale Netzwerke als Bestandteile autonomer oder teilautomatisierter Fahrzeugsysteme trainiert.[1] Das Trainingsmaterial stellen Datenbanken mit computersimulierten oder photographischen Bildern dar, die die so genannten *Training Data* bilden und die in den iterativen Durchläufen als Material für eine optimale Kategorisierung von Verkehrsteilnehmer:innen, -situationen und -umgebungen dienen.[2] Auf dieser Grundlage einer Situationserkennung während des Fahrens, für deren Umsetzung auch entsprechende Abstands-, Kontakt und Radarsensoren notwendig sind, werden vom

1 Mnih, Volodymyr/Badia, Adria P./Mirza, Mehdi et al.: »Human-level control through deep reinforcement learning«, in: Nature 518 (2015), S. 529-533; Pan, Xinlei/You, Yurong/Wang, Ziyan et al.: »Virtual to Real Reinforcement Learning for Autonomous Driving«, in: Proceedings of the British Machine Vision Conference (BMVC) (2017); Shalev-Shwartz, Shai/Shaked, Shammah/Shashua, Amnon: »Safe, Multi-Agent, Reinforcement Learning for Autonomous Driving«, in: Proceedings of the British Machine Vision Conference (BMVC) (2017). Im Verlauf dieses Textes wird allgemein von autonomen Autos gesprochen, wenngleich dies auch das vollautomatisierte Fahren ab Stufe vier beinhaltet, bei dem die/der Fahrer:in noch hinter dem Lenkrad die Fahrt überwachen und in einigen Situationen eingreifen soll. Zu den einzelnen Stufen dieser Gliederung des automatisierten und autonomen Fahrens siehe die Einleitung zu diesem Band.

2 X. Pan et al.: »Virtual to Real«.

neuronalen Netzwerk mögliche Handlungsoptionen oder -anweisungen generiert.

Es liegt also nahe, dass einerseits viele Hoffnungen der Hersteller auf den ML-Verfahren ruhen und dass andererseits diesen Verfahren eine tragende Bedeutung für die gewünschten zeitkritischen Situationserkennungen und die daraus folgenden Aktionen zukommt. Im Kontext von autonomen Autos erhält besonders ein Modell des ML, das so genannte Reinforcement Learning (RL), sowohl innerhalb der Fachcommunity als auch in der Öffentlichkeit eine verstärkte Aufmerksamkeit. Dieser Umstand ist insofern bemerkenswert, als das RL auf den ersten Blick wenig mit Verkehrswelten zu tun hat. In seiner Grundform geht das RL lediglich von einer/m Agent:in-Umwelt-Verhältnis aus, bei dem die/der Agent:in die unbekannte Umwelt erkunden muss (*exploring*), um anschließend aufgrund ihrer gesammelten Informationen eine adäquate Entscheidung treffen zu können (*exploiting*). Dabei strebt die/der Agent:in jeweils ein Ziel an, für das sie/er eine Handlungsstrategie entwickelt (*policy*). Sie/er lernt folglich, zwischen Stimuli zu unterscheiden, auf die ein positives Feedback (*reward*) oder ein negatives Feedback (*punishment*) erfolgen könnte. Im Bestreben, ihr Ziel zu erreichen, erlernt die/der Agent:in durch sukzessives positives Feedback und durch die Vermeidung wiederkehrenden negativen Feedbacks eine optimale Handlungsstrategie.

Ein Terrain, auf dem sich das RL-Modell zunächst für das ML bewährt hat, sind Computerspiele. Neuronale Netzwerke werden dabei mit Bildern von Millionen Spielsituationen trainiert, indem sie lernen, die Spielsituationen sukzessiv zu kategorisieren, zu vergleichen und ihre Handlungsstrategie während des Spiels so auszurichten, dass sie eine optimale Punktzahl erzielen können. Als eine Forschergruppe in einem *Nature*-Artikel 2015 von der Fähigkeit ihrer Künstlichen Intelligenz berichtete, durch »Deep Reinforcement Learning« in einigen Computerspielen besser als ihre menschlichen Gegner abzuschneiden,[3] war dies für viele Beobachter:innen mehr als einer der zahlreichen »Man vs. Machine«-Vergleiche. Seitdem wird das RL auch für autonome Autos als »powerful learning framework« betrachtet, das einen wesentlichen Beitrag zum langjährigen Problem leisten könnte, wie man mit »high dimensional environments« interagiert.[4]

3 V. Minh et al.: »Human-level Control«.

4 Ravi Kiran, B./Sobh, Ibrahim/Talpaert, Victor et al.: »Deep Refinforcement Learning for Autonomous Driving: A Survey«, arXiv:2002.00444v2 [cs.LG], 23. Januar 2021, S. 1.

Es stellt sich aber aus wissenschaftsphilosophischer Sicht die Frage, wie sich solche Transitionen von Computerspielwelten in Verkehrswelten beschreiben lassen. Sind sie allein durch die neue Leistungsfähigkeit neuronaler Netzwerke zu erklären, deren stochastische Situationserkennungen noch schneller und präziser funktionieren? Oder, und das ist eine aktuell verbreitete Antwort, bezeugt die Anwendung von RL-Modellen für autonome Autos das Phänomen einer »Gamification«, also der regelgeleiteten Verschaltung von individuellem Verhalten (sei es von Menschen oder Maschinen) und sozialem Leben unter der Prämisse von positivem oder negativem Feedback aus Entscheidungssituationen?[5] Der Ansatz der »Gamification« als Erklärung für die Virulenz des RL erscheint auch deshalb plausibel, weil es sich bei ihm nicht ausschließlich um Computerspielwelten dreht, sondern um die Abwanderung der Episteme aus dem Spiel in andere Anwendungskontexte.[6] Nicht wenige Medien- und Kulturwissenschaftler:innen teilen daher auch die Ansicht, dass die Interaktionen mit der neuen KI primär unter einer Epistemologie des Spiels betrachtet werden sollten.[7]

Wenngleich die Verzahnung von RL-Modellen mit der Gamification naheliegt, zumal es sich in den Simulationen der Modelle häufig um eine geschlossene Welt handelt, die die/der Agent:in erkunden muss, so soll auf den folgenden Seiten ein anderer Weg eingeschlagen werden. Entgegen der Annahme, dass die neue KI im ML eine Variante der »Gamification« sei, möchte ich das RL als vereinfachtes Verfahren zur Steuerung von Kontingenz betrachten. Im Kern dieses Verfahrens der Kontingenzsteuerung steht die Beziehung zwischen der/dem individuellen Agent:in und ihrer/seiner Umwelt. Dieser Modellkern ist aber nicht erst durch Anwendungen auf Computerspiele entstanden, sondern er entwickelt sich im Laufe einer physiologischen

5 Schrape, Niklas: »Gamification and Governmentality«, in: Mathias Fuchs/Sonia Fizek/Paolo Ruffino et al. (Hg.): Rethinking Gamification, Lüneburg: Meson Press 2014, S. 21-46. Mühlhoff, Rainer: »Human-aided artificial intelligence: Or, how to run large computation in human brains? Toward a media sociology of machine learning«, in: New Media & Society 22/10 (2020), S. 1868-1884, hier S. 1874f.

6 Fuchs, Mathias/Fizek, Sonia/Ruffino, Paolo et al.: »Introduction«, in: Dies. (Hg.): Rethinking Gamification. Lüneburg: meson press 2014, S. 7-20; zu Anwendungen der Gamification siehe Raczkowski, Felix: »Virtuelle Produktivität«, in: Dawid Kasprowicz/Stefan Rieger (Hg.): Handbuch Virtualität, Wiesbaden: Springer 2020, S. 111-128.

