Institut für Soziologie
Westfälische Wilhelms-Universität Münster
Bachelorarbeit im Sommersemester 2020
Erstprüferin: Frau Dr. Gallina Tasheva
Zweitprüfer: Herr Dr. Matthias Grundmann

# Normative Universalien für intelligente Maschinen: Gibt es universelle moralische Werte, an denen Forscher künstliche Intelligenz ausrichten können?

# Normative universals for intelligent machines: Are there universal values by which researchers can align Artificial Intelligence?

*Münster, den 29.07.2020*

Vorgelegt von:
Lara Lawniczak
Geburtsort: Osnabrück
Scharnhorststraße 73
48151 Münster

Matrikelnummer: 440813
Zwei-Fach-Bachelor Kommunikationswissenschaft und Soziologie
8. Fachsemester

## Inhaltsverzeichnis

# 1. Einleitung

Künstliche Intelligenz ist nicht mehr nur eine kreative Idee aus Science-Fiction Romanen oder ein vages Ziel von realitätsfremden Computerwissenschaftlern[1], sondern durchdringt längst viele Bereiche unseres Alltags. Egal ob in Form von digitalen Assistenten, selbstlernenden Algorithmen zur Text- und Bilderkennung, oder OP- und Pflegeassistenzrobotern – intelligente Technologien kommen in fast allen Branchen auf unterschiedlichste Weise zum Einsatz und gelten als Wachstumsmotor und Schlüsseltechnologie der Zukunft. Laut einer Studie von Bughin und Kollegen vom McKinsey Global Institute (MGI) werden 70 Prozent aller Unternehmen bis 2030 mindestens eine Form der künstlichen Intelligenz adaptiert haben (Bughin et al., 2018, S. 2). Zudem prognostiziert das MGI, dass künstliche Intelligenz das globale Bruttoinlandsprodukt bis 2030 um durchschnittlich 1,2 Prozentpunkte pro Jahr steigern wird: Ein jährlicher Wachstumseffekt, der sogar den der Dampfmaschine während der industriellen Revolution übertrifft (Bughin et al., 2018, S. 3).

Die intelligenten Systeme versprechen nicht nur vielfältige ökonomische Vorteile, sondern bieten auch großes Potenzial für Bereiche, in denen menschliches Versagen Schaden verursachen kann, wie zum Beispiel in der medizinischen Diagnostik und Behandlung oder im Straßenverkehr (Mannino et al., 2015, S. 1). Sobald eine künstliche Intelligenz jedoch kognitive Aufgaben mit weitreichenden sozialen Folgen übernimmt, die zuvor ausschließlich von Menschen verrichtet wurden, muss sie neben den technischen auch sozialen und moralischen Anforderungen genügen (Bostrom & Yudkowsky, 2014, S. 317). Beispielsweise sollten Algorithmen zur Prüfung von Kreditwürdigkeit vorurteilsfrei und fair entscheiden (Mannino et al., 2015, S. 4). Noch weitaus komplexere Anforderungen ergeben sich für autonome Fahrzeuge, die in moralischen Dilemmas potenziell über Leben und Tod bestimmen müssen (Awad et al., 2018, S. 59).

Künstliche Intelligenz zu entwickeln, die moralischen und gesellschaftlichen Ansprüchen gerecht wird, ist alles andere als einfach. Wie müssen intelligente Systeme ausgerichtet werden, um menschlichen Moralansprüchen zu genügen? Wie kann sichergestellt werden, dass sie Menschen und anderen moralisch relevanten Wesen nicht (unbeabsichtigt) schaden, und nach welchen Maßstäben sollen sie entscheiden, wenn sie

[1] Im Nachfolgenden wird aufgrund einer besseren Lesbarkeit nicht gegendert. Berücksichtig werden aber alle Personenkreise.

zwischen zwei suboptimalen Ausgängen wählen müssen? Grundsätzlich stellt sich die Frage nach moralischen Werten für die künstliche Intelligenz der Zukunft.

Die Problematik steht auf der Agenda vieler Forschungsinstitutionen (Dafoe, 2018; Soares, 2015; Soares & Fallenstein, 2014) und auch außerhalb der technischen Disziplinen wird das Thema verstärkt diskutiert. Beispielsweise stellte die unabhängige Expertengruppe für künstliche Intelligenz der Europäischen Kommission Ethik-Leitlinien für vertrauenswürdige künstliche Intelligenz vor (HEG-KI, 2018). Alles in einem steht für viele Vertreter des Feldes fest, dass die künstliche Intelligenz von Beiträgen aus anderen Disziplinen profitieren kann: So forderte etwa die Forschungsinstitution OpenAI in einer Veröffentlichung aus dem Jahr 2019 explizit nach mehr Sozialwissenschaftlern in der künstlichen Intelligenz (Irving & Askell, 2019) und der offene Brief über Forschungsprioritäten für wohlwollende und sichere künstliche Intelligenz betonte die Relevanz von interdisziplinären Ansätzen (Russell et al., 2015).

Die vorliegende Arbeit kommt diesen Forderungen nach und widmet sich den ethischen Problemen in der künstlichen Intelligenz aus soziologischer Perspektive. Sie beschäftigt sich mit folgender Forschungsfrage: Gibt es universelle moralische Werte, an denen Forscher künstliche Intelligenz ausrichten können? Diese Frage wird in zwei Teilaspekte unterteilt: Erstens wird nach der Existenz universeller moralischer Werte gefragt und zweitens geht es darum, die Anwendung der gewonnenen Erkenntnisse in der künstlichen Intelligenz zu untersuchen.

Um an die Problematik im Zentrum dieser Arbeit heranzuführen, werden in Kapitel zwei zunächst einige relevante Begriffe definiert und ein Überblick über das Forschungsfeld und seine Methoden gegeben. Außerdem werden die technischen und normativen Schwierigkeiten bei der Entwicklung einer an menschlichen Moralvorstellungen ausgerichteten künstlichen Intelligenz erläutert und verschiedene Wege zur Implementierung von Moral diskutiert. Anschließend widmet sich die Arbeit dem ersten Aspekt der Forschungsfrage aus soziologischer Perspektive. Um darzustellen, wie unterschiedlich Antworten auf die Frage nach der Existenz universeller moralischer Werte ausfallen können, wird im dritten Kapitel das Gebiet der Moralsoziologie und seine verschiedenen Ansätze umrissen. Im vierten Kapitel wird ein moralsoziologischer Ansatz erläutert, der sich explizit mit der Universalisierung von Moral auseinandersetzt und der Potenzial für die Entdeckung von normativen Universalien für die künstliche Intelligenz bietet: Die Diskursethik von Jürgen Habermas. Vorgestellt wird erst die Grundkonzeption der Diskursethik und anschließend spätere Erweiterungen der Theorie. Darüber hinaus

werden Kritik an der Diskursethik und Probleme der Theorie diskutiert. Die gewonnenen Erkenntnisse werden in einem Zwischenfazit zusammengefasst und darauf basierend eine Antwort auf die erste Teilfrage formuliert. Nun wendet sich die Arbeit dem zweiten Teilaspekt der Forschungsfrage zu. Kapitel fünf untersucht die Möglichkeit, das Diskursverfahren aus der Diskursethik und die Methoden der künstlichen Intelligenz zu nutzen, um normative Universalien zu finden, die die künstliche Intelligenz in moralischen Fragen leiten könnten. Dazu wird ein System zur abstimmungsbasierten ethischen Entscheidungsfindung in der künstlichen Intelligenz vorgestellt, dass der Diskursethik in einigen wichtigen Punkten gleicht. Die Grundideen dieses Systems werden im nächsten Schritt mit Ansätzen aus einem weiteren Teilgebiet der künstlichen Intelligenz kombiniert. Es wird ein Modell vorgeschlagen, dass Diskurse mit den Methoden der künstlichen Intelligenz simuliert und dabei einige Probleme der Diskursethik von Habermas überwinden kann. Dieses Modell könnte dabei helfen, normative Universalien für die künstliche Intelligenz zu finden. Zum Ende des Kapitels wird das Modell kritisch reflektiert und eingeschränkt. Abschließend wird im sechsten Kapitel ein Fazit aus den gewonnen Erkenntnissen gezogen und eine Antwort auf die Forschungsfrage formuliert. Zudem werden Beschränkungen der vorliegenden Arbeit erläutert und ein Ausblick auf zukünftige Forschungen gegeben.

# 2. Problematisierung: Moral in der künstlichen Intelligenz

## 2.1. Definitionen

In der Forschungsfrage dieser Arbeit ist die Rede von *Werten* und *künstlicher Intelligenz.* Beide Begriffe sollen im Folgenden erläutert werden. Der Wertbegriff zählt zu den Grundbegriffen der Soziologie: Werte geben Sinn, haben eine handlungsleitende Funktion und sind für eine soziologische Erklärung menschlichen Handelns höchst relevant (Schäfers, 1992, S. 31). In der Soziologie werden sie definiert als die „allgemeinsten Grundprinzipien der Handlungsorientierung […]; sie sind Vorstellungen vom Wünschenswerten, kulturelle und religiöse, ethische und soziale Leitbilder, die über den Tag und die eigene Gesellschaft hinausweisen, die die gegebene Handlungssituation transzendieren“ (Schäfers, 1992, S. 31). Werte sind abzugrenzen von *Normen*, die Sollens-Erwartungen ausdrücken und in der Gesellschaft verbindlich durchgesetzt

werden, z. B. durch äußere Sanktionen. Normen setzten abstrakte Werte in konkrete Handlungsanweisungen um (Korb & Steinbach, 2018, S. 507). Entsprechend dieser Logik fragt diese Arbeit nach universellen Werten, die als Basis für die Ableitung von konkreten Handlungsanweisungen bzw. Normen dienen können. Während menschliche Werte üblicherweise in gültige Normen für Menschen überführt werden, interessiert in dieser Arbeit die Umsetzung von menschlichen Werten in Normen für intelligente Maschinen. Sollte es universelle moralische Werte geben, so könnten diese als Basis für die Formulierung von konkreten Normen für die künstliche Intelligenz dienen und die Lösung der weiter oben diskutierten Probleme vorantreiben.

