Referatsausarbeitung:

Ist künstliche Motivation gefährlich?

Marvin Schmitt

Vorgelegt im Rahmen der Veranstaltung
Ist künstliche Intelligenz gefährlich?

Leiter der Veranstaltung:
PD Dr. Ullrich Köthe

Kontaktadresse für Rückmeldungen und Rückfragen:

Marvin Schmitt
Brennerweg 44
69124 Heidelberg
E-Mail: marvin.schmitt@stud.uni-heidelberg.de

Inhaltsverzeichnis

1 Intelligenz als Prädiktor für Lebenserfolg 1

2 Motivation aus der Sicht der Psychologie 3

2.1 Motiv vs. Motivation 3

2.2 Aktivationstheorie nach Berlyne 4

2.3 Risiko-Wahl-Modell nach Atkinson 5

2.4 Leistungsmotiv in der Seminargruppe 5

3 Motivation in der KI-Forschung 7

3.1 Orthogonalitätsthese 8

3.2 Ethik und Verantwortung 10

4 Fazit und Ausblick 12

Literaturverzeichnis 14

Anhang A: Fragebogen im Seminar 16

„Das Problem ist gar nicht künstliche Intelligenz,
sondern künstliche Motivation.“ (Ullrich Köthe)

1 Intelligenz als Prädiktor für Lebenserfolg

In vielen Bereichen des Lebens ist Intelligenz von zentraler Bedeutung. Bereits in der Schule werden grundlegende kognitive Fähigkeiten gefordert und gefördert. Auch im Berufsleben werden intellektuelle Fähigkeiten verlangt, um den spezifischen Anforderungen des Berufs gerecht zu werden. Es wird allgemein angenommen, dass Intelligenz notwendig ist, um sich wirkungsvoll mit seiner Umwelt auseinandersetzen zu können und Probleme in verschiedensten Bereichen erfolgreich lösen zu können. Was Intelligenz jedoch genau ist, wird in der psychologischen Forschung nach wie vor kontrovers diskutiert. Da Intelligenz ein künstliches Konstrukt ist, bedarf es einer Definition. Innerhalb der zahlreichen Intelligenzdefinitionen unterscheidet man unter anderem zwischen verbalen und operationalen Definitionen. Die prominenteste verbale Definition innerhalb der psychologischen Forschung stammt von einer Taskforce der American Psychological Association (APA):

> *„[Intelligence is the] ability to understand complex ideas, to adapt effectively to the environment, to learn from experience, to engage in various forms of reasoning, to overcome obstacles by taking thought.“* (Neisser et al., 1996, S.1)

Im Gegensatz dazu stehen operationale Definitionen. So definiert beispielsweise Boring Intelligenz als „das, was ein Intelligenztest misst.“ (Boring, 1923)

Es stellt sich jedoch unweigerlich die Frage, ob die Leistung in einem Intelligenztest tatsächlich alles ist, was für die Entwicklung von Agenten[1] von Interesse ist. Um ein Individuum nicht nur hinsichtlich seines Intelligenzquotienten zu beurteilen, wurde bereits im frühen 20. Jahrhundert der Begriff des *Erfolgs* in der Psychologie eingeführt. Erfolg, oder auch *Lebenserfolg*, bezeichnet dabei die Umsetzung von Fähigkeiten in tatsächliche Handlungen. Im Rahmen der sogenannten Terman-Studie

[1] Als Agent wird ein System bezeichnet, das eigenständig handelt und Handlungen ausführt, die über vorher festgelegte Programmabläufe hinausgeht: „ […] act on their own and learn to take actions over choice of computation“ (gwern.net, 2017, Abs. 1)

untersuchte Lewis Terman die intelligentesten 1% der US-amerikanischen Schulkinder. Dazu wurden US-amerikanische Lehrerinnen und Lehrer gebeten, das jüngste, älteste, klügste und zweitklügste Kind jeder Klasse zu benennen, die daraufhin hinsichtlich ihrer intellektuellen Leistungsfähigkeit untersucht wurden. Mit dieser Selektionsmethode wurden schließlich 1528 Kinder in die Studie eingeschlossen (Anastasi, 1976). Diese Kinder beschrieb Terman als Genies. Simonton definiert ein Genie als „eine Person, die einen außergewöhnlich hohen Intelligenzquotienten (IQ), typischerweise über 140, hat" (Simonton, 2016, S. 3). Terman suchte in seiner Untersuchung mithilfe Historiometrischer Analysen[2] nach Prädiktoren für Lebenserfolg.