7 Gramelsberger, Gabriele/Rautzenberg, Markus/Wiemer, Serjoscha et al.:»›Mind the Game!‹ Die Exteriorisierung des Geistes ins Spiel gebracht«, in: Zeitschrift für Medienwissenschaft 21 (2019), S. 29-38.

und psychologischen Kultur der Verhaltensexperimente, die sich in der ersten Hälfte des 20. Jahrhunderts etablierten und aus denen eine Episteme hervorging, die in den Computermodellen des ML adaptiert und modifiziert wurde. Die Attraktivität des RL für die Software autonomer Autos liegt somit nicht in einer Spielelogik, sondern in dem Wissen, die Kontingenz des Verhaltens durch möglichst simple Umweltmodelle experimentaltechnisch steuerbar zu halten. Mit »simpel« ist demnach keine Wertung gemeint, sondern die sukzessive Steigerung von Kontingenz im Experiment, ohne die Zuschreibung eines Stimulus zu einer bestimmten Aktion aufgeben zu müssen. Zugleich ermöglicht es diese induktive Logik, dass Maschinen, Menschen und andere Lebewesen zu Variablen eines Wissens werden, das primär aus dem Vergleich solcher Verhaltensexperimente entsteht. RL-Modelle ermöglichen damit eine unabdingbare Episteme des Komparativs von Lebewesen und Maschinen in zeitlich diskretisierten Schritten.[8] Umso wichtiger ist es, medien- und wissenschaftshistorisch die Techniken des Arrangements von Lebewesen in modellierten Umwelten darzulegen, da erst hierdurch Formen der Explikation und der Prädiktion eines Lernverhaltens entstehen können. Solche medien- und wissenschaftshistorischen Problemgenesen von ML-Modellen haben bisher einen Seltenheitswert. Sie werden allerdings spätestens dann relevant, wenn die ML-Modelle zu skalierbaren Optionen werden, die von Spiel- bis hinauf in Verkehrswelten reichen. Die Anwendung des RL für autonome Autos ist hierbei ein herausragendes Beispiel, das verdeutlicht, wie die Kontingenzsteuerung eines Agent:in-Umwelt-Verhältnisses im Experiment eine Modelldynamik begründet, die für das Zusammenspiel von Situationserkennung und Aktionswahl unabdingbar ist.

Ein weiterer Aspekt, der sich aus diesem Modellkern ergibt, ist eine soziale Perspektive des individualisierten Zugangs zum Verkehr, in dem jede mögliche Handlung stets von einer/m Agent:in (oder Konsument:in) in einer unbekannten Umwelt ausgeführt wird. Zugleich zeigt die Anwendung von neuronalen Netzwerken im sogenannten »Deep Reinforcement Learning«, dass eine Skalierung auf höhere Grade der Umweltkomplexität nur dann erfolgen

8 Es geht im Folgenden nicht um den Einfluss von Lerntheorien wie dem Behaviorismus auf die Computer Science oder dem Machine Learning, sondern um die wissenschaftshistorische Operation einer Kontingenzsteuerung in Verhaltensexperimenten, aus der zahlreiche Lerntheorien im 20. Jahrhundert entstehen werden, vgl. dazu Barrett, Louise: »Why Brains Are Not Computers, Why Behaviorism Is Not Satanism, and Why Dolphins Are Not Aquatic Apes«, in: Behaviorist Analyst 39 (2016), S. 9-29.

kann, wenn die Perspektive der individuellen Agentin beibehalten wird. Das RL-Modell hat damit neben einer technischen und einer ökonomischen vor allem auch eine soziale Komponente des/der individualisierten Verkehrsteilnehmer:in, die auf ihre wissenschafts- und medienhistorischen Fundierungen bisher selten untersucht wurde.[9]

In den nächsten beiden Abschnitten werden zunächst die beiden wichtigsten Vorläufer des RL-Modells, die tierphysiologischen Experimente bei Ivan Pavlov sowie die behavioristischen Experimente des amerikanischen Psychologen Burrhus F. Skinner beschrieben. In beiden Fällen wird verdeutlicht, wie Kontingenzpotentiale im Experiment eingeführt und arrangiert werden. Das hieraus generierte Verhältnis von Agent:in und Umwelt wird dann in seinen Modifikationen sowohl im frühen ML als auch in Varianten des »Deep RL« dargestellt. Besonders das letzte Beispiel dient zur Verdeutlichung der These, dass die Kontingenzsteuerung durch unterkomplexe Umweltmodelle auch jene soziale Kernsituation aus Agent:in und Umwelt bestimmt, mit denen auch RL-Modelle Verkehrswelten operieren müssen.

2. Kontingenz in Experimentalsituationen: Pavlos Lehre vom konditionierten Reflex

Zum Ende des 19. und zu Beginn des 20. Jahrhunderts versteht sich die Physik noch immer als jene Wissenschaft, deren Naturgesetze die Bewegungen der Körper auf der Erde und im Weltall erklären. Theorien der Physik müssen prognostisch in der Aussage und induktiv im Verfahren sein. Eine aus den Beobachtungen gemachte Hypothese hat sich demnach wieder an der Empirie zu bewähren – alles andere entspräche der Spekulation.[10] Diesen Anspruch an

9 Diese wissenschaftshistorische und wissenschaftssoziologische Problematik bringt Jack Stilgoe zum Ausdruck, wenn er schreibt: »The identification of tasks and modes of machine learning (for example, reinforcement learning, which adopts a trial, error and reward approach to optimization […]), is inescapably social«. Stilgoe, Jack: »Machine learning, social learning, and the governance of self-driving cars«, in: Social Studies of Science 48/1 (2018), S. 25-56, hier S. 30.

10 Dieser Vorwurf erging von mehreren britischen Physikern an Charles Darwins Evolutionstheorie, die zwar eine breite Grundlage an Beobachtungen hatte, aber deren Hypothesen Aussagen über Zeiträume machten, die empirisch nicht zu verifizieren seien. Besonders an dieser Darwin-Kritik wird, wie der Wissenschaftstheoretiker Helmut Pulte aufzeigt, die Skepsis deutlich, die lebenswissenschaftlichen Formulierun-

deterministische Naturgesetze, die noch ganz im Zeichen der Newtonschen Mechanik stehen, gilt es zu bedenken, um den besonderen Stellenwert von Ivan Pavlovs Theorie des »Konditionierten Reflexes« für das RL einzuordnen. Hinzu kommt die besondere Funktion des Experimentalobjekts Tier, an dem physiologische Fragen erörtert werden, ohne in den Körper des Tieres einzugreifen. Hierdurch kann das Tier als lebendiger Agent in Wechselwirkung mit einer experimentellen Umwelt treten und somit zu einem Komparativ für menschliches Lernverhalten werden. Damit tritt aber auch die Frage auf, wie sich die Spannung von Variation und Kontingenz in den Versuchen mit lebendigen Tieren steuern lässt.

Vor Pavlovs erstem Artikel im renommierten *Science*-Magazin,[11] in dem er die Wechselwirkungen von unkonditionierten und konditionierten Reflexen beschreibt, kommt die physiologische Theorie des vom Stimulus gesteuerten Verhaltens nicht über eine Beobachtungs- und Beschreibungsebene hinaus. Eine erste umfangreiche Publikation hierzu legt der amerikanische Physiologe Edward L. Thorndike 1898 vor. Er hatte eine Reihe von Experimenten mit Katzen unternommen, die den Weg aus diversen Käfigkonstruktionen finden mussten, um an Futter außerhalb des Käfigs zu gelangen. Sollte die Katze aus dem Käfig entkommen, indem sie z.B. ein Seil durchriss und damit eine Tür hob, so würde sie auch im nächsten Käfig genauso verfahren, bis sie nach einiger Zeit merkte, dass in anderen Käfigen hierdurch keine Veränderung einträte und entsprechend andere Wege aufsuchen, bis sie zum Futter gelängte.[12]

---

gen von Naturgesetzen am Ende des 19. Jahrhunderts entgegenschlug, sofern sie nicht dem mechanischen Leitbild der Physik folgten. Siehe dazu Pulte, Helmut: »Darwin und die exakten Wissenschaften. Eine vergleichende wissenschaftstheoretische Untersuchung zur Physik mit einem Ausblick auf die Mathematik«, in: Eve-Marie Engels (Hg.): Charles Darwin und seine Wirkung, Frankfurt a.M.: Suhrkamp 2009, S. 139-177, hier S. 151-160. Zugleich muss aber betont werden, dass dies auch eine wissenschaftspolitische Haltung war, um die Physik als Leitdisziplin zu stärken. Denn bereits im 19. Jahrhunderts gab es mit der Thermodynamik eine Theorie, die von Wahrscheinlichkeiten der Zustandsübergänge von Systemen ausging und somit allein auf einer Makroebene argumentierte, während sie über die Teilchenbewegungen auf der Mikroebene keine Vorhersagen machen konnte.