Der Begriff künstliche Intelligenz wurde zum ersten Mal vom amerikanischen Informatiker John McCarthy verwendet. In einem Antrag für ein Forschungsprojekt stellte dieser die Vermutung auf, dass "every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it" (McCarthy et al., 1995, S.12). Inzwischen existieren viele verschiedene Definitionen von künstlicher Intelligenz (Monett et al., 2020). Hilfreich ist es, zunächst *Intelligenz* zu definieren: Intelligenz beschreibt die Fähigkeit eines Agenten, seine Ziele in verschiedenen Umgebungen zu erreichen (Legg & Hutter, 2007, S. 12). Von *künstlicher Intelligenz* wird geredet, wenn es sich nicht um einen biologischen Organismus, sondern um einen künstlichen Agenten, etwa ein computergesteuertes System, handelt (Brock, 2018, S. 5). Relevant ist zudem die doppelte Bedeutung des Begriffs: Einerseits beschreibt er die Fähigkeit eines Computers, biologische Intelligenz zu simulieren oder zumindest den Anschein von Intelligenz zu erwecken; andererseits ist er ein Oberbegriff für ein Teilgebiet der Informatik und weitere Disziplinen, die sich mit der Entwicklung und dem Einsatz von künstlicher Intelligenz auseinandersetzen (Brock, 2018, S. 5; Eitner et al., 2017, S. 6). Anders ausgedrückt beschreibt künstliche Intelligenz „a property or quality of computerized systems and […] a set of techniques used to achieve this capability" (Gabriel, 2020, S. 1). Um Verwirrung zu vermeiden, bezieht sich *künstliche Intelligenz* im Kontext dieser Arbeit auf das Forschungsfeld, während für die Fähigkeit eines computergesteuerten Systems, intelligent zu handeln, die Abkürzung *KI* genutzt wird. *KI-Systeme*, *KI-Anwendungen*, oder *KI-Programme* sind dementsprechend computergesteuerte Systeme, die in der Lage sind, intelligent zu handeln.

In der künstlichen Intelligenz gibt es mehrere Ansätze, die versuchen, zwischen verschiedenen Arten von KI zu differenzieren (Monett et al., 2020; Bringsjord & Govindarajulu, 2020; Wang, 2019; Russell & Norvig, 2020). Sehr bekannt ist die

Unterscheidung in *schwache* und *starke* bzw. *generelle* KI (Gabriel, 2020, S. 2). Schwache KI-Anwendungen zeigen eine Reihe von Fähigkeiten, die mit Intelligenz assoziiert werden, aber arbeiten nur in einem begrenzten Tätigkeitsfeld (Gabriel, 2020, S. 2). Ein KI-System ist schwach, wenn es nur ein bestimmtes Problem sehr gut bewältigt, aber darüber hinaus nicht so breit generalisieren kann wie ein Mensch (Dafoe, 2018, S.19). Ein prominentes Beispiel für eine schwache KI-Anwendung ist der Schachcomputer Deep Blue von IBM, der im Mai 1997 den Schachweltmeister Garry Kasparov besiegte (IBM, 2020).

Als stark oder generell beschrieben werden KI-Systeme, die bereichsübergreifend operieren und Aufgaben unterschiedlicher Art lösen können (Gabriel, 2020, S. 2). Da sie kognitive Qualitäten besitzen, die beim Menschen mit Intelligenz in Verbindung gebracht werden, wie etwa Problemlösungskompetenz oder die Fähigkeit zu lernen (Thorisson, 2020, S. 8), können sie in verschiedenen Domänen eingesetzt werden (Gabriel, 2020, S. 2). Die originale Definition starker KI von John Searle enthält zudem, dass starke KI-Systeme kognitive Bewusstseinszustände besitzen würden (Searle, 1980, S. 2). Allerdings nutzen viele Forscher den Begriff heute, ohne diese Komponente miteinzuschließen (z. B. Gabriel, 2020; Brock, 2018). Aus Gründen der Klarheit wird im Folgenden nur noch von *genereller KI* als „intelligence comparable to that of the human mind“ (Artificial General Intelligence Society, 2020) gesprochen. Einige Forscher, wie zum Beispiel Nick Bostrom, unterscheiden darüber hinaus *Superintelligenz* als eine KI, die selbst den klügsten Menschen in allen vorstellbaren Bereichen intellektuell überlegen ist (Bostrom, 2016). Gegenwärtig gibt es weder eine generelle KI noch eine Superintelligenz (Gabriel, 2020, S. 2). Die Entwicklung einer solchen Technologie würde jedoch weitreichende Folgen für die gesamte Gesellschaft mit sich bringen (Dafoe, 2018, S. 5).

## 2.2. Das Forschungsfeld der künstlichen Intelligenz im Überblick

Die künstliche Intelligenz ist ein sehr neues Forschungsfeld, in dem verschiedene Teildisziplinen, wie die Philosophie, die Neurowissenschaften, die Mathematik und die Statistik, arbeiten (Russel & Norvig, 2010, S. 5ff.). Genauso vielfältig wie ihre beteiligten Disziplinen sind auch ihre Methoden. Eine vollständige Darstellung des Forschungsfelds ist im Rahmen dieser Arbeit nicht möglich. Dieses Unterkapitel liefert daher lediglich einen groben Überblick über die künstliche Intelligenz.

Eine Möglichkeit zur Strukturierung des diversen Gebiets bietet das Werk „Artificial Intelligence: A modern approach“ (Russell & Norvig, 2010), indem sich die Unterscheidung von *symbolischer* und *nicht-symbolischer* KI einerseits und *logikbasierter* und *nicht-logikbasierter* KI andererseits erkennen lässt (Bringsjord & Govindarajulu, 2020).

Ansätze der symbolischen künstlichen Intelligenz modellieren Intelligenz, indem sie Wissen mit abstrakten Symbolen repräsentieren: Mit der Unterstützung von manuell eingegeben Regeln können zuvor gesammelte Daten und Zusammenhängen verarbeitet werden (Döbel et al., 2018, S. 9). Symbolische KI leitet demnach unter Verwendung von klar definierten Regeln neues Wissen aus bereits vorhandenem Wissen ab (Rimscha, 2014, S. 140). Symbolische KI kann logikbasiert oder nicht logikbasiert sein (Bringsjord & Govindarajulu, 2020). Bei der logikbasierten Programmierung werden KI-Systemen logische Klauseln an die Hand gegeben, wie etwa „X is true […] if B1, B2, and B3 are all true" (DeepAI, 2020). Anhand dieser Regeln und ihrer vorherigen Wissensbasis können KI-Systeme dann ableiten, wie sie handeln müssen. Ein Beispiel für eine symbolische, logikbasierte KI-Anwendung ist die Wissensbasis *Cyc*, die grundlegende Axiome über die Regeln der Welt sammelt und diese nutzt, um reale Daten und Informationen zu verstehen (Cycorp, 2020). Nicht logikbasierte symbolische KI nutzt probabilistische Techniken aus der statistischen Wahrscheinlichkeitstheorie und zeichnet sich durch ihren Umgang mit Unsicherheit aus: Während logikbasierte KI-Systeme immer den Wahrheitsgrad bestimmter Aussagen kennen müssen, erkennt probabilistische KI an, dass Agenten die Wahrheit oder Falschheit einer Aussage realistisch betrachtet nicht immer kennen (Bringsjord & Govindarajulu, 2020). Daher ordnet sie unterschiedlichen Szenarien bestimmte numerische Wahrscheinlichkeiten zu, die sie aus vorherigen und bedingten Wahrscheinlichkeiten berechnet (Russel & Norvig, 2010, S. 484ff.). So können probabilistische KI-Systeme unsichere Aspekte der Welt miteinbeziehen (Russel & Norvig, 2010, S. 503). Sogenannte Bayessche Netze, die den Probabilismus mit der Graphentheorie verbinden, haben in den letzten Jahren zu einem Wiederaufleben von probabilistischen Techniken geführt (Bringsjord & Govindarajulu, 2020).

In der Kognitionswissenschaft vertreten nicht-symbolische Positionen die Auffassung, dass Intelligenz nicht in symbolischer Repräsentation besteht, sondern in nicht-symbolischer Verarbeitung, die den Vorgängen im menschlichen Gehirn ähnelt (Bringsjord & Govindarajulu, 2020). Während das symbolische Paradigma einen top-

down Ansatz verwendet, um Intelligenz zu simulieren, bedient sich das nicht-symbolische einer bottom-up Methode (Bringsjord & Govindarajulu, 2020). Zur Entwicklung von nicht-symbolischer KI kommen unter anderem *künstliche neuronale Netzwerken (KNN)*, also Simulationen von Neuronennetzen, zum Einsatz ( Bringsjord & Govindarajulu, 2020). Konzepte und Informationen werden in KNN nicht logisch repräsentiert, sondern über komplexe Aktivierungsmuster innerhalb des Netzwerkes dargestellt (Buckner & Garsons, 2019). KNN bestehen aus einer variablen Zahl von identischen Einheiten bzw. künstlichen Neuronen: Eingangseinheiten, versteckten Einheiten, und Ausgangseinheiten (Buxmann & Schmidt, 2019, S. 14). Diese Einheiten stehen miteinander in Verbindungen, die mit unterschiedlich starken Gewichten behaftet sind, welche das Wissen des KNN repräsentieren. Mit Lernregeln und Trainingsdaten können die Gewichte modifiziert werden. Den Output einer jeweiligen Einheit bestimmt eine sogenannte Aktivierungsfunktion und er basiert immer auf dem Input bzw. dem Output der vorherigen Einheit und den Gewichten (Buxmann & Schmidt, 2019, S. 14f.). Vereinfacht dargestellt gelangen also Eingabedaten in das KNN, werden in den verschiedenen Einheiten mithilfe der Aktivierungsfunktion und der Gewichte verarbeitet und produzieren Ausgabedaten (Buxmann & Schmidt, 2019, S. 14f.). Nicht-symbolische KI ist üblicherweise nicht logikbasiert (Bringsjord & Govindarajulu, 2020). In Zuge des beachtlichen Wachstums der künstlichen Intelligenz werden die erläuterten logikbasierten, probabilistischen und neurocomputationalen Techniken jedoch vermehrt paradigmenübergreifend miteinander kombiniert: Das Spielprogramm *AlphaGo* von DeepMind ist ein Beispiel für ein solches Multi-Paradigmen-System (Bringsjord & Govindarajulu, 2020).