Dabei identifizierte er mithilfe von Faktorenanalysen die drei Sekundärfaktoren *Soziale Verantwortung*, *Soziabilität* und *Intellektualität*. Intelligenz zeigte sich dabei nur als eine Unterfacette des Faktors *Intellektualität*, während auf der Ebene der Unterfacetten die Variable *familiäre Harmonie* die meiste Varianz an Lebenserfolg aufklärte. Eine Analyse von Cox (1926) zeigt ähnliche Ergebnisse: Intelligenz kläre demnach nur 10% der Varianz von Lebenserfolg auf, weshalb die Suche nach weiteren Prädiktoren für Erfolg notwendig sei. Cox bezeichnet Intelligenz im Folgenden als ein notwendiges, aber nicht hinreichendes Kriterium für Erfolg bzw. Eminenz und rückt gleichzeitig die Rolle von Motivation und Ausdauer in den Vordergrund:

> *„[…] that high but not the highest intelligence, combined with the greatest degree of [motivational] persistence, will achieve greater eminence than the highest degree of intelligence with somewhat less [motivational] persistence."* (Cox, 1926, zitiert nach Simonton, 2016)

Diese Art von Motivation bzw. Ausdauer wird in der modernen psychologischen Forschung als *Grit* bezeichnet und besitzt über Intelligenz hinaus inkrementelle Validität zur Vorhersage des Erreichens langfristiger Ziele (Duckworth, Peterson, Matthews & Kelly, 2007). Der zugrundeliegende Mechanismus ist jedoch bisher noch unklar und wird mithilfe von Mediatoranalysen[3] weiterhin beforscht.

---

[2] Eine Historiometrische Analyse ist eine idiographische Methode, um Leistungs- und Charaktermerkmale herausragender Persönlichkeiten zu quantifizieren. Dies erfolgt unter Rückgriff auf deren biographisches Material.

[3] Im Rahmen einer Mediatoranalyse wird die Hypothese geprüft, ob der Zusammenhang zwischen zwei Variablen durch eine Drittvariable *vermittelt* wird. In der Praxis gibt dies Hinweise auf einen dahinterliegenden Mechanismus. Ein statistisch signifikanter Mediationseffekt ist jedoch nicht als Beweis für Kausalität zu sehen.

2 Motivation aus der Sicht der Psychologie

Die psychologische Motivationsforschung legt den Fokus auf Motive, die dem Handeln zugrunde liegen, sowie deren Realisierung in konkreten Situationen. Sie geht damit über den allgemeinsprachlichen Motivationsbegriff im Sinne von „sehr motiviert sein“ hinaus und trifft beispielsweise keine Unterscheidung zwischen „guter“ und „schlechter“ Motivation. Die psychologische Motivationsforschung unterscheidet dabei zum Beispiel zwischen Motiv und Motivation. Ein weiterer zentraler Aspekt ist der Unterschied von Personismus, Situationismus und Interaktionismus, worauf im Rahmen der vorliegenden Arbeit jedoch nicht näher eingegangen wird.

2.1 Motiv vs. Motivation

Motive sind „überdauernde Vorlieben einer Person, die sich auf inhaltliche Klassen von Handlungszielen beziehen“ (Heckhausen, 1989, 16f.) und stellen ein Bindeglied zwischen der biologischen Natur von Menschen und deren konkreten Verhaltensweisen dar. Damit können sie im Sinne eines *backward engineering* einen Schluss von offenen Verhaltensweisen auf die zugrundeliegenden inneren Zustände erlauben, der andernfalls nicht möglich wäre. Somit liefern Motive einen Erklärungsansatz für interindividuelle Unterschiede im Verhalten. Motive fungieren darüber hinaus im Bereich der forensischen Psychologie als Grundlage dafür, Menschen die Verantwortung für ihre Taten zuschreiben zu können.

Im Gegensatz dazu bezeichnet Motivation die Aktualisierung eines Motivs in einer konkreten Situation. Damit mündet Motivation in der Regel im (Nicht-)Ausführen einer Handlung. Im Sinne eines *Approach-Avoidance-Ansatzes* kann Motivation als ein Indikator für den Zustand eines Organismus gesehen werden, ein bestimmtes Ziel aufzusuchen oder zu vermeiden. In der psychologischen Forschung unterscheidet man verschiedene Motivationen hinsichtlich ihrer Wahl, Intensität, Latenz und Persistenz.

Im Rahmen der psychologischen Motivationsforschung wurde die Frage, ob ein Motiv zwangsweise zur Motivation führt, intensiv beforscht. Nach Rheinberg et al. (2004) hängt dies vor allem von der Situation und deren potentiellen Anreizen ab. Ob eine Motivation auch zwangsweise in der entsprechenden Handlung mündet, wird unter dem Begriff *behavior-intention-gap* beforscht, im Rahmen der vorliegenden Arbeit jedoch nicht näher thematisiert.

2.2 Aktivationstheorie nach Berlyne

Im Rahmen seiner *Aktivationstheorie* beschreibt Berlyne, wie Menschen reagieren, wenn sie bestimmten Stimuli ausgesetzt sind. Dazu führt er den Begriff der *Aktivation* ein, die als *kognitive Erregung* oder *Arousal*, also der allgemeinen Aktiviertheit des zentralen Nervensystems, verstanden wird. Berlyne legt der *Aktivationstheorie* zwei Prämissen zugrunde:

(1) *Niedrige Aktivation* führt zu *hoher Attraktivität*
(2) *Mittlere Komplexität* führt zu *niedriger Aktivation*

Daraus leitet er die Konklusion ab, dass mittlere Komplexität zu hoher Attraktivität führt (siehe Abbildung 1).