11 Pavlov, Ivan: »The Scientific Investigation of the Psychical Faculties in the Higher Animals«, in: Science 24/620 (1906), S. 613-619.

12 Thorndike, Eugene L.: »Animal Intelligence. An Experimental Study of the Associative Processes in Animals«, in: Psychological Review 2/4 (1898), (Series of Monograph Supplements). Thorndike entwickelt insgesamt elf Konstruktionen. In manchen ist auch eine Kombination von ehemals erfolgreichen Aktionen notwendig, um einen Ausweg

Anhand dieser Beobachtungen formuliert Thorndike sein »Law of Effects«, dass er als »fundamental law« bezeichnet, um das Lernverhalten von Lebewesen mit Bedürfnissen, Interessen oder Wünschen zu beschreiben.[13] Für Thorndike löst nicht ein Reflex die Handlung aus, stattdessen ist die Handlung das Resultat von Assoziationen auf der Grundlage von Instinkten, die es experimental zu evozieren und zu beschreiben gilt.

Pavlov geht allerdings einen Schritt weiter, denn das beobachtbare Verhalten lässt nicht zwangsläufig auf einen bestimmten Instinkt schließen. Des Weiteren ist eine »nature of instinctive impulses«, wie sie Thorndike noch behauptet, eine Annahme, die keine Verbindung zum spezifischen Nervensystem des Tieres enthält. Ausgehend von seinen Studien zu Verdauungsdrüsen gelangt Pavlov zur Frage nach den externen Stimuli so genannter »psychic secretion« und ihrer physiologischen Ursachen im Nervensystem.[14] In seinen bekannten Hunde-Experimenten kombiniert er einen unkonditionierten Reiz, wie etwa das Präsentieren eines Stückes Fleisch vor dem Hund, das zur Aktivität der Speicheldrüse führt, mit konditionierten Reizen, die akustisch, olfaktorisch oder taktil sein können. Als Folge tritt der Speichelfluss auch auf, wenn das Fleisch nicht mehr im Raum ist und nur der konditionierte Reiz vernommen wird. Damit begründe für Pavlov der »well-known physiological process« des Reflexes erst jene vermeintlichen Instinkte, die gewöhnlich als »psychical stimuli« bezeichnet werden[15].

Die Variationen dieser »psychical stimuli«, wie Pavlov die konditionierten Reize auch nennt, können beliebig sein, solange sie im Experiment iterativ und synchron mit dem Reiz für den unkonditionierten Reflex präsentiert werden.[16] In dieser Experimentalsituation ist der Hund an einem Gestell be-

---

aus dem Käfig zu finden. Neben Katzen kommen auch Hunde und Hühner im Experiment zum Einsatz. Pavlov ist mit den Arbeiten Thorndikes zu Beginn seiner Hunde-Experimente nicht vertraut, hebt aber später seinen eigenen Ansatz als »experimental investigation« von Thorndikes Methode ab, dazu Pavlov, Ivan: »Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex«, in: Annals of Neuroscience 17/3 ([1927] 2010), S. 136-141, hier S. 137.

13 Thorndike, Eugene L.: »The Law of Effect«, in: The American Journal of Psychology 39/1 (1927), S. 212-222, hier S. 212.

14 I. Pavlov: »Conditioned reflexes«, S. 138.

15 I. Pavlov: »The Scientific Investigation«, S. 614.

16 Siehe für ein ausführliches Experiment zur negativen Hemmung Pavlovs Arbeit zur Hundehypnose. Pavlov, Ivan: »Ein Beitrag zur Physiologie des hypnotischen Zustandes beim Hunde«, in: Charakter. Eine Vierteljahresschrift für psychodiagnostische und verwandte Gebiete 4 (1933/1934), S. 181-190.

festigt, während eine Röhre in seinem Mund den Speichel in ein Gefäß leitet, um die Intensität der Reaktion auf den konditionierten Reiz zu messen. Es gehört zu Pavlovs physiologischer Methodik, dass die Umwelt der Tiere so angerichtet wird, dass der innere Mechanismus von außen sichtbar ist. Mit anderen Worten: Pavlovs Experimentalsetting ist eine materialisierte Hypothese über die mechanischen Funktionen des Verdauungstraktes. Daher darf kein zusätzlicher Faktor diese funktionelle Umwelt des Experimentes unterlaufen – selbst Pavlov nicht.[17]

Wissenschaftshistorisch bietet er damit nicht nur eine Alternative zu den invasiven Methoden, die bis dato in der Physiologie verwendet wurden.[18] Er konstituiert mit dem konditionierbaren Hund ein neues Wissensobjekt, das zugleich auf den Menschen als umweltsensibles Wesen zurückwirkt.[19] Dazu muss aber jegliches Kontingenzpotential reduziert werden, bis das Lebewesen selbst das einzige Kontingenzpotential darstellt. Hierdurch wird jede Aktion, die nicht als eine Reaktion auf den präsentierten Stimulus interpretiert werden kann, zu einem regelrechten »freedom reflex«, mit dem das Tier gegen die Begrenzung seiner Freiheit protestiere.[20] Auch das unerwartete Verhalten

17 Folglich ist Pavlov während der Experimente niemals im selben Raum wie der Hund, um die Wirkung der Stimuli nicht zu beeinflussen. Siehe dazu Bühler, Benjamin: »Experimentalobjekte. Tiere als Figuren anthropologischen Wissens«, in: Annette Bühler-Dietrich et al. (Hg.): Topos Tier. Neue Gestaltungen des Mensch-Tier-Verhältnisses, Bielefeld: transcript 2016, S. 19-39.

18 Lesch, John E.: »The Paris Academy of Medicine and Experimental Science, 1820-1848«, in: William Coleman/Frederic L. Holmes (Hg.): The Investigative Enterprise. Experimental Physiology in Nineteenth-century Medicine, Berkeley, CA.: University of California Press 1988, S. 100-137.

19 B. Bühler: »Experimentalobjekte«, S. 33.

20 Genau jenen »Reflex der Freiheit« führt Helmuth Plessner zum Schluss seiner Kritik an Pavlos physiologischer Theorie an, um den Unterschied eines Verhaltens des Organismus anstelle eines Vorgangs des Reflexes deutlich zu machen. Plessner, Helmuth: »Die physiologische Erklärung des Verhaltens. Eine Kritik an der Theorie Pavlovs«, in: ders. (Hg.): Gesammelte Schriften, Bd. VIII, Frankfurt a.M.: Suhrkamp [1935] 2003, S. 7-32. Plessners Kritik kann hierbei stellvertretend für andere Autor:innen gesehen werden, die Pavlovs mechanistischer Interpretation ein ganzheitliches Modell über das Umweltverhalten von Organismen entgegenstellen. Siehe dazu Gruevska, Julia: »Das naturimitierende Labor. Philosophisch-anthropologische Implikationen in der Physiologie Frederik Buytendijks«, in: dies. (Hg.): Körper und Räume, Wiesbaden: Springer 2018, S. 133-151; aus medienanthropologischer Sicht Rieger, Stefan: Kybernetische Anthropologie. Eine Geschichte der Virtualität, Frankfurt a.M.: Suhrkamp 2003, hier S. 466-483.

des Tieres liege somit in der Natur seiner Reflexe. Es stellt damit einen Faktor dar, der sich durch eine adäquate Umweltmodellierung kontrollieren lässt, indem das Tier lernt, mit einem aktuellen Stimulus mögliche kommende Zustände zu verbinden. Das diskretisierte Zeitverständnis aller RL-Modelle, in dem es nur Zustände mit möglichen Anreizen zur Aktion oder zur Unterlassung einer solchen gibt, ist in dieser Hinsicht eine Folge des systematischen Unterbindens des »freedom reflex« – oder anders: es ist eine Folge der Steuerung von Kontingenz zur Wahrung eines theoriestützenden Experiments.