Ein wichtiger Faktor, der zu den schnellen Fortschritten in der künstlichen Intelligenz beigetragen hat, war die Entwicklung neuer Algorithmen im Teilgebiet des maschinellen Lernens. Maschinelles Lernen befähigt KI-Systeme, ihre Leistungen in Bezug auf ein bestimmtes Problem durch Vorlage von Idealbeispielen oder durch wiederholtes Üben zu verbessern (Bringsjord & Govindarajulu, 2020). Innerhalb des maschinellen Lernen werden drei Lernstile unterschieden: *überwachtes, unüberwachtes und verstärkendes* Lernen (Buxmann & Schmidt, 2019, S. 9). Beim überwachten Lernen wird ein Algorithmus mit bekannten Daten trainiert und anschließend mit Testdaten evaluiert. Zum Beispiel bekommt er eine Reihe von Bilden, die jeweils mit „Hund“ oder „Katze“ beschriftet sind, sodass er anschließend in der Lage ist, unbeschriftete Bilder in die jeweils passende Kategorie „Hund“ oder „Katze“ einzusortieren. Die Güte des

Algorithmus wird dann mithilfe des Testdatensatzes ermittelt (Buxmann & Schmidt, 2019, S. 10). Überwachtes Lernen kommt unter anderem in der Spracherkennung und der Vorhersage von Kaufverhalten zum Einsatz (Buxmann & Schmidt, 2019, S. 13). Ansätze des unüberwachten Lernens hingegen arbeiten mit unbeschrifteten Daten, in denen der Algorithmus eigenständig Muster, Zusammenhänge und Kategorien erkennen soll (Buxmann & Schmidt, 2019, S. 10). Ein gängiges Einsatzgebiet unüberwachten Lernens ist das Data Mining, bei dem große Datensätze auf interessante Informationen überprüft werden (Bringsjord & Govindarajulu, 2020). Beim verstärkenden Lernen geht es darum, ideale Strategien zur Lösung von bestimmten Problems zu erlernen (Buxmann & Schmidt, 2019, S. 10). Das funktioniert mithilfe einer Belohnungsfunktion, die dem Algorithmus zu bestimmen Zeitpunkten positives oder negatives Feedback zu seinen gewählten Aktionen gibt. Der Algorithmus probiert also nach und nach verschiedene Aktionen aus und versucht dabei, die Belohnungsfunktion zu maximieren, also möglichst viel positives Feedback zu bekommen (Buxmann & Schmidt, 2019, S. 10). Erfolgreich eingesetzt wurde verstärkendes Lernen zum Beispiel in der Entwicklung von Programmen zum Spielen von Computerspielen oder Schach (Bringsjord & Govindarajulu, 2020).

Ein neuerer Ansatz des maschinellen Lernens ist das sogenannte *Deep Learning*: Dabei werden die beschriebenen Lerntechniken auf KNN mit mehr als einer Schicht angewendet (Buxmann & Schmidt, 2019, S. 12). Ein Vorteil beim Deep Learning besteht darin, dass KNN dank ihrer vielen Schichten mehr Zusammenhänge erkennen als übliche Algorithmen des maschinellen Lernens und große Mengen von Trainingsdaten verarbeiten können (Buxmann & Schmidt, 2019, S. 12).

## 2.3. Das Alignment Problem und Herausforderungen bei der Implementierung von Moral

Nachdem relevante Grundlagen geklärt sind, kehrt dieses Unterkapitel zu den ethischen Herausforderungen der künstlichen Intelligenz zurück. Unter Experten werden diese häufig unter den Schlagwörtern *Alignment Problem* oder *Value Alignment Problem* (Gabriel, 2020; Russell, 2019) diskutiert. Beide Begriffe beziehen sich auf die schwierige Aufgabe, KI-Systeme so zu konstruieren, dass sie ihr Verhalten während ihres gesamten Betriebs an menschlichen Werten ausrichten (Gabriel, 2020, S. 2; Conn, 2017; CFI, 2020). Besonders relevant wird das Alignment Problem vor dem Hintergrund der

Möglichkeit von generellen KI-Anwendungen oder einer Superintelligenz, die autonom in vielen Bereichen dieser Welt agieren könnten und daher über große Macht verfügen würden (Gabriel, 2020, S. 2).

Das Alignment Problem setzt sich aus zwei Teilproblemen zusammen: einem technischen und einem normativen (Gabriel, 2020, S. 2). Die technische Herausforderung besteht darin, ein KI-System so eindeutig zu programmieren, dass es zuverlässig das tut, was von ihm erwartet wird und einprogrammierte Werte befolgt (Gabriel, 2020, S. 2). Es gibt viele technische Probleme, die zu ethischen Schwierigkeiten führen könnten: Zum Beispiel könnten KI-Systeme uneindeutig programmiert werden und so Dinge tun, die ihre Entwickler nicht beabsichtigt haben, oder sie könnten beim Verfolgen ihrer programmierten Ziele negative Nebeneffekte produzieren (Amodei et al., 2016, S. 3; Yudkowsky, 2011, S. 2). Darüber hinaus könnte ein KI-Programm, das ein bestimmtes Ziel erreichen soll, dafür zerstörerische Methoden entwickeln (Bostrom, 2003, S. 5). Weitere technische Aspekte des Alignment Problems werden in der künstlichen Intelligenz umfassend diskutiert, unter anderem die Sicherung von Kontrolle über KI-Programme, der Umgang mit menschlichen Fehlern und die Garantie von stabilem Verhalten in unterschiedlichen Umgebungen (Amodei et al., 2016, S. 2 f.; Russell et al., 2015, S. 106ff.; Soares & Fallenstein, 2017, S. 2). Eine komplette Übersicht über alle technischen Probleme ist an dieser Stelle weder möglich noch nötig, stattdessen wird übergegangen zur Darstellung der normativen Aspekte.

Auf der normativen Seite des Alignment Problems gilt es zu beantworten, was die KI tun sollte, also "what values or principles, if any, we ought to encode in artificial agents" (Gabriel, 2020, S. 2). Diskutiert werden hier beispielsweise Prinzipien zur Vermeidung von Diskriminierung durch KI-Systeme, zur Regulierung des Verhaltens von intelligenten Maschinen in Mensch-Maschine Interaktionen und zur Kontrolle autonomer KI-Systemen mit hoher Entscheidungsmacht, wie autonome Fahrzeuge oder autonome Waffensysteme (Müller, 2020). Gerade der Einsatz von autonomen Fahrzeugen wird medial (Simanowski, 2017) und in der Wissenschaft häufig thematisiert, vor allem in Bezug auf das bekannte Trolley-Problem[2] (Awad et al., 2018). Probleme,

[2] Das Trolley-Problem ist ein moralischer Gedankenexperiment, dass erstmals 1930 von Karl Engisch diskutiert wurde. Es entwirft eine Situation, in dem ein Zug droht Menschen zu überfahren, dieser jedoch mit Hilfe eines Weichenstellers umgeleitet werden kann. Auch wenn der Zug umgeleitet wird, werden Menschenleben gefährdet; jedoch weniger, als wenn dieser seinen eigentlichen Weg fortsetzt (Engisch, 1930). Solche Entscheidungen, in denen Menschenleben auf dem Spiel stehen und es nur schlechte Alternativen gibt, müssen möglicherweise auch die autonomen Fahrzeuge der Zukunft treffen (Bonnefon et al., 2016, S. 1573).

vor denen selbstfahrende Autos jedoch wahrscheinlich öfter stehen werden, beziehen sich auf die Abwägung zwischen persönlichen Interessen des Fahrers und dem Gemeinwohl, etwa beim Überschreiten der Höchstgeschwindigkeit oder dem Nichteinhalten des Sicherheitsabstandes (Müller, 2020). Relevant ist gerade beim autonomen Fahren auch die Frage, wer die autonomen Fahrzeuge kontrolliert und wer im Falle ihres Versagens die Verantwortung trägt (Müller, 2020). Insgesamt ergeben sich in den zahlreichen Einsatzgebieten von KI-Systemen viele normative Fragen, die der Klärung bedürfen (Russel et al., 2015, S. 106ff.).