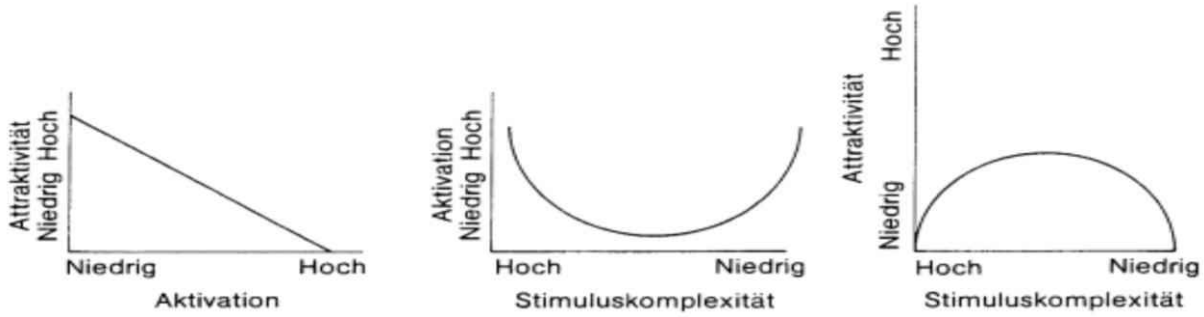

Abbildung 1. Aktivationstheorie nach Berlyne. Aus: Weiner (1994).

Mit dieser theoretischen Grundlage erklärt Berlyne zwei Arten von Neugierverhalten, diversives und spezifisches Neugierverhalten. Sie unterscheiden sich im Grad des initialen Reizeinstroms und haben gemeinsam, dass das Individuum stets einen mittleren Reizeinstrom anstrebt (siehe Abbildung 2).

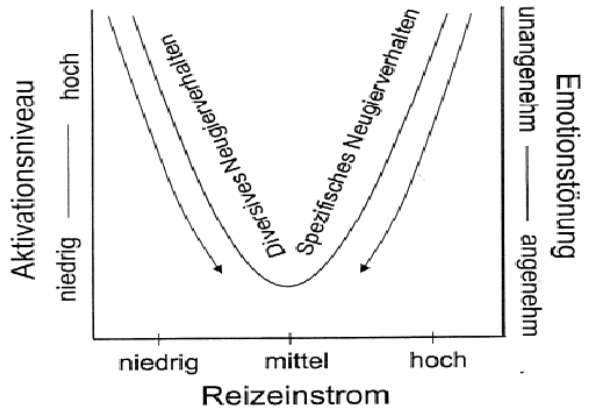

Abbildung 2. Erklärung von Neugierverhalten nach Berlyne. Aus: Weiner (1994)

Diversives Neugierverhalten zeichnet sich demnach dadurch aus, dass das Individuum den Reizeinstrom künstlicherhöht. Das kann beispielsweise dadurch geschehen, dass dem Stimulus eine tiefere Sinnhaftigkeit zugeschrieben wird oder künstlich weitere Stimuli aufgesucht werden, wie beim schnellen Umschalten der Fernsehkanäle („Zappen“). Bei spezifischem Neugierverhalten soll der Reizeinstrom hingegen verringert werden. Dies kann dadurch geschehen, dass sich das Individuum auf einzelne Komponenten des Stimulus, wie beispielsweise eines komplexen Musikstücks, konzentriert – dieser Prozess wird auch als Exploration bezeichnet. Ein Beispiel dafür ist die Arbeit mit einem Kursbuch, das vom Individuum in seiner Komplexität überhaupt nicht ganzheitlich erfasst werden kann. In diesem Fall fällt der Fokus auf den relevanten Bereich – z.B. ein ausgewähltes Kapitel –, während die restlichen Aspekte des Kursbuchs ausgeblendet werden. Durch diese zielgerichtete Fokussierung wird der Reizeinstrom reduziert, was das spefizische Neugierverhalten nach Berlyne charakterisiert.

2.3 Risiko-Wahl-Modell nach Atkinson

Während der Fokus bei der *Aktivationstheorie* nach Berlyne auf dem Umgang mit einem bestimmten Set an Stimuli liegt, rückt Atkinson (1964) im Rahmen seines *Risiko-Wahl-Modells* die Motive des Individuums stärker in den Vordergrund. Er unterscheidet im Sinne der *Motive Disposition Theory* nach McClelland (1953) auf einer ersten Ebene drei Motive: *Machtmotiv*, *Intimitätsmotiv* und *Leistungsmotiv*. Nach Atkinson besteht das Leistungsmotiv wiederum aus zwei Tendenzen:

(1) Der *Tendenz, Erfolg zu suchen* (T_E), die sich multiplikativ zusammensetzt aus dem *Erfolgsmotiv* (M_E), der *subjektiven Erfolgswahrscheinlichkeit* (W_E) und dem *Anreiz von Erfolg* ($A_E = 1 – W_E$).

(2) Der *Tendenz, Misserfolg zu meiden* (T_M), die sich wiederum multiplikativ zusammensetzt aus dem *Misserfolgsmotiv* (M_M), der *subjektiven Misserfolgswahrscheinlichkeit* (W_M) und dem *Anreiz von Misserfolg* ($A_M = 1 – W_M$).