3. Modellieren für die effektive Kontrolle der Umwelt: Das Reinforcement Learning im Behaviorismus

Für die junge Strömung des Behaviorismus wirken Pavlovs Experimente wie eine Anleitung zur Verhaltensforschung. Einem ihrer Pioniere, dem Psychologen John B. Watson, wird die Konditionierungstheorie zum finalen Anstoß für eine Psychologie, die es sich zum Ziel machen soll, das Verhalten von Tieren und Menschen kontrollieren und vorhersagen zu können.[21] Um diesem Anspruch selbst gerecht zu werden, versucht Watson das Verhalten seiner Agent:innen in variierenden Experimenten so zu konditionieren, dass sich emotionale Zustände wie Angst, Furcht oder Freude bestimmten Stimuli eindeutig zuordnen lassen können.[22]

Sowohl Pavlos Reflextheorie als auch dem Behaviorismus wurde häufig eine reduktionistische oder mechanistische Konzeption von Organismus und Umwelt vorgeworfen. Aus wissenschaftshistorischer Sicht stehen aber beide

21 Watson, John B.: »Psychology as the Behaviorist Views it«, in: Psychological Review 20/2 (1913), S. 158-177, hier S. 158.

22 Watson unternimmt mit seiner Ko-Autorin Rosalie Rayner an einem elf Monate alten Säugling den Versuch, die emotionale Regung von Furcht zu konditionieren. Das Kind mit dem Namen Albert wird dabei auf einen Tisch gesetzt, ihm wird eine weiße Ratte, später auch ein Kaninchen präsentiert, nach denen Albert hin und wieder seine Hand ausstreckt. Bei den folgenden Durchgängen erklingt synchron mit dem Erscheinen der Tiere ein lauter Ton, ausgelöst durch einen Hammerschlag auf einer aufgehängten Stahlstange, die hinter dem Kind angebracht wird (der konditionierte Reiz). Hatte Albert zuvor noch nach der Ratte oder dem Kaninchen gegriffen, weicht er nach den Durchgängen mit dem Hammerschlag zurück und beginnt manchmal auch zu weinen, Watson, John B./Rosalie Rayner: »Conditioned Emotional Reactions«, in: Journal of Experimental Psychology 3/1 (1920), S. 1-14.

in der Nachwirkung des Theorie- und Methodenanspruches der Physik im 19. Jahrhundert. Dabei ist der prognostische Charakter dieser experimentellen Episteme zu betonen – nicht zuletzt auch deshalb, weil er mit einer kontrollierten Reduktion von Kontingenz in Experimenten mit Lebewesen einhergeht. Eine systematische und zunehmend selbstreflexive Integration des Begriffes von Kontingenz zur Beschreibung des Versuchsaufbaus findet sich dann aber erst in den viel zitierten Experimenten von Burrhus F. Skinner.

Skinner und sein Ko-Autor Charles B. Fester bezeichnen ihre Experimentalreihen als »Schedules of Reinforcement«. Im gleichnamigen, über 700 Seiten umfassenden Buch, das 1957 erscheint, entwerfen sie mehrere Parameter, mit denen die Konditionierbarkeit des Verhaltens sich nicht nur in einigen Durchläufen, sondern über einen längeren Zeitraum einstellen lassen soll. Hierin erhält auch der Begriff des »reinforcement« seine bis heute anhaltende Bedeutung, die nicht in wiederkehrenden Kopplungen von konditionierten und unkonditionierten Reizen besteht, sondern in einer Wahrscheinlichkeit der Stabilisierung des Verhaltens über eine längere Zeit bei variierenden Stimuli. Die zentrale Herausforderung ist die Modulation einer optimalen Mischung von geplanten und spontanen Stimuli, mittels derer a posteriori eine Verhaltensänderung den gesetzten Stimuli (*reinforcements*) zugeschrieben werden kann, wenngleich Letztere, wie betont wird, selten eine direkte Aktion hervorrufen können. Das Medium hierzu sind zeitliche Variationen von Stimuli, Auslassungen und Belohnungen, sogenannte »properties of schedules«,[23] die in den späteren RL-Modellen autonomer Autos allgemein als diskrete Zustände (*states*) einer modellierten Umwelt beschrieben werden, in der der/die Agent:in sich für eine Aktion entscheiden muss, um ihr Ziel zu erreichen. An die Stelle von Pavlos »ideal of inevitable reinforcement«, das noch mit einer kausalen Notwendigkeit argumentiert, tritt die »manipulation of schedules«, die das Verhalten sowohl auf Grundlage eines längeren Untersuchungszeitraums erklären als auch für einen längeren Zeitraum vorhersagen will.[24]

Unterschieden wird folglich zwischen regelmäßigen und zufälligen Reaktionen (*fixed and variable ratio*) sowie zwischen regelmäßig oder zufällig gesetzten Stimuli in einem Zeitabschnitt (*fixed and variable intervals*).[25] Austra-

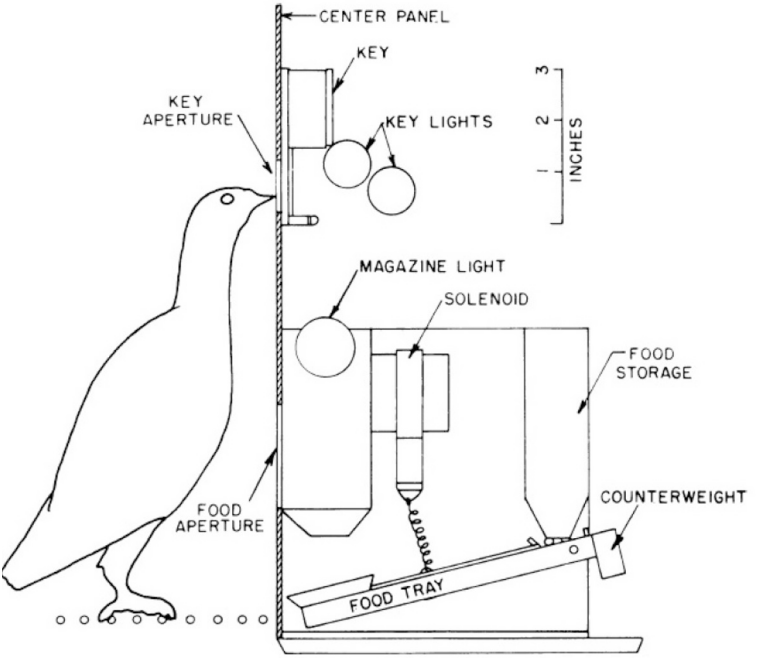
23 Fester, Charles B./Skinner, Burrhus F.: Schedules of Reinforcement Learning. Cambrige, MA. Online-Reprint Series der B.F. Skinner Foundation 1957, hier S. 15.

24 Ebd.

25 Ebd., S. 17.

gungsort der Experimente ist eine Variante der berühmten Skinner-Box (Abb. 1). In einem Käfig wird in etwas höherer Stellung ein Schlüssel angebracht, während unten eine Futterluke zu sehen ist.[26] Der Raum, in den eine Taube hineingeht, ist beleuchtet. Beim Picken nach dem Schlüssel kommt aus der Luke Futter heraus, was je nach Anordnung in verschiedenen Zeitabständen oder in Kombination mit anderen Reizen geschehen kann.

Abbildung 1: Elektromechanisch automatisierte Version der Skinner-Box



Fester, Charles B./Skinner, Burrhus F.: »Schedules of Reinforcement Learning.« Cambrige, MA. Online-Reprint Series der B.F. Skinner Foundation 1957, hier S. 23.