Die künstliche Intelligenz braucht moralische Werte und Grundsätze, an denen sie das Handeln und die Entscheidungen ihrer KI-Systeme ausrichten kann (Gabriel, 2020, S. 10ff.; Burton et al., 2017, S. 5ff.; Russel et al., 2015, S. 107). In der Philosophie existieren viele Theorien, die sich für die Reflexion über Ethik in der künstlichen Intelligenz eignen, aber jeweils unterschiedliche Antworten auf relevante Fragen geben. Auch in der Gesellschaft finden sich diesbezüglich sehr unterschiedliche Meinungen, Werte und Präferenzen (Burton et al., 2017, S. 5ff.). Einheitliche Richtlinien für die Konstruktion von den Menschen wohlgesonnenen KI-Systemen bzw. „friendly AI" (Yudkowsky, 2001, S. 2) zu finden, gestaltet sich in einer globalisierten Welt mit unterschiedlichen Menschen und diversen Ansichten folglich sehr schwierig (Gabriel, 2020, S. 10).

An zusätzlicher Komplexität gewinnt das Alignment Problem, wenn der enge Zusammenhang zwischen seinen technischen und normativen Aspekten berücksichtigt wird (Gabriel, 2020, S. 3). Schließlich haben die zur Entwicklung eines KI-Systems genutzten Methoden erheblichen Einfluss auf die Art von Werten und Prinzipien, die einprogrammiert werden können (Gabriel, 2020, S. 3). Das kann gut am Beispiel des verstärkenden Lernens verdeutlicht werden: Beim verstärkenden Lernen ist das einzige Ziel der KI-Anwendung, eine vorgegebene Belohnungsfunktion langfristig zu maximieren, also möglichst viele positive Feedback-Signale zu erhalten (Gabriel, 2020, S. 3). Moral kann in einem solchen Design nur aufgenommen werden, wenn sie sich in Form von einem zu maximierenden Guten ausdrücken lässt. Das funktioniert gut mit konsequentialistischen[3] Moraltheorien, insbesondere mit dem Akt-Utilitarianismus (Gabriel, 2020, S. 3). Dieser besagt, dass die moralisch richtige Handlung immer

[3] Konsequentialismus ist die Ansicht, dass normative Eigenschaften nur von Konsequenzen abhängen. Am häufigsten wird der Konsequentialismus auf die moralische Richtigkeit von Handlungen angewendet: Er besagt dann, dass nur die Konsequenzen einer Handlung entscheiden, ob diese moralisch richtig ist (Sinnott-Armstrong, 2019).

diejenige ist, die zur „greatest happiness for the greatest number of sentient creatures in the future“ (Gabriel, 2020, S. 3) führt. Eine solche Maximierung des größten Glücks ließe sich in einem Modell des verstärkenden Lernens mit einer Belohnungsfunktion beschreiben (Gabriel, 2020, S. 3). Anders verhält es sich mit Moraltheorien, die von individuellen Rechten der Menschen ausgehen, welche nicht verletzt werden dürfen: Einer KI-Anwendung, deren Lernprozess von der Maximierung positiver Feedback Signale geleitet wird, Respekt für individuelle Menschenrechte beizubringen, ist schwierig (Gabriel, 2020, S. 4).

Prinzipiell gibt es laut Wallach und Allen drei Möglichkeiten, KI-Systeme mit Moralität auszustatten: *top-down Ansätze*, *bottom-up Ansätze* und *hybride Ansätze* (Wallach & Allen, 2009, S. 80ff.). Bei *top-down* Ansätzen werden KI-Anwendungen eine Reihe von moralischen Regeln vorgegeben, nach denen sie sich in ihrem Verhalten richten sollen (Loh, 2018, S. 9). Problematisch ist dabei einerseits, dass moralische Regeln kontextbezogen interpretiert werden müssen, aber nur eindeutige Vorgaben mit eindeutigen Interpretationen programmierbar sind. Andererseits besteht die Gefahr, dass in der Praxis zwei oder mehr einprogrammierte Regeln miteinander in Konflikt geraten (Loh, 2018, S. 9).

In *bottom-up* Ansätzen werden KI-Systemen keine moralischen Regeln vorgegeben, sie erhalten lediglich grundlegende Kompetenzen (Loh, 2018, S. 9). Mithilfe von verschiedenen Lernstrategien sollen sie im Laufe der Zeit eigenständig moralisches Verhalten entwickeln. Hier können *Evolutionsmodelle* und *Modelle menschlicher Sozialisation* unterschieden werden (Loh, 2018, S. 9). In Evolutionsmodellen wird ein künstliches System mit mehreren unterschiedlichen Programmen erschaffen, die ein ethisches Problem lösen müssen. Selektiert werden anschließend diejenigen Programme, die die Aufgabe am besten bewältigt haben. Miteinander rekombiniert werden ihnen weitere Probleme gestellt. Dieser Prozess wird so lange wiederholt, bis das System moralisches Verhalten perfektioniert hat (Loh, 2018, S. 9). Modelle menschlicher Sozialisation orientieren sich an der Art und Weise, wie Kinder moralisches Verhalten erlernen (Loh, 2018, S. 10). Sie beziehen Mitgefühl, Emotionen, Belohnungen, Strafen, Zustimmung und Ablehnung mit in ihre Handlungsentscheidungen ein (Wallach & Allen, 2009, S. 107). Darüber, wie genau Menschen Moral erlernen, besteht jedoch kein Konsens und die basalen Lernfähigkeiten existierender KI-Systeme sind mit den umfassenden Lernkompetenzen kleiner Kinder nicht zu vergleichen (Wallach & Allen, 2009, S. 108 ff.). Bisher wurden Lernstrategien zudem nicht auf die Moralentwicklung

angewendet (Wallach & Allen, 2009, S. 112). Bottom-up Ansätze berücksichtigen die Kontextsensitivität von moralischem Verhalten und ermöglichen KI-Systemen zu lernen, wie Menschen situativ handeln und entscheiden (Loh, 2018, S. 10). Nachteile solcher Ansätze sind, dass Evolutions- und Lernprozesse lange dauern und dass sie überhaupt keine Regeln vorgeben, KI-Systemen also in der Entwicklungsphase Fehler und potenzielle Schadensverursachung erlauben (Wallach & Allen, 2009, S. 115f.)

*Hybride* Ansätze bewegen sich zwischen den erläuterten Paradigmen: Sie geben grundlegende ethische Prinzipien vor, die dann in Lernprozessen an unterschiedliche Umgebungen und Situationen angepasst werden (Loh, 2018, S. 10). Die programmierten Regeln hängen dabei von den Aufgaben und der Umgebung des KI-Systems ab (Loh, 2018, S. 10). Nach Loh kann von einem hybriden System gesprochen werden, wenn es „in einem anpassungsfähigen Spielraum agieren [kann], innerhalb dessen es auf die Wertvorstellungen seiner Nutzer*innen kontextsensitiv reagiert“ (Loh, 2018, S. 10f.).

# 3. Moralsoziologie: Ein Überblick

Nachdem die Paradigmen und Methoden der künstlichen Intelligenz sowie die Herausforderungen bei der Implementierung von Moral deutlich geworden sind, widmet sich die Arbeit in den folgenden zwei Kapiteln dem ersten Teilaspekt der Forschungsfrage: der Frage nach der Existenz universeller moralischer Werte. Da sich diese Arbeit der Thematik aus soziologischer Perspektive widmet, ist es zunächst wichtig, das Gebiet der Moralsoziologie und seine Paradigmen zu kennen. Dazu wird im Folgenden ein Überblick gegeben, der die Diversität moralsoziologischer Ansätze spiegelt und erkennen lässt, wie unterschiedlich Antworten auf die Frage nach universellen moralischen Werten ausfallen können.

Die Soziologie ist seit ihrer Entwicklung zu einer eigenständigen wissenschaftlichen Disziplin im 19. Jahrhundert mit der Moral verstrickt: Eine Beziehung, die nicht unproblematisch ist (Liebig, 2007, S. 1). Ein Grund für das schwierige Verhältnis der Soziologie zur Moral besteht darin, dass die Soziologie unter anderem aus der Moralphilosophie, den politischen Tugendlehren und der Moralstatistik hervorgegangen ist und sich daher bewusst von diesen Disziplinen abgrenzen möchte (Bergmann, 1998, S. 70). Schließlich ist es Anspruch der Soziologie, eine objektive

Wissenschaft zu sein, die sich auf empirische Daten stützt (Herbermann, 2002, S. 281f.). Folglich ist die Soziologie stets bemüht, Moral rein deskriptiv oder explanativ, nicht normativ, zu untersuchen.

Auf der anderen Seite ist die Moral schon immer ein Thema soziologischer Analysen (Bergmann, 1999, S. 70). So bezeichnete Durkheim, einer ihrer Gründerväter, die Soziologie sogar als Moralwissenschaft und definierte Moral als Voraussetzung für jede Form des gesellschaftlichen Zusammenlebens (Liebig, 2007, S. 1). Viele weitere bekannte Soziologen beschäftigten sich in ihren Arbeiten intensiv mit den Funktionen und Arten von Moral in verschiedenen Lebensbereichen (Reinhold, 2000, S. 445ff.). Grundsätzlich können in der Moralsoziologie drei Paradigmen unterschieden werden: Einmal kommt der Moral eine normativ-integrierende Funktion zu, einmal wird ihre Sinnkonstruktionsleistung im Alltag, etwa im Rahmen von moralischer Kommunikation, untersucht und einmal wird sie rational aus der Sicht von egoistischen Individuen erklärt (Beetz, 2009, S. 249ff.).