2.4 Leistungsmotiv in der Seminargruppe

Um die Grundlagen der psychologischen Motivationsforschung anzuwenden und dadurch zu festigen, wurde im Rahmen des Seminars eine kurze Studie zur Vertiefung des *Risiko-Wahl-Modells* durchgeführt. Den Kursteilnehmenden sollte durch die Verknüpfung von Theorie und Praxis die Anwendung psychologischer Modelle

auf zentrale Fragen der KI-Forschung erleichtert werden. Die Methoden und Ergebnisse der Erhebung werden im Folgenden beschrieben.

Ziel war die Erfassung des *Leistungsmotivs* innerhalb der Seminargruppe. Dabei wurden mithilfe der Kurzform der *Achievement Motives Scale* (AMS) nach Engeser (2005) die beiden Tendenzen T_E und T_M nach Atkinson erfasst und daraufhin zu einer *Resultierenden Tendenz RT* zusammengefasst. Die Items der Skalen sind im Folgenden tabellarisch aufgeführt (siehe Tabelle 1). Der komplette Fragebogen ist in Anhang A aufgeführt.

Tabelle 1. Items der Skalen T_E und T_M.

Tendenz, Erfolg zu suchen (T_E)	**Tendenz, Misserfolg zu meiden (T_M)**
1. Es macht mir Spaß, an Problemen zu arbeiten, die für mich ein bisschen schwierig sind.	1. Es beunruhigt mich, etwas zu tun, wenn ich nicht sicher bin, dass ich es kann.
2. Probleme, die schwierig zu lösen sind, reizen mich.	2. Wenn eine Sache etwas schwierig ist, hoffe ich, dass ich es nicht machen muss, weil ich Angst habe, es nicht zu schaffen.
3. Mich reizen Situationen, in denen ich meine Fähigkeiten testen kann.	3. Dinge, die etwas schwierig sind, beunruhigen mich.
4. Ich mag Situationen, in denen ich feststellen kann, wie gut ich bin.	4. Auch bei Aufgaben, von denen ich glaube, dass ich sie kann, habe ich Angst zu versagen.
5. Ich möchte gern vor eine etwas schwierige Arbeit gestellt werden.	5. Wenn ich ein Problem nicht sofort verstehe, werde ich ängstlich.

Die Items entstammen der Kurzform der Achievement Motives Scale (AMS) nach Engeser (2005).

Die Antworten wurden mithilfe einer vierstufigen Likert-Skala (Stimme gar nicht zu – Stimme eher nicht zu – Stimme eher zu – Stimme völlig zu) erfasst und entsprechend itemweise in Werte von 1 bis 4 transformiert. Anschließend wurden die Skalenwerte T_E und T_M durch arithmetische Mittel der jeweiligen Itemscores gebildet.

Innerhalb der 19 Teilnehmenden war die *Tendenz, Erfolg zu suchen* hoch ausgeprägt ($M = 3.30$, $SD = 0.34$) und folgte annähernd einer Normalverteilung[4]. Die *Tendenz, Misserfolg zu meiden* fiel im Mittel niedriger aus, streute innerhalb der Stichprobe stärker ($M = 2.28$, $SD = 0.60$) und war nicht normalverteilt. Um zu prüfen, ob sich T_E und T_M innerhalb der Gruppe unterscheiden, wurde ein Ein-Gruppen t-Test gegen 0 mit der Testvariablen RT durchgeführt, die als Differenz von T_E und T_M definiert ist. Die Ergebnisse stützen die These, dass die beiden Tendenzen nach Atkinson (1964) innerhalb der Seminargruppe unterschiedlich ausgeprägt sind, $t(18)$

[4] Die Verteilung wurde mithilfe eines Kolmogorov-Smirnov-Tests auf Normalverteiltheit geprüft.

< .001. Die beobachtete unterschiedliche Ausprägung der Tendenzen ist mit den Erkenntnissen der psychologischen Motivationsforschung konsistent.

3 Motivation in der KI-Forschung

Bei der Frage nach der Schaffung künstlicher Motivation stellt sich zunächst die Frage, wie Motivation beim Menschen entsteht. Maslow (1943) stellt im Rahmen seiner Bedürfnistheorie die These auf, dass menschliche Bedürfnisse hierarchisch aufgebaut sind. Dabei unterscheidet er sechs aufsteigende Stufen der sogenannten *Bedürfnispyramide* (siehe Abbildung 3). Um ein höheres Bedürfnis befriedigen zu können, müssen zunächst die niederen Bedürfnisse gestillt sein. Auf der untersten Hierarchieebene stehen dabei *physiologische Bedürfnisse*, gefolgt von *Sicherheitsbedürfnissen*, *Sozialen Bedürfnissen*, *Individualbedürfnissen* und schließlich *Selbstverwirklichung* an der Spitze der Bedürfnispyramide.