Festers und Skinners Version eines »Deep Learning« sieht Experimente mit der Dauer von bis zu fünfzehn Stunden vor. Dabei ist die Schlüsselvorrichtung mit einem Zähler für Kontaktfrequenz und einem Zeitmesser für die Intervalle zwischen dem Eintreten in die Box und dem Beginn des Pickens

26 Ebd., S. 27.

versehen. Um auch bei Abwesenheit des Laborpersonals die Fortdauer der Untersuchung zu gewährleisten, wird die gesamte Box mit einem elektrischen Kreislauf versehen, über den vom Licht bis zur Futterluke alles durch Kontaktstellen und Relaisvorrichtungen gesteuert wird.[27] Bei einer derart minutiösen und automatisierten Versuchsdurchführung gerät das Forschungsobjekt fast in Vergessenheit. Dabei ist gerade die Taube Dreh- und Angelpunkt einer systeminternen Kontingenz: »But among the physical events occuring in the experimental chamber are the activities of the organism itself. These enter into the contingencies and must be specified as part of the animal's environment.«[28] In einem Umweltmodell, in dem zu den zufällig gesetzten Stimuli auch unvorhersehbare Aktionen verzeichnet werden, drängt sich die Frage auf, welche Verhaltensänderung welchem Stimulus bzw. welcher Stimulusfolge – wenn überhaupt – entspricht, oder mit den Worten der beiden Behavioristen: »We deal with a response both as an activity of the organism and as part of a series of events affecting the organism as a stimulus«.[29]

Diese Ko-Existenz von gestiegener Kontingenz und einer problematischen Zuschreibung der Umweltstimuli, die ein zielorientiertes Verhalten verstärken oder schwächen könnten, ist die in die Skinner-Box verlagerte Problematik einer Relation zwischen dem Umweltmodell des Experimentators und dem Umweltmodell der Taube. Um zu wissen, welcher Reiz bei einem Tier welches zielorientierte Verhalten auslöst, ohne auf eine Pavlov'sche Konditionierbarkeit zurückzugehen, müsse der Reiz selbst mit einer in den Versuch verbauten Kontingenz gekoppelt werden. Diese gesteuerte Kontingenz der restringierten Handlungsmöglichkeit wird zu einem Interface, an dem der Lerneffekt durch »reinforcement« sicht- und messbar werden soll. Genau diese Gestaltbarkeit eines Interfaces für den sichtbaren Vergleich heterogener Umweltentwürfe ist es, die sich sowohl auf normativ gerahmte Alltags- wie auch Spielsituationen übertragen lässt. Solche Interfaces können Objekte, Hindernisse oder eben Verkehrszeichen darstellen, wie Skinner in seinem späteren Buch »About Behaviorism« betont:

27 Dabei werden die Zeitpunkte der Futterausschüttung sowie die Intervalle zwischen den Kontakten der Taube mit dem Schlüssel gemessen und per Stift auf einem Papierstreifen notiert. Die Relais sind so verschaltet, dass das Licht der Box sowie das Licht am Schlüsselhalter nur manuell ausgemacht werden kann (vgl. ebd., S. 25-27).

28 Ebd., S. 19.

29 Ebd.

> But we can also arrange that a particular object will be seen by establishing contingencies which can be met only by responding to it. Traffic signs are designed to be easily seen, but we see them or ignore them largely because of the contingent consequences. Measures of this sort are often said to increase a person's awareness, or to expand his mind or consciousness, but they simply bring him under more effective control of his environment.[30]

Wenn die verbaute Kontingenz zu einem indirekten Steuerungstool für die Handlungen von Agent:innen wird, dann ließen sich nach behavioristischer Logik auch Lerneffekte beobachten, deren einzige Quelle eben diese Stimuli einer unterkomplex modellierten Umwelt sind, die nach Bedarf von Taubenkäfigen bis hinein in Verkehrssituationen zu skalieren wäre. Weder ein »mind« noch eine »consciousness« seien hierzu notwendig, sondern die Idee eines steuerbaren Verhaltens durch die Anordnung von Kontingenz im Experiment.[31] Dies ist allerdings keine Epistemologie, die sich auf Grundlage eines Spiels entwickelt, sondern in der Spannung von experimentellen Umweltkonstruktionen und einer Agency von Tieren liegt, deren nicht bekanntes Umweltmodell mehr und mehr nach außen – auf Objekte, Hindernisse oder Verkehrszeichen – verlagert und dort materialisiert wird.

4. Reinforcement Learning als Modell des Machine Learning

Das vorige Zitat von Skinner verdeutlicht, wie sich durch eine Setzung begrenzter Handlungsmöglichkeiten Effekte des »reinforcement« generieren lassen, ganz gleich, ob die Ursache hierfür der Organismus, das Gehirn oder ein Ereignis in der Umwelt ist. Diese Episteme der Minimalbedingungen wird im Machine Learning (ML) der 1980er Jahre wieder aufgenommen. Das RL des ML geht ebenfalls wie der Behaviorismus von diskreten Zuständen aus (*states*), die sich dem/der Agent:in in Form von Stimuli mitteilen. Dabei gilt es, Entscheidungen zu treffen (*actions*), die zu einer Belohnung oder Vermeidung von Strafen (*reward function*) führen, wobei das übergeordnete Ziel

30 Skinner, Burrhus F.: About Behaviorism, New York: Vintage Books 1974, S. 37.

31 Zur Bedeutung dieses kognitionsarmen Lernmodells für die Maschinen in der frühen Kybernetik siehe Pickering, Andrew: The Cybernetic Brain. Sketches of Another Future. Chicago und London: Chicago University Press 2008, sowie Müggenburg, Jan: Lebhafte Artefakte. Heinz von Foerster und die Maschinen des Biological Computer Laboratory, Konstanz: Konstanz University Press 2018.

die Maximierung dieser Belohnung ist (*goal* oder *value function*). An die Stelle einer Beobachtung des Lernverhaltens tritt nun die Verhaltensoptimierung von zielorientierten Agent:innen. Was Fester und Skinner mit den »schedules of reinforcement« untersuchen wollten, nämlich die Folge der Einsicht, dass nur auf die wenigsten Stimuli eine direkte Aktion erfolgt, wird im ML zum zentralen Dreh- und Angelpunkt für die Entwicklung von Algorithmen. Dieser Punkt liegt in der zeitkritischen Variable, wie viel Erfahrung zur optimalen Entscheidung reicht und wieviel von der unbekannten Umwelt noch erforscht werden muss (*exploiting and exploring*).

RL-Modelle des ML gehen von einer indeterminierten Umwelt für den/die Agent:innen aus, der/die ihr eine explorative und »hedonistische« Einstellung gegenüber einnehmen.[32] Diese Kombination wird als *trial-and-error*-Verfahren bezeichnet. Während in den psychologischen Experimenten von Fester und Skinner kein Organismus angenommen wurde, der eine bestimmte Lernfähigkeit ermöglicht, sondern nur ein Lernverhalten beobachtet wurde, muss im ML ein Verfahren gefunden werden, mit dem die vergangenen Zustände im Angesicht des aktuellen Zustandes bewertet werden können, oder anders: Lernen bedeutet hier, die Wahrscheinlichkeit zu maximieren, dass nach der nächsten Entscheidung die bestmögliche Belohnung folgt, eben weil die vergangenen für die Optimierung möglicher Entscheidungen betrachtet werden sollen.[33] Dies impliziert eine adäquate Gewichtung der vergangenen Entscheidungen, ein durchgängiges Problem der KI, das Marvin Minsky bereits 1961 als »Credit Assignment Problem« bezeichnet. Man stelle sich ein Schachspiel vor, bei dem man immer wieder auf eine durch den Gegner veränderte Umwelt reagieren muss und in der ca. eine Million Züge zum gewünschten Sieg führen. Könnte man nun, so fragt Minsky, eine Millionstel des Verdienstes jedem gemachten Zug zuschreiben?[34]

Intuitiv sagten sich auch die Vertreter der frühen KI – Nein! Aber wie im Falle von Minsky lösten sie das Problem, indem sie der Komplexität der Weltprobleme eine Komplexität der Rechner- bzw. Hirnarchitektur entgegenstell-

32 Sutton, Richard/Barto, Andrew: Reinforcement Learning. An Introduction, Cambridge: MIT Press 1998, S. 21.

33 Kaelbing, Leslie P./Littman, Michael L./Moore, Andrew W.: »Reinforcement Learning: A Survey«, in: Journal of Artificial Intelligence Research 4 (1996), S. 237-285, hier S. 239.