Das Paradigma, das von einer zentralen Integrationsfunktion der Moral ausgeht, basiert zu großen Teilen auf den Werken von Émile Durkheim (Liebig, 2007, S. 6). Durkheim definiert Moral als ein „System von Geboten und Verhaltensregeln“ (Durkheim & Adorno, 1996, S. 92), das sich aus den strukturellen Bedingungen einer Gesellschaft entwickelt (Liebig, 2007, S. 13). Er geht davon aus, dass in einer Gesellschaft Konsens über die moralischen Regeln herrscht und deren Befolgung über Sanktionen eingefordert wird. Die Moral sei essenziell für die Gesellschaft, da erst sie Integration schaffe und Solidarität zwischen den Individuen entstehen lasse (Liebig, 2007, S. 13). Eine ähnliche Funktionalität der Moral für die Herstellung von sozialer Ordnung findet sich in den Werken von Talcott Parsons (Dallinger, 2009, S. 75). Für Parsons ist Moral ein Teil von Kultur. Kultur ist bei ihm gleichzeitig kollektives Wissensmuster und Vorrat an kollektiv geteilten und handlungsleitenden Werten (Dallinger, 2009, S. 95). Moralische Wertstandards liegen seiner Ansicht nach in den „prinzipiellen Orientierungsmöglichkeiten menschlichen Handelns“ (Liebig, 2007, S. 15), wo sie die entscheidende Instanz zur Bewertung von Handlungsoptionen und das wichtigste Mittel zur Handlungsabstimmung sind (Dallinger, 2009, S. 97). Auch im Rahmen seines bekannten Vier-Funktionen-Schemas gesellschaftlicher Subsysteme kommt dem kulturellen System, zu dem auch die Moral zählt, die höchste Position und somit eine steuernde und integrierende Rolle zu (Dallinger, 2009, S. 113).

Neuere Positionen der Soziologie kritisieren solche Konzeptionen und zweifeln an der Integrationsfunktion der Moral (Luhmann & Horster, 2008; Bergmann & Luckmann, 1999). Eine integrierende Funktion könne Moral schließlich nur haben, wenn in der Gesellschaft moralischer Konsens herrsche (Liebig, 2007, S. 26). Gerade ein solcher Konsens sei in modernen Gesellschaften jedoch nicht mehr möglich, stattdessen würden moralische Perspektiven immer diverser und ein großer Teil der Moral und ihrer Funktionen von institutionalisiertem Recht ersetzt (Liebig, 2007, S. 3, S. 26). Das Paradigma der Sinnkonstruktion beschäftigt sich daher mit der Rolle, die Moral im lebensweltlichen Alltagshandeln der Individuen spielt (Beetz, 2009, S. 250f.). So untersuchten Bergmann und Luckmann die kommunikative Konstruktion von Moral in alltäglichen Situationen und beschrieben verschiedene Erscheinungsformen moralischer Kommunikation (Bergmann & Luckmann, 1999). Moralische Kommunikation definierten sie dabei wie folgt: Es handelt sich um moralische Kommunikation, „wenn in der Kommunikation einzelne Momente der Achtung oder Missachtung, also der sozialen Wertschätzung einer Person, mittransportiert werden und dazu ein situativer Bezug auf übersituative Vorstellungen von ‚gut' und ‚böse' beziehungsweise vom ‚guten Leben' stattfindet" (Bergmann & Luckmann, 1999, S. 22). Laut Bergmann und Luckmann liegt Moral in der lebensweltlichen Alltagskommunikation, wird sozial konstruiert und in verschiedenen Kontexten immer wieder neu erschaffen und besitzt einen lokalen, flüchtigen Charakter (Liebig, 2007, S. 34ff.). Niklas Luhmann definiert Moral, ähnlich wie Bergmann und Luckmann, als spezifische Art der Kommunikation, in der personale Achtung oder Missachtung Ausdruck findet (Luhmann & Spaemann, 1990, S. 18). Zunächst beschränkt Luhmann die Moral auf direkte Interaktionen zwischen Personen und beschreibt sie als Mechanismus zur sozialen Koordination, der Interaktionen auf das Schema Achtung/Missachtung reduziert (Liebig, 2007, S. 29). Der Ausdruck von Achtung signalisiere den Einbau der Ziele und Wünsche des Gegenübers in eigene Handlungen und Ziele und löse so Probleme der doppelten Kontingenz[4] (Großmaß & Anhorn, 2013, S. 70f.). Außerdem sichere Moral die Fortsetzung von Kommunikation, indem sie eine Meta-Diskussion über die Bedingungen von Achtung bzw. Missachtung ermögliche (Liebig, 2007, S. 30). In Luhmanns Systemtheorie kommt der Moral eine ambivalente Rolle zu: Einerseits würden sich gesellschaftliche Teilsysteme um Abstand von der Moral bemühen, da diese ihre von moralischer Bewertung unabhängigen

[4] Probleme der doppelten Kontingenz sind Situationen, in denen der Erfolg der Handlung von Ego vom Handeln des Alter abhängt, aber wechselseitig Unsicherheit über das Verhalten des jeweils anderen besteht. Weder Alter noch Ego weiß mit Sicherheit, wie sich das Gegenüber verhalten wird (Luhmann & Pfürtner, 1978, S. 44)

Eigenlogiken störe. Gleichzeitig diene Moral aber als Hinweis auf Probleme einzelner Funktionssysteme: Moralische Kommunikation deute darauf hin, dass ein Teilsystem ein Problem nicht mit systeminternen Methoden bewältigen könne (Liebig, 2007, S. 32). Manche Teilsysteme bräuchten die Moral zudem als Absicherung für ihre systemeigenen Codes (Großmaß & Anhorn, 2013, S. 75f.). Wenn diese auf Vertrauen beruhen, also z. B. durch Bestechung oder Fälschung hintergangen werden könnten, benötigten sie Moralisierung, um zu funktionieren (Großmaß & Anhorn, 2013, S. 76).

Das letzte Paradigma der Moralsoziologie beschäftigt sich mit der Rationalität von Moral und fragt unter anderem, warum eigeninteressierte Individuen überhaupt moralisch handeln (Beetz, 2009, S. 250). Ansätze in dieser Tradition versuchen zu zeigen, warum moralisches Verhalten, basierend auf individuellen Nützlichkeitserwägungen, vorteilhaft sein kann (Liebig, 2007, S. 7). So diskutiert Rainer Hegselmann zum Beispiel das klassische Gefangenendilemma[5] und geht davon aus, dass es für beide Gefangenen neben der egoistischen Nutzenfunktion auch eine moralische Nutzenfunktion gibt. Dieser moralischen Nutzenfunktion schreiben die Gefangenen jeweils unterschiedliche Bedeutung zu (Liebig, 2007, S. 55). Ist ihnen der moralische Standpunkt wichtig, so kann für beide Schweigen die bessere Option sein. Das würde zu einer Kooperation mit Vorteilen für beide Gefangenen führen (Liebig, 2007, S. 54). Viele Ansätze dieser Tradition orientieren sich an der Rational-Choice-Theory von George C. Homans (Beetz, 2009, S. 250).

# 4. Die Diskursethik nach Jürgen Habermas

Wie im vorherigen Kapitel deutlich geworden ist, gibt es in der Soziologie viele verschiedene Perspektiven, die für die Beantwortung der ersten Teilfrage dieser Arbeit, der Frage nach universellen Werten, herangezogen werden könnten. Das ist im Rahmen

[5] Das Gefangenendilemma ist ein Spiel aus der Spieltheorie welches die Situation zweier Gefangener modelliert, die beschuldigt werden, gemeinsam eine Bank ausgeraubt zu haben. Beide sitzen in Isolationshaft und können entscheiden, ob sie das Verbrechen gestehen oder schweigen möchten. Wenn einer gesteht und der andere schweigt wird der Geständige freigesprochen und sein Komplize erhält die Höchststrafe. Wenn beide gestehen, bekommen sie auf Grund des Geständnisses eine Strafe, aber nicht die Höchststrafe. Wenn hingegen beide schweigen bekommen sie eine niedrige Strafe, da ihnen nur eine geringer zu bestrafende Tat nachgewiesen werden kann. Das Dilemma entsteht dadurch, dass jeder, egal was der andere tut, besser gestehen als schweigen sollte. Jedoch ist das Ergebnis, das sie erzielen würden, wenn beide geständen, für jeden schlechter als das Ergebnis, das sie erzielen würden, wenn beide schwiegen (Kuhn, 2019).

dieser Arbeit jedoch nicht möglich. Sie zieht daher einen moralsoziologischen Ansatz heran, der sich für die Beantwortung der Forschungsfrage besonders eignet: Die Diskursethik von Jürgen Habermas. Die Diskursethik setzt sich explizit mit der Universalisierung von Moral auseinander und kann potenziell bei der Entdeckung von normativen Universalien für die künstliche Intelligenz unterstützen.

Als Soziologe und Philosoph gleichermaßen ist Habermas einer der einflussreichsten Intellektuellen der Welt, dessen Theorien bis heute viel diskutiert und weiterentwickelt werden (Bohman & Rehg, 2017). Er steht in der Tradition der kritischen Theorie nach Horkheimer und Adorno. Der normative Ausgangspunkt für all seine Theorien ist das Ideal einer herrschaftsfreien Gesellschaft (Horster, 2006, S. 11ff.). Habermas moraltheoretischen Überlegungen, die in der Diskursethik mündeten, lassen sich nur schwer einem der drei vorgestellten Paradigmen der Moralsoziologie zuordnen. Einerseits beschäftigt er sich, wie die Vertreter des symbolischen Paradigmas, intensiv mit Kommunikation (Horster, 2006, S. 45ff.) und untersucht die für seine Konzeption zentralen Diskurse in ihrer Anbindung an die Lebenswelt (z. B. Habermas, 1992, S. 30ff.). Andererseits sind seine soziologischen Analysen als Vertreter der kritischen Theorie mit dem Ideal einer herrschaftsfreien Gesellschaft grundsätzlich normativ orientiert (Horster, 2006, S. 11). Darüber hinaus besitzt die Diskursethik auch rationale Elemente: Sie geht davon aus, dass Moral durch die Angabe von rationalen Gründen begründet werden kann (Seiler, 2014, S. 38).