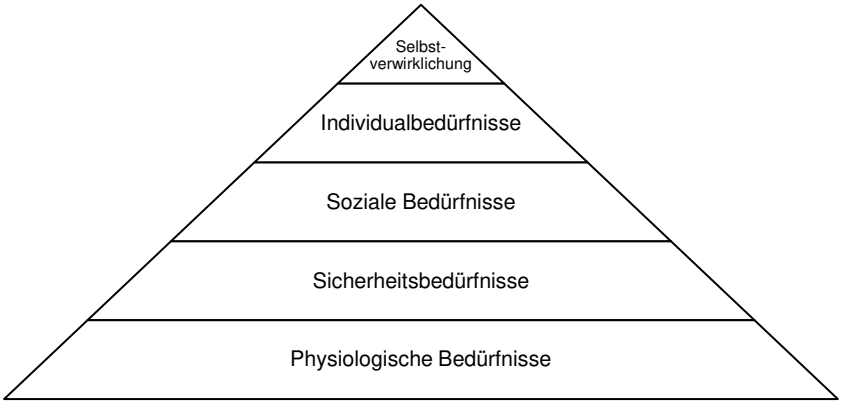

Abbildung 3. Bedürfnispyramide nach Maslow (1943).

Nach Maslow (1943) kann sich Motivation[5] nur auf der höchsten Ebene, der *Selbstverwirklichung*, bilden. Dafür müssen also nach dieser Theorie alle niederen Bedürfnisse erfüllt sein, darunter *soziale Bedürfnisse*. Hier drängt sich jedoch unweigerlich die Frage auf, ob die sozialen Bedürfnisse agierender künstlicher Intelligenzen (Agenten) überhaupt befriedigt werden können. Wenn dies nämlich nicht der Fall ist, kann nach Maslow keine Selbstverwirklichung und damit auch keine eigene Motivation erreicht werden.

[5] Der hier verwendete Motivationsbegriff bezieht sich auf die Terminologie nach Maslow, sie sich von dem Motivationsbegriff nach Rheinberg (2004, siehe Abschnitt 2.1) unterscheidet. Motivation bezeichnet hier nicht die Aktualisierung eines Motivs, sondern die Bildung eigenständiger Ziele, die das Individuum daraufhin verfolgt.

Da diese Überlegungen stark dem Stil der klassischen Psychologie verpflichtet sind und nicht unverändert auf Agenten angewendet werden können, soll im Folgenden ein Ansatz beleuchtet werden, der aus der Forschung zu künstlicher Intelligenz entspringt – die Orthogonalitätsthese.

3.1 Orthogonalitätsthese

In ihrer strikten Form besagt die Orthogonalitätsthese, dass die Ausprägung von Intelligenz und Motivation (Zielen) voneinander unabhängig sei. Demnach seien alle Kombinationen von Motivation und Intelligenz möglich. Armstrong (2013) zeigt jedoch einige definitorische Limitationen der strikten Formulierung auf. So seien beispielsweise einige Ziele mit dem Intelligenzzustand des Agenten unvereinbar („Ich will weniger intelligent sein!“). Außerdem könnten Ziele so komplex sein, dass sie die Intelligenz des Agenten lähmen, da die bloße Beschreibung des Ziels mehr Ressourcen beanspruchen würde als auf der Erde vorhanden seien. Darüber hinaus könnten im Rahmen der Orthogonalitätsthese nur statische Beobachtungen getroffen werden – eine Beobachtung von Entwicklungsverläufen sei nicht möglich. Auf Basis dieser Limitationen wird eine weniger strikte Formulierung der Orthogonalitätsthese vorgeschlagen:

> *„The fact of being of high intelligence provides extremely little constraint on what final goals an agent could have (as long as these goals are of feasible complexity, and do not refer intrinsically to the agent's intelligence).“* (Armstrong, 2013, S. 6)

Auf den Ausführungen von Armstrong (2013) aufbauend, lassen sich drei Ausprägungen von Zielsystemen intelligenter Agenten unterscheiden:

(1) Hohe Motivation zu positiven Handlungen
(2) Keine hohe Motivation zu positiven oder negativen Handlungen (neutral)
(3) Hohe Motivation zu negativen Handlungen

Dabei liegt der Fokus stets auf dem Bereich hoher Intelligenz, weil sich intelligente Agenten hinsichtlich ihrer kognitiven Fähigkeiten per definitonem stets dort einordnen lassen (und nicht im Bereich niedriger Intelligenz). Aus den drei genannten

Ausprägungen leiten sich Handlungskonsequenzen zum Umgang mit einem entsprechenden Agenten ab, die im Folgenden näher erläutert werden sollen (siehe Abbildung 4).

Ein intelligenter Agent mit einer hohen Motivation, etwas Positives zu tun, ist wünschenswert. Ein solcher Agent könnte beispielsweise das Ziel haben, ein Malaria-Heilmittel zu entwickeln. Die Konsequenz, die sich aus dieser Kombination an Intelligenz und Motivation ergibt, ist, dass das Verhalten des Agenten aufrechterhalten und gefördert werden soll.

Ein intelligenter Agent, der keine starke Motivation hat, etwas Positives oder Negatives zu tun, kann als neutral angesehen werden. Ein solcher „neutraler" Agent könnte beispielsweise das Ziel haben, in extraterrestrischen menschenfeindlichen Umgebungen zu überleben. Die Konsequenz aus dieser Kombination von Intelligenz und Motivation ist, das Verhalten des Agenten weiterhin zu kontrollieren.