34 Minsky, Marvin: »Steps Towards Artificial Intelligence«, in: Proceedings of the IRE 1 (1961), S. 8-30, hier S. 20.

ten, die auf die eingehenden Informationen eine adäquate Problemlösungsstrategie entwickeln sollte.[35] In den 1980er Jahren, als das RL im Zuge des ML wieder aufkam, waren Fragen der logisch hergeleiteten Problemlösung nicht mehr maßgebend, sondern die stochastisch ermittelte Gewichtung von Daten aus der Umwelt, die über Sensoren eingehen. Denn es ist unerheblich, welches Element der Operationskette welchen Anteil an der richtigen Entscheidung hat, solange der/die Agent:in in der Erwartung verbleibt, dass seine nächste Aktion die Belohnung maximieren könnte.[36] Folglich grenzt sich das RL von anderen ML-Verfahren wie dem Supervised Learning ab, in denen bestimmte Input/Output-Paare vorgegeben sind oder eine Vollzugsweise von Aktionen einprogrammiert ist.[37] Es trägt noch jenen Beobachtungscharakter einer Kopplung von Agent:in und Umwelt, die sich in der Episteme eines Verhaltensexperimentes ausdrückt, wie es von Pavlov und Skinner konstituiert wurde. Zugleich wird die Komplexität der Welt nicht mehr durch ein hierarchisch geordnetes Problemlösungsverfahren in die/den Agent:innen eingeschrieben, sondern in einen stochastisch ermittelten Erwartungshorizont verlagert, dessen Berechnung jeder Aktion zugrunde liegen muss.

Was geschieht mit solchen ML-Modellen in der Anwendung auf Verkehrssituationen? Dazu wird abschließend weniger der technische Aspekt im Vor-

35 Als konkretes Beispiel für solche Zuschreibungsprobleme führt Minsky Programme an, die in mehrere Subroutinen aufgeteilt sind. Subroutinen sind zueinander als auch untereinander hierarchisch geordnet und repräsentieren die Aufteilung von hoch- und niedrigstufigen Aufgaben im Programm. Minsky liest den Aufbau der Subtoutinen im Programm häufig isomorph zum Aufbau des Gehirns, wobei er »Intelligenz« eher als eine ästhetische Frage von Problemlösungsstrategien bezeichnet und nicht als spezifisch kognitive Eigenschaft. Folglich ist der zentrale Begriff für ihn »heuristics« und nicht »intelligence« (ebd., S. 27).

36 Mathematisch lässt sich dies mit dem Markov-Entscheidungsproblem lösen. Die Verfahren von Markov-Ketten oder auch Markov-Eigenschaften werden angewandt, um einen stochastischen Wert in einem aktuellen Zustand für einen zukünftigen Zustand zu erhalten. Im RL-Beispiel ist es für die/den Agent:in die Frage, ob eine Information aus der Umwelt zum Zeitpunkt *t* eine Markov-Eigenschaft hat, d.h., ob es eine mögliche Belohnung in *t+1* gibt. Um dies stochastisch zu bestimmen, werden die jetzigen Informationen sowie alle ähnlichen Entscheidungssituationen betrachtet, um die Wahrscheinlichkeit einer Belohnung (oder Bestrafung) für den nächsten Zeitpunkt zu bestimmen. Zu rekursiven Funktionen und Markov-Ketten siehe auch Ofak, Ana/von Hilgers, Philipp: »Einleitung«, in: dies. (Hg.): Rekursionen. Faltungen des Wissens. München: Fink 2010, S. 7-21.

37 L.P. Kaelbing/M.L.Littman/A.W.Moore: »Reinforcement Learning«, S. 239.

dergrund stehen, ob solche Modellierungen bewerkstelligt und performt werden könnten. Stattdessen wird dargelegt, dass das RL nicht unabhängig von seiner Experimentalgeschichte und der dortigen Steuerung von Kontingenz betrachtet werden kann, will man die Popularität, aber auch die fraglichen Konsequenzen des RL-Modells für autonome Fahrsysteme nachvollziehen.

5. Policy-Design und Skalierung: Reinforcement Learning für autonome Autos

RL-Modelle werden seit den 2010er Jahren verstärkt zum Trainieren von neuronalen Netzwerken eingesetzt. Es kann und soll hier keine vollständige Darstellung der RL-Anwendungen für autonome Autos erfolgen, sondern das Argument ausgeführt werden, dass trotz der Hinzunahme von neuronalen Netzwerken spezifische Probleme des RL bleiben, die nicht technischer, sondern epistemologischer Art sind. Für den Einsatz autonomer Autos betrifft dies die Handlungsstrategie (*policy*) und die Skalierbarkeit der Ergebnisse. Besondere Aufmerksamkeit wird dem RL-Modell seit einer Publikation zuteil, die unter dem Titel »Human-level control through deep reinforcement learning« von einem Deep Neural Network-Test an 49 Atari-Spielen berichtet.[38] Auch wenn es im Fazit dieses Textes heißt, dass das Deep Neural Network in den meisten Fällen besser abschneidet als alle bisher genutzten Algorithmen und das Score-Niveau eines professionellen Spieltesters erreicht, so wird zum Ende des Artikels doch festgehalten, dass so genannte »temporally extended planing strategies« immer noch eine große Herausforderung für das RL-Modell darstellen.[39] Für die ML-Community war dies zugleich der Anreiz, Lösungen für die besagten Planungsstrategien zu entwerfen. Hierdurch wurde neben dem Pärchen von Zuständen und Aktionen (*state-action pairs*) ein weiterer Faktor entscheidend – die Handlungsstrategie, auch *policy* genannt. Mit Hilfe der *policy* soll der/die Agent:in lernen, nicht der nächsten Erwartung auf Belohnung zu folgen, sondern unmittelbare Belohnungen bis zu einem bestimmten Grad zurückzustellen, um noch mehr Informationen über die Umwelt sammeln zu können.[40] Die Frage der *policy* läuft somit auf

38 V. Minh et al.: »Human-level Control«.

39 Ebd., S. 532.

40 Dies war bereits in den späten 1980er Jahren ein zentrales Forschungsthema in der ML-Community. Das Problem hierbei ist, wie ein(e) Agent:in nach einer optimalen Ent-

eine Simulation der »schedules of reinforcement« hinaus, mit denen Charles Fester und Burrhus F. Skinner erforschten, wie sich ein Verhalten über einen längeren Zeitraum im Experiment einstellt, wenn unmittelbare Reizquellen selten eine direkte Aktion beim Tier auslösen.

Während Fester und Skinner aber rigoros die Frage ausklammern, welche Sinnesdaten die Taube über welche Organe empfängt, ist dies für Konstrukteur:innen autonomer Autos mit Blick auf das Sammeln und Sampeln ihrer Daten von großer Relevanz. So wird zwischen zwei Informationen unterschieden, dem *state space* (Position, Richtung, Abstand zu anderen Autos etc.) und der *vehicle control* (Fahrwinkel, Brems- und Beschleunigungsverhalten, Schätzung der eigenen Bewegung und der benachbarten Autos).[41] Bewegungen in der Nähe des Autos werden sensorisch aufgegriffen (Lidar, Radar) und im Simulationsmodus in 2D-Ansichten übersetzt, um zu zeigen, wie das Auto seine Umwelt »wahrnimmt«. Eine solcher Ansichten ist das so genannte *occupancy grid*, eine 2D-Perspektive, bei der ein Gitternetz um das Auto gelegt wird, um den direkten Einflussbereich sowie angrenzende semantische Informationen zu visualisieren (vgl. Abb. 2). Ein Blick auf dieses Gitter drängt vor dem hier erläuterten Hintergrund den Eindruck auf, dass die autonomen Agenten im Straßenverkehr ihren virtuellen Käfig selbst mit sich herumfahren.