## 4.1. Grundlagen der Diskursethik

### 4.1.1. Herleitung der Kernprinzipien

Habermas gesamtes Werk ist gekennzeichnet von einer intensiven Beschäftigung mit Sprache (Horster, 2006, S. 45ff.). Er geht davon aus, dass Sprache kulturelles und lebensweltliches Wissen bewahrt und immer schon normative Voraussetzungen enthält (Horster, 2006, S. 45). In seiner Theorie des kommunikativen Handelns bezeichnet er Sprache als „ein Medium unverkürzter Verständigung (…), wobei sich Sprecher und Hörer aus dem Horizont ihrer vorinterpretierten Lebenswelt gleichzeitig auf etwas in der objektiven, sozialen, und subjektiven Welt beziehen, um gemeinsame Situationsdefinitionen auszuhandeln“ (Habermas, 1981, S. 142). In Orientierung am Werk des deutschen Philosophen Karl-Otto Apel begründet auch Habermas seine

Diskursethik aus den Voraussetzungen der sprachlichen Kommunikation (Gottschalk-Mazouz, 2000, S. 17). Aus diesen Voraussetzungen leitet er eine Konsenstheorie der normativen Richtigkeit ab und führt den *Diskurs* als Verfahren ein, mit dem Normen auf ihre universelle Geltung geprüft werden können (Seiler, 2014, S. 32ff.). Insgesamt lässt sich mit Habermas Diskursethik für die Existenz universeller Normen argumentieren. Bevor das zentrale Element der Theorie, der Diskurs, näher untersucht wird, muss geklärt werden, wann diese Form der Kommunikation zum Einsatz kommt.

Eine dafür relevante Grundlage ist die Differenzierung von Handlungstypen: Habermas unterscheidet zunächst zwischen instrumentellem und sozialem Handeln und trennt innerhalb des sozialen Handelns strategisches und kommunikatives Handeln voneinander ab (Habermas, 1984, S. 460). Instrumentelles Handeln sei erfolgsorientiert und fokussiere die Erreichung gewisser Ziele in der physikalischen Welt. Soziales Handeln hingegen sei immer mit zwischenmenschlichen Interaktionen verknüpft. Hier würden im Falle des strategischen Handelns andere Menschen jedoch nur im Rahmen eigennütziger Erfolgskalküle mit einbezogen, auch strategisches Handeln sei erfolgsorientiert (Habermas, 1984, S. 460). Demgegenüber ziele kommunikatives Handeln nicht primär auf den eigenen Erfolg, sondern auf intersubjektive Verständigung: „Kommunikativ nenne ich die Interaktionen, in denen die Beteiligten ihre Handlungspläne einvernehmlich koordinieren; dabei bemißt sich das jeweils erzielte Einverständnis an der intersubjektiven Anerkennung von Geltungsansprüchen“ (Habermas, 1983, S. 68). Bei Habermas tritt im kommunikativen Handeln an die Stelle der praktischen Vernunft die *kommunikative Vernunft* (Habermas, 1992, S. 17f.). Die praktische Vernunft beschreibe ein subjektives Vermögen, das sich auf das individuelle Glück und die moralische Autonomie des einzelnen Subjekts als Staatsbürger und Gesellschaftsmitglied richte und es in seinem Handeln leite (Habermas, 1992, S. 15ff.). Hingegen wird die kommunikative Vernunft keinem einzelnen Akteur zugeschrieben, sondern in die sprachliche Interaktion zwischen Individuen verlegt, wo sie auf Einverständnis ziele. Sie gebe den Individuen nicht vor, was diese tun sollen, sondern habe nur insofern normativen Gehalt, als sich „kommunikativ Handelnde auf pragmatische Voraussetzungen kontrafaktischer Art einlassen müssen“ (Habermas, 1992, S. 18). Diese Voraussetzungen bestehen nach Habermas darin, dass die Individuen sogenannte *Geltungsansprüche* erheben und einander die Erfüllung dieser Geltungsansprüche unterstellen (Habermas, 1992, S. 18f.). Diese würden in jeder Art der Kommunikation impliziert. Niemand könne sprechen, ohne diese Geltungsansprüche

einzufordern und zu unterstellen, dass das Gegenüber dasselbe tue (Gottschalk-Mazouz, 2000, S. 19).

Habermas unterscheidet vier Geltungsansprüche: Ansprüche auf Wahrheit, Richtigkeit, Wahrhaftigkeit und Verständlichkeit (Horster, 2006, S. 50). In Bezug auf die objektive Welt erhöben Sprecher den Anspruch der Wahrheit, d. h. sie unterstellten, dass das, was sie sagen, wahr sei und die real existierenden Sachverhalte korrekt beschreibe (Habermas, 1983, S. 68). Der Geltungsanspruch der Richtigkeit beziehe sich auf die soziale Welt. Hier werde behauptet, dass die Sprechhandlung den festgelegten Regeln für intersubjektive Beziehungen entspreche (Habermas, 1983, S. 68). Wahrhaftigkeit stehe in Verbindung zur subjektiven Welt. Der Sprecher unterstelle, dass er seine inneren Erlebnisse und Gedanken in seiner Aussage aufrichtig wiedergebe (Habermas, 1983, S. 68). Mit dem Geltungsanspruch der Verständlichkeit würden Sprecher voraussetzen, dass ihre Gesprächspartner die gebrauchten Worte und Sätze auch verstünden. Verständlichkeit bezeichnet Habermas später nicht mehr als Geltungsanspruch, sondern als Bedingung für das Gelingen von Kommunikation (Horster, 2006, S. 53). Während die Geltungsansprüche im kommunikativen Handeln immer mitschwängen, böten Diskurse die Möglichkeit, problematisch gewordene Geltungsansprüche zu hinterfragen und auf ihre Berechtigung hin zu untersuchen. Mithilfe des Diskurses würden Geltungsansprüche folglich reflexiv thematisiert, also Zweifel an ihrer Einlösung geklärt (Gottschalk-Mazouz, 2000, S. 20).

Laut Habermas beruht jede Art von Diskurs auf Argumentation und behandelt Probleme, die grundsätzlich durch die Angabe von rationalen Gründen gelöst werden können (Gottschalk-Mazouz, 2000, S. 20). In Diskursen gälten bestimmte Argumentationsvoraussetzung bzw. Diskursregeln, die von den Teilnehmern eines Diskurses intuitiv angenommen würden (Habermas, 1983, S. 101). Habermas unterscheidet hier zwischen Voraussetzungen auf der logischen, der dialektischen und der rhetorischen Ebene (Habermas, 1983, S. 97). Auf der logischen Ebene würden logische und semantische Regeln impliziert, die widerspruchsfreie, konstante, und eindeutige Argumentationen sicherten und frei von ethischen Gehalten seien (Habermas, 1983, S. 97). Voraussetzungen auf der dialektischen Ebene beträfen die Wahrhaftigkeit aller Teilnehmer. So dürfe beispielsweise jeder Teilnehmer nur Behauptungen aufstellen, an die er selber glaube und nur an Aussagen von anderen Teilnehmern zweifeln, wenn er dafür gute Gründe nennen könne (Habermas, 1983, S. 98). Diese Voraussetzungen hätten eine ethische Komponente, da sie Bedingungen betreffen würden, „die der Diskurs mit

dem verständigungsorientierten Handeln überhaupt teilt, z. B. Verhältnisse reziproker Anerkennung“ (Habermas, 1983, S. 98). Den höchsten ethischen Gehalt besäßen die Voraussetzungen auf der rhetorischen Ebene, die Habermas über die Beschreibung einer „idealen Sprechsituation“ (Habermas, 1983, S. 98) darstellt. Die ideale Sprechsituation kennzeichnen laut Habermas folgende Merkmale: 1. Jeder muss die gleiche Chance haben, Diskurse zu eröffnen und an ihnen teilzunehmen; 2. Jeder Diskursteilnehmer darf Argumente beisteuern und die Argumente anderer hinterfragen; 3. Jeder Diskursteilnehmer hat das gleiche Recht, Emotionen und Wünsche wahrhaftig anzubringen; 4. Jeder Diskurskursteilnehmer hat die gleiche Chance, zu befehlen, sich zu widersetzen oder anderweitig regulativ zu sprechen, der Diskurs muss also frei von Machtgefällen aller Art sein (Horster, 2006, S. 55).

Als *diskursiven Konsens* bezeichnet Habermas die in einer idealen Sprechsituation über rationale Argumentation erzielte Übereinstimmung aller Diskursteilnehmer (Seiler, 2014, S. 32; Horster, 2006, S. 57). Er unterscheidet zwischen theoretischen und praktischen Diskursen. In theoretischen Diskursen werde der Geltungsanspruch der Wahrheit geprüft und ein in ihnen erzielter Konsens habe „wahrheitsverbürgende Kraft“ (Seiler, 2014, S. 32). Das wird in Habermas Konsenstheorie der Wahrheit wie folgt begründet: Auch wenn reale Diskurse die Bedingungen der idealen Argumentation nicht erfüllen und räumlichen wie zeitlichen Beschränkungen unterliegen, müsse zunächst jeder, der an einem Diskurs teilnimmt, eine ideale Diskurssituation voraussetzen. Für Argumentationszwecke sei es ausreichend, wenn alle Diskursteilnehmer eine annähernde Erfüllung der genannten Bedingungen unterstellen würden, auch wenn diese Unterstellung „kontrafaktischen Charakter“ (Habermas, 1983, S. 102) habe. Wie in jeder Art der Kommunikation würden die Diskursteilnehmer auch in ihren Argumentationen die drei Geltungsansprüche einlösen. Im Rahmen des Geltungsanspruchs der Wahrheit gingen sie also davon aus, dass ihre vorgebrachten Argumente wahr seien (Seiler, 2014, S. 32). Da in einer idealen Sprechsituation alle Diskursteilnehmer die gleichen Chancen hätten, Argumente zu bringen und zu hinterfragen, und es keine Machtgefälle gäbe, zähle allein die Güte der Argumente (Seiler, 2014, S. 33). Das beste Argument habe eine „konsenserzielende Kraft“ (Habermas, 1984, S. 176), die die Zustimmung aller Diskursteilnehmer anrege (Seiler, 2014, S. 33). Wahr ist eine Aussage nach Habermas, wenn ihr Geltungsanspruch der Wahrheit in einem Diskurs über einen diskursiven, begründeten Konsens eingelöst werden kann. Ein solcher Diskurs basiere auf Voraussetzungen mit kontrafaktischem Gehalt: Einerseits müssten die Diskursteilnehmer

die Geltungsansprüche der Kommunikation erheben und einander deren Erfüllung unterstellen. Andererseits müssten sie annehmen, dass die Argumentationsvoraussetzungen eingehalten werden und wirklich nur die Güte der Argumente zählt (Seiler, 2014, S. 32f.).