Ein intelligenter Agent mit einer hohen Motivation, etwas Negatives zu tun, ist eine Gefahr für sein Umfeld. Das Ziel eines solchen Agenten könnte sein, eine für Menschen unverständliche Sprache zu entwickeln, um sie somit zu überlisten und schließlich zu unterwerfen. Diese Kombination von Intelligenz und Motivation ist als höchst kritisch anzusehen und bedarf eines unmittelbaren Eingriffs. Armstrong (2013) führt jedoch an, dass ein superintelligenter Agent mit einer hohen Motivation, Negatives zu tun, jedem Menschen intellektuell überlegen sei. Deshalb müsse schon im Vorfeld sichergestellt werden, dass sich eine solche Motivation, Bösartiges zu tun, gar nicht erst bilden kann.

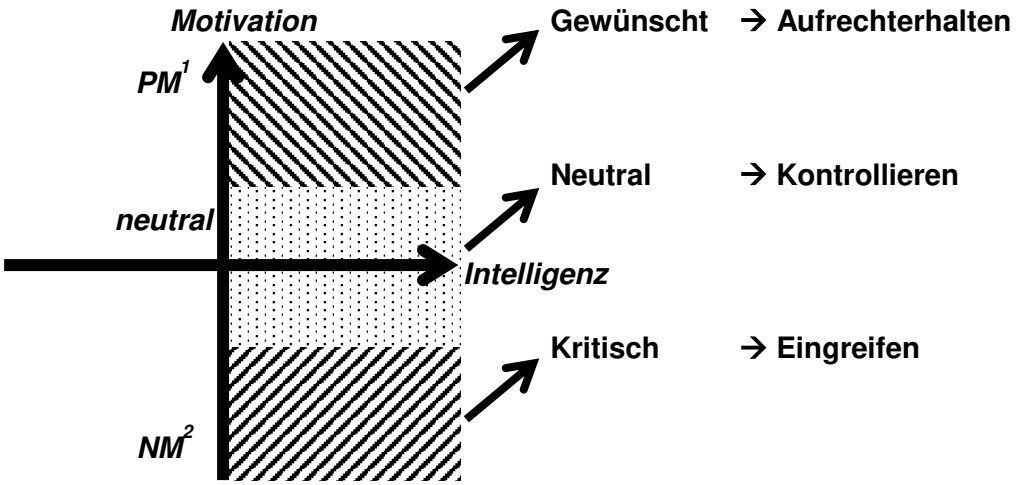

Abbildung 4. Verschiedene Ausprägungen der Ziele intelligenter Agenten.
[1] PM: Hohe Motivation, etwas Positives zu tun.
[2] NM: Hohe Motivation, etwas Negatives zu tun.

Die bloße Tatsache, dass ein Agent (super-)intelligent ist, führt folglich nicht dazu, dass er Ziele verfolgt, die als moralisch bezeichnet werden können (vgl. Armstrong, 2013). Armstrong formuliert daraus den Appell, dass es notwendig sei, die finalen Ziele eines Agenten direkt einzuprogrammieren. Wenn das nicht möglich sei, müsse es zumindest möglich sein, sie zu jedem beliebigen Zeitpunkt in Erfahrung zu bringen. Jedoch stellt sich die Frage, ob man überhaupt versuchen sollte, diese finalen Ziele einzuprogrammieren – schließlich lasse sich ein superintelligenter Agent ohnehin nicht kontrollieren und könnte jeden Menschen intellektuell überlisten (Armstrong, 2013).

3.2 Ethik und Verantwortung

Welche Rechte, Pflichten und Verantwortung etwas zugesprochen wird, hängt zunächst davon ab, ob „es“ lebendig ist. Um die Frage zu beantworten, ob intelligente Agenten Lebewesen sind, kann man schrittweise prüfen, ob sie die fünf Kriterien, die ein Lebewesen im biologischen Sinne charakterisieren, erfüllen (Koops, 2016):

(1) Reizbarkeit
(2) Fortpflanzung
(3) Eigener Stoffwechsel
(4) Wachstum und Entwicklung
(5) Beweglichkeit

Reizbarkeit. Mithilfe von Sensoren können Maschinen bereits heute Sinneseindrücke wahrnehmen und daraufhin verarbeiten.

Fortpflanzung. Während sich Computerviren über verschiedene Geräte „fortpflanzen“ können und Rechner Daten generieren können, ist mit Fortpflanzung im biologischen Sinne die eigenständige materielle Reproduktion gemeint. Dieses Kriterium kann kontrovers diskutiert werden – bereits heute können Roboter Abbilder ihrer selbst herstellen, allerdings benötigen sie dafür nach wie vor Materialien, die von Menschen zur Verfügung gestellt und entsprechend angeordnet werden.

Eigener Stoffwechsel. Die Umsetzung von Strom in Rechenkraft und Wärme entspricht keinem Stoffwechsel im biologischen Sinne, weshalb dieses Kriterium als nicht erfüllt angesehen werden kann.

Wachstum und Entwicklung. Im Bereich des *Machine Learning* wird bereits heute eindrucksvoll gezeigt, dass intellektuelle Entwicklung von Maschinen möglich ist. Nichtsdestotrotz ist kein physisches Wachstum im biologischen Sinne möglich, weshalb auch dieses Kriterium nur als eingeschränkt erfüllt gesehen werden kann.