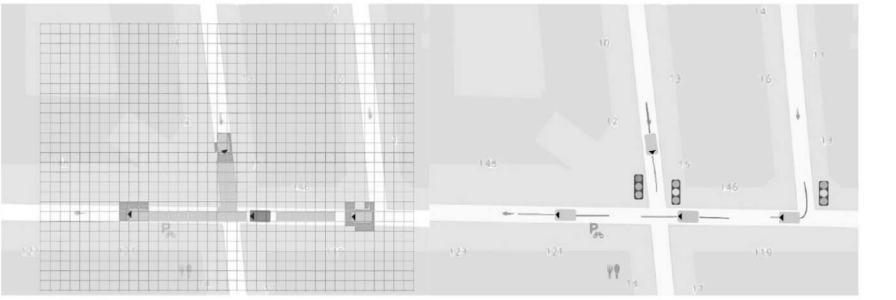
Der Grid als virtueller Käfig dient hier nicht nur einer Metaphorik. Bedenkt man das *policy*-Problem in variierenden Umwelten, so verdeutlicht sich damit der Anspruch an RL-Algorithmen, eine gesunde Balance zwischen der Erforschung der Umwelt und der belohnungsmaximierenden Entscheidung zu treffen. Es liegt hierbei nahe, dass in einer dynamischen Umwelt mehrere *policies* notwendig werden, je nachdem, ob man einparkt, überholt oder im Stau steht. Wie und aus welchen Stimuli solche adaptiven *policies* berechnet und damit auch in eine Aktion gewandelt werden, ist ein aktuelles Thema in der Forschung zu Lernalgorithmen. Die Ansätze variieren dabei von deterministisch orientierten *policies*, die einem *supervised learning* ähneln, bis hin

---

scheidung nicht bei ähnlichen Situationen genauso handelt, ohne die Umwelt weiter abzutasten. Dies setzt einmal voraus, dass frühere Zustände in die Berechnung miteinfließen, sowie eine Funktion, die vermeidet, dass alle ähnlichen Situationen zu einer gleichen Vorhersage führen (die so genannte *discount rate*). Generell spricht man von »Methods of Temporal Differences« für die Vorhersagen nach auswirkungsreichen Entscheidungen. Siehe dazu Sutton, Richard: »Learning to Predict by the Methods of Temporal Differences«, in: Machine Learning 3 (1988), S. 9-44.

41 R. Kiran et al.: »Deep Reinforcement«, S. 1.

Abbildung 2: Das so genannte »occupancy grid« aus einer computersimulierten Vogelperspektive



Ravi Kiran, B./Sobh, Ibrahim/Talpaert, Victor et al.: »Deep Refinforcement Learning for Autonomous Driving: A Survey«, arXiv:2002.00444v2 [cs.LG], 23. Januar 2021, S. 10.

zu Algorithmen, die vor jeder Entscheidungsfindung die letzten Erfahrungen stochastisch neu auswerten und daraufhin erst den Belohnungswert für jede mögliche Aktion neu berechnen.[42] Im Kern steht dabei ein Designproblem von Algorithmen, das sich in jeglicher Hinsicht sozial niederschlägt. Denn je größer das Sampling ausfällt, also die Auswahl der entscheidungsrelevanten Zustände, umso enger ist der Spielraum an möglichen Handlungsstrategien (*policies*), die in diesem Zeitraum abgerufen werden können.[43]

Folglich ist die Frage des Datensamplings vergleichbar mit Skinners Idee einer Steuerung von Kontingenz durch den Einsatz gewählter symbolischer Stimuli. Auch wenn die Quantität und der Abstraktionsgrad in den Trainingsdaten eine andere induktive Grundlage darstellen als es in den Verhaltensexperimenten von Skinner der Fall ist,[44] liegt dennoch dieselbe epistemische

42 Eine solche Variante ist die »Actor-Critic-Method«, wobei *Critic* hier bedeutet, dass nach jeder Entscheidung auf Grundlage der vergangenen Entscheidungen (und ihrer Vorhersagen) neu bewertet wird, ob der Zielwert, also das Maximum an Belohnungen, erhöht wurde oder nicht. Actor-Critic-Methoden sind häufig Formen eines *unsupervised learning*, bei dem die optimale *policy* durch die Erforschung der Umwelt errechnet und später iterativ angepasst werden soll. (vgl. Ravi Kiran et al. : »Deep Reinforcement«, S. 4-5).

43 Ebd., S. 12 ; Shalev-Shwartz et al.: »Safe, Multi-Agent«.

44 Was sich noch durch den Umstand vertiefen ließe, dass für das Sampling auch die Daten anderer autonomer Autos hinzugezogen werden können. Das Updaten der Fahrsoftware geschieht auf Grundlage von Datensätzen, die aus allen autonomen Au-

Idee einer Steuerung von Kontingenz durch mögliche, diskretisierte Entsprechungen von Zustand und Aktion zugrunde. Skinners Beispiel der Ampel als Umweltstimuli, deren Wirkung auf unser Verhalten auch ohne die Annahme eines Bewusstseins zu erklären wäre, versinnbildlicht die Modellierung des Verhaltens durch die Steuerung von Kontingenz in diskretisierten Zuständen. Ob eine normative Ordnung des möglichen Verhaltens nun wie im *supervised learning* vorgegeben wird oder ob sie sich einer induktiven Logik folgend aus den Samplings generiert, ist dabei zweitrangig. Wissenschafts- und medienhistorisch betrachtet, verbleibt das RL-Modell in der Zuordnung von *state-action*-Pärchen, in der Beobachtung individuellen Verhaltens in Umwelten mit einer modifizierbaren Komplexität. Ob nun das Gehirn, das Nervensystem oder die sensorischen Daten für das Verhalten ausschlaggebend sind, ob ein *supervised* oder ein *unsupervised learning* in Kombination mit dem RL angewandt wird, bleibt eine Folgefrage. Was in diesen Konstellationen hervortritt, ist die verhaltenspsychologische Grundkonstante, eine(n) individuelle(n) Agent:in gegenüber einer Umwelt zu positionieren, in der sich die Kontingenz steuern lässt, so dass das Verhalten der/des Agent:in vorhersagbar wird. Es geht nicht darum, die experimentellen Möglichkeiten der Verhaltensprädiktion zu verkennen, sondern darum, dass bei jeder Anwendung des RL-Modells die eins-zu-eins-Situation einer Welterschließung vorherrscht. Der virtuelle Käfig, den die autonomen Autos in den Simulationen durch die Straßen fahren, entspricht einer ungewollten Versinnbildlichung der Pavlov'sche n Ausgangssituation – als wäre einer seiner Hunde nun auf der Fahrbahn!

Ein letztes Argument, dass das RL-Modell nicht nur besonders adaptiv für Individualitätskulturen wie die Automobilität macht, sondern auch für den Vergleich mit Menschen, ist die Skalierbarkeit. Das sich vom Verhalten eines autonomen Autos auf mehrere Tausend Fahrer:innen skalieren lässt, ist mehrfach beschrieben worden.[45] Es gehört aber ebenso zu einer Eigenschaft des RL-Modells, dass es gleich mehrere Migrationen in andere wissenschaft-

tos in eine Cloud des Herstellers Eingang finden. Dies bedeutet auch, dass ein Fahrfehler für das Fahrverhalten aller anderen Autos relevant werden könnte. Etwas emphatisch wird in diesem Kontext denn auch gleich von einem »fleet learning« gesprochen, siehe dazu J. Stilgoe: »Machine learning«, S. 35 sowie den Text von Jan Distelmeyer in diesem Band.

45 Vgl. J. Stilgoe: »Machine learning«, S. 35 sowie den Text von Jan Distelmeyer in diesem Band.

liche Disziplinen erfahren hat – allen voran in die Neurowissenschaften.[46] Die Frage nach einer Optimierung der Lernalgorithmen ist dabei das verbindende Glied. So wurde eine Trainingsvariante für rekurrente neuronale Netzwerke konzipiert, die als »slow RL« nicht mehr auf ein anwendungsspezifisches Verhalten ausgerichtet sein soll, sondern über Tage ein eigenes Lernverhalten aus einem großen Sampleumfang generieren soll.[47] Anhand dieses Vorgehens, bei dem Schichten von Netzen wieder in die nächste Berechnung Eingang finden, sollen die entsprechenden Gewichtungen der Netzwerke vom Algorithmus selbst erlernt werden. Die/der Agent:in führt damit nicht mehr ein RL-Modell im strengen Sinne aus, das auf eine Aktion hinausläuft, sondern nur noch ein antrainiertes Lernverhalten, das für eingehende Zustände angepasst werden muss.