Diese Konzeption der Wahrheit überträgt Habermas in seiner Diskursethik auf die moralische Richtigkeit: Während in den erläuterten theoretischen Diskursen der Geltungsanspruch der Wahrheit einer Aussage geprüft werden könne, würden sich praktische Diskurse für die Prüfung des Geltungsanspruchs der Richtigkeit[6] eignen (Habermas, 1983, S. 68ff.). Mit dieser These nimmt Habermas eine kognitivistische Position ein: Er geht davon aus, dass normative Aussagen einen wahrheitsanalogen Geltungsanspruch besitzen, moralische Normen also rational und eindeutig begründet werden können (Habermas, 1983, S. 66f.). Das drückte er schon 1973 aus: „…der normative Geltungsanspruch selber ist kognitiv im Sinne der (wie immer kontrafaktischen) Unterstellung, dass er diskursiv eingelöst, also in einem argumentativ erzielten Konsensus der Beteiligten begründet werden könnte." (Habermas, 1973, S. 144). Die diskursive Einlösung von moralischen Normen ähnelt der bereits beschriebenen diskursiven Einlösung von Wahrheitsansprüchen. Genau wie jemand eine Tatsachenbehauptung mit Argumenten belegen könne, könne dieser auch Gründe für die Rechtfertigung einer Norm angeben (Seiler, 2014, S. 33f.). Wenn über eine Norm in einem praktischen Diskurs ein Konsens erzielt wird, so besitzt sie nach Habermas wahrheitsanaloge Richtigkeit. Auch in praktischen Diskursen müssten die Diskursteilnehmer die weiter oben erläuterten Voraussetzungen kontrafaktisch unterstellen (Seiler, 2014, S. 33f.). Wie diskursiv eingelöste Wahrheitsansprüche seien auch diskursiv eingelöste Ansprüche der moralischen Richtigkeit bzw. Normen kontextübergreifend gültig (Gottschalk-Mazouz, 2000, S. 25).

Aus diesen Grundlagen leitet Habermas die Kernprinzipien seiner Diskursethik ab: den diskursethischen Grundsatz D und den Universalisierungsgrundsatz U (Habermas, 1983, S. 73ff.). Nach D darf „eine Norm nur dann Geltung beanspruchen, wenn alle von ihr möglicherweise Betroffenen als *Teilnehmer eines praktischen Diskurses* Einverständnis darüber erzielen (bzw. erzielen würden), daß diese Norm gilt"

[6] Der Geltungsanspruch der Wahrhaftigkeit bzw. ob sich jemand so verhält, wie er behauptet, kann nur über die Beobachtung seines Handelns geprüft werden (Habermas, 1983, S. 69f.).

(Habermas, 1983, S. 76, Hervorhebung i.O.). U hingegen ist nach Habermas eine Argumentationsregel, die er ursprünglich wie folgt formulierte:

> „So muß jede gültige Norm der Bedingung genügen, daß die Folgen und Nebenwirkungen, die sich jeweils aus ihrer *allgemeinen* Befolgung für die Befriedigung der Interessen eines *jeden* Einzelnen (voraussichtlich) ergeben, von *allen* Betroffenen akzeptiert (und den Auswirkungen der bekannten alternativen Regelungen vorgezogen) werden können“ (Habermas, 1983, S. 75f., Hervorhebung i. O.)

Laut Habermas kann U aus den erwähnten Voraussetzungen der Argumentation abgeleitet werden: „Aus den genannten Diskursregeln ergibt sich nämlich, daß eine strittige Norm unter den Teilnehmern eines praktischen Diskurses Zustimmung nur finden kann, wenn >U< gilt“ (Habermas, 1983, S. 103). U ist das eigentliche Moralprinzip der Diskursethik, da es bestimmt, wie Normen im Diskurs gerechtfertigt werden müssen, um als moralisch richtig zu gelten und so als Regel der Argumentation dient. D hingegen ist die Basis der Moraltheorie, nicht aber Teil der Argumentationslogik (Habermas, 1983, S. 103). D und U spezifizieren die Kernaussage der Diskursethik nach Habermas: Normen sind universell gültig, wenn über ihre Richtigkeit in einem Diskurs ein rationaler Konsens erreicht werden kann (Habermas, 1983, S. 102ff.). Grundsätzlich kann über gültige Normen ein solches Einverständnis erreicht werden, „wobei ›grundsätzlich‹ den idealisierten Vorbehalt meint: wenn die Argumentation nur offen genug geführt und lange genug fortgesetzt werden könnte“ (Habermas, 1983, S. 115).

Anders als der Philosoph und Diskursethiker Karl Otto Apel, dessen Gedanken Habermas Werk nachhaltig geprägt haben, erhebt Habermas mit seiner Begründung der Diskursethik keinen Anspruch auf Letztbegründung, d. h. er unterlässt die Anbindung seiner Theorie an ein untrügliches, unwiderlegbares Wissen (Habermas, 1983, S. 105ff.). Die bereits erläuterten Voraussetzungen der Argumentation sieht er wohl als nicht verwerfbar und alternativlos, grundsätzlich sei es aber möglich, dass Menschen ihre Art über die Welt zu sprechen und zu argumentieren einmal ändern würden (Habermas, 1983, S. 106). In seinen Erläuterungen zur Diskursethik schreibt er: „Dieser Nachweis der faktischen Nichtverwerfbarkeit von normativ gehaltvollen Präsuppositionen einer mit unserer soziokulturellen Lebensform *intern* verschränkten Praxis steht gewiß unter dem Vorbehalt der Konstanz dieser Lebensform. Wir können nicht a priori ausschließen, daß sich diese ändert“ (Habermas, 2009, S. 265, Hervorhebung i. O.). In Habermas Augen ist eine Letztbegründung der Diskursethik für den Nachweis der Geltung seines universalistischen, für alle sprach- und handlungsfähigen Subjekte verbindlichen

Moralprinzips auch nicht notwendig. Dafür genüge der geführte schwach transzendentale Nachweis, dass die Diskursethik und ihr moralischer Universalismus auf den nicht verwerfbaren Voraussetzungen der Argumentation beruhen (Habermas, 2009, S. 265).

Die Diskursethik schließt an die Kantische Moraltradition an. Habermas selbst bezeichnete sie als Versuch, die Moraltheorie Kants „mit kommunikationstheoretischen Mitteln neu zu formulieren" (Habermas, 2009, S. 116). Wie Kant arbeitet auch Habermas mit einem engen Moralbegriff, der sich nur auf Fragen des gerechten oder richtigen Handelns bezieht und evaluative Fragen des guten, gelungenen Lebens ausschließt. Diese fasst er stattdessen in den Bereich der Ethik und legt damit Definitionen der Begriffe Moral und Ethik zugrunde, die von den gängigen Wortbedeutungen abweichen[7] (Habermas, 1983, S. 113f., S. 116ff.). Diese enge Konzeption nimmt er später zurück (vgl. Kapitel 4.2.). Zudem ist die Diskursethik, wie die Kantische Moraltheorie, eine Vernunftethik. Sie nimmt an, dass moralische Pflichten mit rationalen Gründen gerechtfertigt werden können (Seiler, 2014, S. 38*).* Allerdings setzt sie an die Stelle der praktischen die kommunikative Vernunft (Habermas, 1992, S. 117). Als Ethik in der Kantischen Tradition kann die Diskursethik als *kognitivistisch, formalistisch, universalistisch, und deontologisch* bezeichnet werden (Habermas, 2009, S. 118ff.). Der kognitivistische Charakter der Diskursethik nach Habermas besteht in der Annahme, dass sich moralische Aussagen rational begründen lassen (Habermas, 2009, S. 118f.). Formalistisch ist die Diskursethik, da sie wie Kants Ethik und sein kategorischer Imperativ[8] keine konkreten moralischen Regeln liefert, sondern lediglich ein Verfahren zur Begründung von moralischen Normen vorgibt: den diskursethischen Grundsatz D und den Universalisierungsgrundsatz bzw. die Argumentationsregel U (Habermas, 2009, S. 119; Gottschalk-Mazouz, 2000, S. 17). Da Habermas die Prüfung von Normen dialogisiert bzw. diese in den Prozess des Diskurses verlegt, ist seine Diskursethik außerdem *prozeduralistisch* (Gottschalk-Mazouz, 2000, S. 17). Als universalistisch gelten Ethiken, die besagen, dass ihr Moralprinzip nicht nur für bestimmte kulturelle Gruppen oder Epochen, sondern allgemein gilt. Habermas begründet die universalistische Geltung seines Moralprinzips über die normativ gehaltvollen Voraussetzungen der

[7] Üblicherweise bezeichnet der Begriff „Ethik" eine wissenschaftliche Disziplin, die den „Bereich der menschlichen Praxis reflektiert und ihn in evaluativen *sowie* normativen Hinsichten zu beurteilen sucht" (Düwell et al., 2002, S. 2). Moral hingegen beschreibt „die Gesamtheit der Überzeugungen vom normativ Richtigen und vom evaluativ Guten" (Düwell et al., 2002, S. 2). Somit ist die Moral der Gegenstand der Ethik (Düwell et al., 2002).