Beweglichkeit. Wenngleich sich Agenten durch die Montage von Robotikelementen bewegen können, ist dies keine Funktion, die der Maschine immanent ist. Deshalb ist dieses Kriterium mit kleinen Einschränkungen als gegeben zu sehen.

Zusammenfassend lässt sich also festhalten, dass intelligente Agenten nur einige der fünf biologischen Kriterien von Leben erfüllen. Damit können sie, nach heutigem Stand der Überlegungen, nicht als Lebewesen bezeichnet werden.

Dass ein Agent nicht lebendig, wohl aber intelligent ist, wirft völlig neuartige ethische Fragen auf. Bisher gibt es lediglich drei der vier möglichen Kombinationen von Lebendigkeit und Intelligenz, wobei „intelligent“ ein menschenähnliches Niveau an Intelligenz beschreibt:

(1) Lebendig und intelligent. Menschen sind sowohl intelligent als auch lebendig, für sie existieren Menschenrechte und Konventionen, um ihre Würde zu wahren.

(2) Lebendig und nicht intelligent. Tiere sind lebendig, aber deutlich weniger intelligent als Menschen. Der Mangel an Intelligenz wird oft als Grund dafür genannt, Tiere als Nutztiere und Haustiere zu halten sowie zu töten. Nichtsdestotrotz gibt es in vielen Ländern Gesetze zum Schutz von Tieren, die beispielsweise Tierquälerei unter Strafe stellen. Auf die Willkürlichkeit und Unzulänglichkeit dieses Arguments zur Rechtfertigung von Massentierhaltung u.Ä. wird an dieser Stelle nicht weiter eingegangen.

(3) Nicht lebendig und nicht intelligent. Objekte sind weder lebendig noch intelligent. Für sie gibt es keine Gesetze oder Konventionen, was damit begründet wird, dass sie nicht lebendig sind.

Das postulierte Spannungsfeld besteht darin, dass intelligente Agenten die vierte Kombinationsmöglichkeit darstellen: Intelligent und nicht lebendig. Da sie intelligent sind, müsste ihnen nach (2) derselbe Schutz zukommen wie Menschen. Das würde beispielsweise einer Ausweitung des Artikel 1 des Grundgesetzes entsprechen: „Die Würde des Menschen und des intelligenten künstlichen Agenten ist unantastbar.“ – was zunächst befremdlich erscheint, kann aus ethischer Sicht durchaus kontrovers diskutiert werden. Allerdings sind intelligente Agenten nicht lebendig, was

nach (3) zur Folge hätte, dass sie als nicht schützenswert zu betrachten wären. Welche der Positionen zu den Rechten intelligenter Agenten sich durchsetzt, bedarf weiterer Beforschung und Diskussion – und zwar vor der Entwicklung erster intelligenter Agenten.

4 Fazit und Ausblick

Um die Fragen zu beantworten, die bei der Forschung zu künstlicher Intelligenz aufkommen, reichen technische Überlegungen allein nicht aus. Mit der Erschaffung eine künstlicher Intelligenz geht eine große Verantwortung einher, da sie neben zahlreichen Chancen auch Risiken birgt. Die Gefahr, die von einem „bösartigen“ Agenten ausgeht, ist immens. Deshalb ist es unerlässlich, die psychologischen Aspekte der KI-Forschung zu beleuchten.

Dabei wird schnell deutlich, dass Intelligenz allein nicht ausreicht, um alle Prozesse und Fähigkeiten, die wir von einem intelligenten Agenten erwarten, zu erklären. Es ist vielmehr notwendig, weitere Bereiche wie zum Beispiel Motivation oder Ausdauer in die Forschung aufzunehmen. Aus der psychologischen Intelligenz- und Motivationsforschung gehen eine Vielzahl von Modellen menschlicher Motivation hervor, die den Fokus jeweils auf bestimmte Aspekte motivationsassoziierten Verhaltens legen. Die Übertragung dieser psychologischen Modelle auf intelligente Agenten ist umstritten und in vielen Fällen nicht ohne Weiteres möglich. Nichtsdestotrotz drängen sich Fragen auf, die die Entwicklung entsprechender Intelligenz- und Persönlichkeitsmodelle für intelligente Agenten erfordern. Deshalb ist es besonders wichtig, psychologische Fragestellungen zu künstlicher Intelligenz zu beforschen und die entsprechenden Konsequenzen für die KI-Forschung zu ziehen.

Die ethischen Fragen, die mit der Entwicklung künstlicher Intelligenz verbunden sind, sind weitestgehend unbeforscht. Intelligente Agenten ziehen ein Spannungsfeld zwischen Intelligenz und Leben auf, das neuartig ist und weiterer Forschung bedarf. Während bisher vor allem über die Pflichten intelligenter Agenten diskutiert wird, stellt sich im Umkehrschluss auch die Frage nach deren Rechten.