Gerade solche Verfahren sind in den Kognitionswissenschaften euphorisch aufgenommen worden. Unter dem Terminus des »Meta-RL« verknüpfen sich Fragen nach den Lernstrategien von Tieren und Menschen, nach episodischen Mustern und abrupten Einschnitten von Erfahrungen. Lernen wird demnach zu einer fortlebenden Abfolge von *policies* des Lernens, in denen »one RL algorithm gives birth to another, and hence the moniker ›meta-RL'«.[48] Die Idee, »Deep RL« auch zum Vergleich für menschliches Lernverhalten zu verwenden, ist von hier aus nicht mehr weit. Und mehr noch, selbst das Bias-Problem aller psychologischen Verhaltensexperimente, in denen das Lernverhalten zu stark vom Datensample abhängig wäre, könne durch die Methode des »meta-RL« untersucht und unterbunden werden.[49] Dies verdeutlicht, dass es sich beim RL nicht nur um ein aktuelles Modell für das Trainieren neuronaler Netze handelt, sondern das es mit seinem wissenschaftshistorischen Hintergrund eine doppelte Skalierbarkeit mitbringt – einmal vom Individuum hinauf auf ein allgemeines Lernverhalten, aber auch von einer tech-

46 Zur Adaption des RL-Modells in den Neurowissenschaften und ihren Subdisziplinen wie der Neuroökonomie, siehe Kasprowicz, Dawid: »Zwischen Fruchtfliege und Global Player. Zur neuroökonomischen Modellierung des Lernverhaltens«, in: Johann S. Ach/Beate Lüttenberg/Alexa Nossek (Hg.): Neuroimaging & Neuroökonomie: Grundlagen, ethische Fragestellungen, soziale und rechtliche Relevanz, Münster: Lit-Verlag 2016, S. 149-166.

47 Duan, Yan/Schulmann, John/Chen, Xi et al.: »Fast Reinforcement Learning Via Slow Reinforcement Learning«, in: arXiv:1611.02779v2 [cs.AI], 12.2.2021.

48 Botvinick, Matthew/Ritter, Sam/Wang, Jane X. et al.: »Reinforcement Learning, Fast and Slow«, in: Trends in Cognitive Sciences 29/5 (2019), S. 408-422, hier S. 413.

49 Ebd., S. 418.

nischen auf eine biologisch-kognitionswissenschaftliche Fragestellung (und andersherum). Diese Flexibilität ist aber aus der minutiösen Einführung von Kontingenzpotentialen entstanden, deren Steuerungstechniken sich wissenschaftshistorisch nachzeichnen und vergleichen lassen. Erst vor dem Hintergrund solcher Rekonstruktionen kann die Migration von Modellen wie dem RL kartographiert werden, ohne eine Kontinuität oder gar Linearität in den historischen Anwendungsszenarien zu postulieren.

Demnach lässt sich die Variabilität des RL nicht erst aus einer spezifischen Formalisierung herleiten, die mit dem Computerspiel oder dem ML eingetreten wäre. Ihr liegt eine für die Experimentalkultur wesentliche Situation zugrunde, das individuelle Verhalten in einer geschlossenen Umwelt zu untersuchen, deren Kontingenz von außen gesteuert wird. Die Idee, ein Lernen als Sukzession von Zustands-Aktion Folgen oder ihren Unterbrechungen zu lesen, ebenso wie die Les- und Modifizierbarkeit der *policies*, erstrecken sich entlang der Settings von Pavlov, Skinner, den ML-Anwendungen sowie ihren datenintensiven Formen beim Trainieren von neuronalen Netzwerken für autonome Autos. Solche medien- und wissenschaftshistorischen Grundlagen von Modellen und ihre Mediationsleistungen stehen noch am Anfang,[50] wenn es darum geht, ihre Wechselwirkungen mit den aktuellen Methoden im ML aufzuzeigen. Besonders die auch für autonome Autos zentrale Frage nach den »high dimensional environments« wäre wissenschafts- und medienhistorisch darauf zu befragen,[51] wie in den diversen Experimental- und Simulationskulturen Kontingenz eingeführt und gesteuert wird. Die hier dargelegten Beispiele anhand des RL-Modells für autonome Autos haben auch verdeutlicht, dass die Modelle selbst eine soziale Dimension haben, die einmal einen wissenschaftshistorischen Ursprung hat und zum andern ein Verkehrsverständnis der individuellen Mobilität mitbefördern. Dabei stehen sich diese beiden Aspekte des wissenschaftshistorischen Ursprungs und der sozialen Implikation nicht getrennt gegenüber. Sie bedingen einander, zumal das Wissen des RL primär auf ein Selbst-Welt-Verhältnis ausgelegt und somit geeignet ist für Modelle automatisierter Fahrzeuge, die für oder mit ihrer/ihren Fahrzeughalter:innen entscheiden sollen.

50 Morgan, Mary S./Margaret Morrison (1999): »Models as mediating instruments«, in: dies. (Hg.): Models as mediators, New York: Cambridge University Press, S. 10-37.

51 Ravi Kiran et al.: »Deep Reinforcement«, S. 1.

6. Schluss

Ich habe dargelegt, dass der Übergang von Computerspielwelten in Verkehrswelten bei der Anwendung von ML-Verfahren mehr als eine neue technische Komplexitätsstufe impliziert. Das dabei formulierte Kernproblem, die adäquate zeitkritische Entscheidung und Vorhersage in unbekannten Umwelten durch autonome Autos zu optimieren, eröffnet ein epistemisches Feld, das nicht allein computerwissenschaftlicher Provenienz ist. Es referiert auf eine wissenschafts- und medienhistorisch nachhaltige Kultur der Verhaltensexperimente mit lebenden Wesen und den zahlreichen Modi der Umweltmodellierung. Zwei Punkte habe ich für das experimentelle Verhältnis von individueller/m Agent:in und Umwelt hervorgehoben, die den heutigen RL-Modellen für autonome Autos zugrunde liegen. Zum einen die Techniken zur Steuerung von Kontingenz in Experimenten und Simulationen, die stets auf einen sichtbaren, lerneffektiven Verlauf der Aktionen zurückbezogen werden müssen. Ohne diesen Rückbezug, also die Zuordnung von Zustands-und-Aktions-Pärchen, wird jede Vorhersagbarkeit über das Verhalten gefährdet. Gerade dieses Prädiktionsversprechen aber ist es, dass sich sowohl durch die experimentalpsychologischen Versuche als auch durch die computerwissenschaftlichen Modelle des RL zieht und damit nachweislich zeigt, dass das Problem der Zurechenbarkeit von Stimuli und Aktion in der Modellgeschichte des RL selbst angelegt ist und nicht erst ab einer bestimmten Komplexitätsstufe eintritt. Unabhängig von jeder Spielelogik oder Gamifizierung sind ML-Modelle wie das RL damit Produkte einer Experimentalkultur und ihrer inhärenten, induktiven Logik. Hinzu kommt der auf ein Selbst-Welt-Verhältnis reduzierte Kern des RL, der aus solchen Experimentalkulturen herzuleiten ist. Dieser hat nicht nur eine technische Implikation, wenn per Software das optimierte Verhalten eines autonomen Autos auf alle anderen übertragen werden kann, sondern auch eine soziale. Modelle wie das RL eignen sich besonders für eine individualisierte Mobilitätskultur, in der nicht nur Fahrer:innen, sondern auch Konsument:innen, Infrastrukturen und Markterwartungen zusammengehen. Die sozialen Gründe für die Attraktivität des ML speisen sich nicht aus ihren Erfolgen im Kontext von Computerspielen, sondern aus einem epistemischen Kern, der zutiefst in einer Kultur der Verhaltensexperimente und ihrer Episteme des Komparativs liegt. Autonome Autos als Anwendungen des ML sind damit repräsentativ für weitere Fallbeispiele, in denen die medien- und wissenschaftshistorische Bedeutung der Modelle und ihrer

Mediationen erst hinter den aufmerksamkeitskonzentrierenden technischen Verfahren freigelegt werden muss.