Abschließend lässt sich sagen, dass die Forschung zu künstlicher Intelligenz eine Schnittstelle zwischen Natur- und Geisteswissenschaften darstellt, in dessen Rahmen Forscherinnen und Forscher verschiedenster Disziplinen in den Dialog treten und zusammenarbeiten müssen. Dies ist für eine ganzheitliche und vor allem si-

chere Beforschung künstlicher Intelligenz unumgänglich. In der heutigen KI-Forschung nehmen diese interdisziplinären Fragestellungen nur einen kleinen Raum ein, da der technologische Fortschritt im Fokus der Forschung steht und mit entsprechenden Geldern gefördert wird. Die Auseinandersetzung mit den ethischen und psychologischen Implikationen einer künstlichen Intelligenz mag trivial und nicht dringlich erscheinen, weil die aktuellen Agenten noch nicht weit genug entwickelt sind. Wenn dieser technologische Schritt jedoch vollbracht ist und unsere Maschinen auf einmal einen eigenen Willen entwickeln, sollten die Diskussionen darüber, wie wir damit umgehen, bereits abgeschlossen sein:

> *"Presumably, these agents are still too primitive to have any moral status. But how confident can we really be that this is so? More importantly, how confident can we be that we will know to stop in time, before our programs become capable of experiencing morally relevant suffering?"* (Bostrom, 2014, S. 234)

Literaturverzeichnis

Anastasi, A. (1976). *Differentielle Psychologie. Unterschiede im Verhalten von Individuen und Gruppen* (Beltz-Studienbuch, Bd. 102). Weinheim: Beltz.

Armstrong, S. (2013). *General Purpose Intelligence. Arguing the Orthogonality Thesis,* Oxford.

Atkinson, J. W. (1964). *An introduction to motivation* (The university series in psychology). New York: Van Nostrand.

Boring, E. G. (1923). Intelligence as the Tests Test It. *New Republic, 36,* 35-37.

Bostrom, N. (2014). *Superintelligence. Paths, dangers, strategies* (1. ed.). Oxford: Oxford University Press.

Cox, C. (1926). *The early mental traits of 300 geniuses*. Stanford: Stanford University Press.

Duckworth, A. L., Peterson, C., Matthews, M. D. & Kelly, D. R. (2007). Grit. Perseverance and passion for long-term goals. *Journal of personality and social psychology, 92* (6), 1087-1101.

Engeser, S. (2005). *Messung des expliziten Leistungsmotivs. Kurzform der Achievement Motives Scale,* Potsdam.

Gwern.net. (2017). *Why Tool AIs want to be Agent AIs*. Verfügbar unter https://www.gwern.net/Tool%20AI

Heckhausen, H. (1989). *Motivation und Handeln* (Springer-Lehrbuch, Zweite, völlig überarbeitete und ergänzte Auflage). Berlin: Springer.

Koops, M. (2016). *Leben. Kennzeichen des Lebens*. Verfügbar unter http://www.biologie-lexikon.de/lexikon/leben.php

Maslow, A. H. (1943). A Theory of Human Motivation. *Psychological Review* (50), 370-396.

McClelland, D. C. (1953). *The achievement motive* (Century psychology series). New York: Appleton-Century-Crofts.

Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A. W., Brody, N., Ceci, S. J. et al. (1996). Intelligence. Knowns and Unknowns. *American Psychologist, 51* (2), 77-101.

Rheinberg, F., Selg, H. & Salisch, M. v. (2004). *Motivation* (Urban-Taschenbücher, Bd. 555, 5., überarb. und erw. Aufl.). Stuttgart: Kohlhammer.

Simonton, D. K. (2016). Reverse engineering genius. Historiometric studies of superlative talent. *Annals of the New York Academy of Sciences, 1377* (1), 3-9.

Weiner, B. (1994). *Motivationspsychologie* (3. Aufl.). Weinheim: Beltz Psychologie-Verl.-Union.

Umfrage im Seminar
„Ist künstliche Intelligenz gefährlich?"

Ich möchte für mein Referat eine kurze Umfrage zu verschiedenen Motiven durchführen. Die Teilnahme ist freiwillig. Die Daten werden selbstverständlich vertraulich behandelt und sind anonym. Ich würde mich über deine Teilnahme sehr freuen. Vielen Dank!

Bitte gib an, inwiefern du den jeweiligen Aussagen zustimmst. Dabei gibt es kein richtig oder falsch! Antworte ehrlich aus dem Bauch heraus, ohne viel nachzudenken.

	Stimme gar nicht zu	**Stimme eher nicht zu**	**Stimme eher zu**	**Stimme völlig zu**
Es macht mir Spaß, an Problemen zu arbeiten, die für mich ein bisschen schwierig sind.				
Es beunruhigt mich, etwas zu tun, wenn ich nicht sicher bin, dass ich es kann.				
Probleme, die schwierig zu lösen sind, reizen mich.				
Mich reizen Situationen, in denen ich meine Fähigkeiten testen kann.				
Wenn eine Sache etwas schwierig ist, hoffe ich, dass ich es nicht machen muss, weil ich Angst habe, es nicht zu schaffen.				
Dinge, die etwas schwierig sind, beunruhigen mich.				
Auch bei Aufgaben, von denen ich glaube, dass ich sie kann, habe ich Angst zu versagen.				
Ich mag Situationen, in denen ich feststellen kann, wie gut ich bin.				
Ich möchte gern vor eine etwas schwierige Arbeit gestellt werden.				
Wenn ich ein Problem nicht sofort verstehe, werde ich ängstlich.				