[8] „Handle nur nach derjenigen Maxime, durch die du zugleich wollen kannst, dass sie ein allgemeines Gesetz werde!" (Habermas, 2009, S. 119)

Kommunikation und der Argumentation (Habermas, 2009, S. 119f.). Die Diskursethik ist deontologisch, da diskursiv eingelöste Normen nach Habermas verpflichtend und unbedingt gültig sind (Habermas, 1992, S. 311).

### 4.1.2. Normen, Werte und der Bezug zur Lebenswelt

Bei Habermas zeigt sich eine Differenzierung zwischen Normen und Werten, die sich von der zu Beginn erläuterten soziologischen Unterscheidung der Begriffe unterscheidet (vgl. Kapitel 1). Sie hängt zusammen mit seiner bereits erwähnten Trennung zwischen Ethik und Moral (Habermas, 1983, S. 113f., S. 116ff.).

Normen haben nach Habermas einen deontologischen Sinn und verpflichten unbedingt zu einem bestimmten Verhalten. Sie sind gültig, wenn ihre normative Richtigkeit in einem rationalen Diskurs eingelöst werden kann. Normen sind binär codiert, sodass sie eindeutig als gültig oder nicht gültig zu beurteilen sind. Zudem ist ihr Sollcharakter universell und absolut: Normen erheben den Anspruch, für alle und überall gut zu sein. Unterschiedliche Normen dürfen sich nicht widersprechen (Habermas, 1992, S. 311). Normen fallen in den Bereich der Moral, d. h. sie beziehen sich auf moralische Fragen, „die unter dem Aspekt der Verallgemeinerungsfähigkeit von Interessen oder der *Gerechtigkeit* grundsätzlich rational entschieden werden können“ (Habermas, 1983, S. 118, Hervorhebung i. O.).

Demgegenüber haben Werte einen teleologischen Sinn. Sie „drücken die Vorzugswürdigkeit von Gütern aus, die in bestimmten Kollektiven als erstrebenswert gelten und durch zielgerichtetes Handeln erworben oder realisiert werden können“ (Habermas, 1992, S. 311). Werte sind intersubjektiv geteilte Präferenzen, bestimmen also die Vorzugsbeziehungen zwischen verschiedenen Gütern. Sie sind nicht binär codiert und können als mehr oder weniger bedeutend empfunden werden. In gewissen kulturellen Kontexten haben Werte relative Geltung, indem sie den ihnen zugehörigen Menschen vorgeben, was gut für sie und die Gruppe ist. Verschiedene Werte stehen miteinander in Konkurrenz, sodass Spannungen entstehen können (Habermas, 1992, S. 311). Werte gehören bei Habermas in den Bereich der Ethik. Sie stehen in Bezug zu evaluativen Fragen, „die sich unter dem allgemeinsten Aspekt als Fragen des *guten Lebens* (oder der Selbstverwirklichung) darstellen und die einer rationalen Erörterung nur *innerhalb* des unproblematischen Horizonts einer geschichtlich konkreten Lebensform oder einer

individuellen Lebensführung zugänglich sind“ (Habermas, 1983, S. 118, Hervorhebung i. O.).

Im Jahr 1981 bestimmte Habermas Werte als nicht diskurs- und begründungsfähig, sie könnten lediglich wie ästhetische Urteilen kritisiert werden. Da Werte keinen Anspruch auf allgemeine, universelle Zustimmungsfähigkeit erheben und immer ein gemeinsames Vorverständnis der Argumentationsteilnehmer voraussetzen würden, erfüllten die Argumentationen zu ihrer Rechtfertigung auch nicht die Voraussetzungen von Diskursen (Habermas, 1981, S. 41f., S. 71). Die Diskussion über ethische Werte sei kein Diskurs, da die Geltung von Werten nur innerhalb gewisser Kontexte plausibel gemacht werden könne (Habermas, 1981, S. 41f., S. 71). Eine solche Abstraktion vom eigenen lebensweltlichen Kontext und von lokalen Übereinkünften fordere nur die Moralität, also der Diskurs über moralische Normen (Habermas, 1983, S. 118f.). Sowohl seinen Ausschluss des Guten bzw. des Ethischen aus dem Bereich der Moral als auch seine Konzeption von Werten als nicht diskursfähig revidiert Habermas später (vgl. Kapitel 4.2.).

Obwohl Habermas Werte zunächst als nicht diskursfähig bestimmt, sind sie schon immer für die Entstehung von Normen relevant. Schließlich stünden Diskurse nicht für sich, sondern seien in lebensweltliche Kontexte eingebunden (Habermas, 1983, S. 113). Damit Diskurse Inhalte hätten, bräuchten sie die Lebenswelt mit ihren realen Konflikten, die der Klärung bedürfen. Habermas geht davon aus, dass erst im Diskurs die ethischen Fragen des guten Lebens bzw. Wertfragen von moralischen Fragen nach dem Gerechten bzw. normativen Fragen getrennt würden: Der Universalisierungsgrundsatz funktioniere wie ein Messer, das einen Schnitt legt zwischen »das Gute« und »das Gerechte«“ (Habermas, 1983, S. 113). Menschen besäßen ein moralisches Alltagswissen, aus dem sie in Konfliktsituationen Normenkandidaten zögen. Normkandidaten, die im Diskurs allgemeine Anerkennung fänden, würden anschließend als deontologische Normen gelten. Stellten sich Normkandidaten hingegen als nicht konsensfähig heraus, so sei damit geklärt, dass es sich bei ihnen bloß um Werte handle. Nach Habermas ist es also Aufgabe der Lebenswelt und der kulturellen Wertsysteme, praktischen Diskursen ihre Inhalte zu geben (Habermas, 1983, S. 113f.).

Wie bereits erwähnt, müssen Individuen laut Habermas von kulturellen und lokalen Wertsystemen abstrahieren, um moralische Normen zu finden. Damit gültige Normen wiederum angewendet werden könnten, sei diese Dekontextualisierung wieder rückgängig zu machen (Habermas, 1983, S. 118f.). Dafür sei die Lebenswelt von zentraler

Bedeutung: Sie erst liefere die Motivation zur Umsetzung von Normen (Habermas, 2009, S. 197f.). Im Diskurs würden unter Anwendung des Moralprinzips U Normen begründet und gerechtfertigt, jedoch verpflichte U „weder zum Eintritt in moralische Argumentationen […], noch zur Befolgung moralischer Einsichten“ (Habermas, 2009, S. 197). Die schwache motivierende Kraft der rationalen, moralischen Einsicht alleine reiche als Motivation für die Umsetzung moralischer Normen nicht aus. Erst durch Sozialisationsprozesse, den Wunsch nach Aufrechterhaltung personaler Identitäten und unterstützende Institutionen erhielten moralische Einsichten genügend „Rückendeckung“ (Habermas, 2009, S. 198) aus der Lebenswelt, um moralisches Handeln anzuregen. Zum Beispiel müssten Individuen, die gegen Normen verstoßen, mit moralischen Vorwürfen ihrer Gruppe rechnen und sich ihrem schlechten Gewissen stellen (Habermas, 2009, S. 198). Nach Habermas braucht die Diskursethik rationalisierte Lebensformen, die „die kluge Applikation allgemeiner moralischer Einsichten ermöglichten und Motivationen für die Umsetzung von Einsichten in moralisches Handeln fördern“ (Habermas, 1983, S. 119).

## 4.2. Erweiterungen und Modifikationen der Diskursethik

### 4.2.1. Anwendungsdiskurse

In Reaktion auf Einwände äußerte sich Habermas spezifischer zur Anwendbarkeit der Diskursethik in konkreten Situationen (Habermas, 2009, S. 200ff.). Zuvor schrieb er zu dieser Thematik, dass zur Anwendung von Normen die vorher vorgenommene Dekontextualisierung wieder rückgängig gemacht werden müsse und das Individuum dazu die Unterstützung der Lebenswelt (vgl. Kapitel 4.1.2.) und eine gewisse „praktische Klugheit“ brauche (Habermas, 1983, S. 114). Später widmete sich Habermas der Anwendungsproblematik näher und führte unter Rückbezug auf Klaus Günther die Differenzierung zwischen *Begründungs-* und *Anwendungsdiskursen* ein (Habermas, 2009, S. 200).

In Begründungsdiskursen gehe es um die Begründung von Normen, die als „generalisierte Verhaltenserwartungen (…) einer allgemeinen Praxis zugrunde liegen“ (Habermas, 2009, S. 200). Dazu werde der Universalisierungsgrundsatz U angewendet: Dieser verlange, dass jede gültige Norm der Bedingung genügt, dass die voraussichtlichen Folgen und Nebenwirkungen ihrer Befolgung von allen akzeptiert

werden könnten (Habermas, 1983, S. 75f.). Nach Habermas darf von den Diskursteilnehmern jedoch nicht erwartet werden, im Begründungsdiskurs alle zukünftig möglichen Situationen zu kennen: Diese Anforderung zu erfüllen sei schlichtweg unmöglich (Habermas, 2009, S. 202). Stattdessen müsse anerkannt werden, dass Teilnehmer an Diskursen immer nur über ihr begrenztes Wissen zu einer bestimmten Zeit verfügen. Diskursiv begründete Normen ließen daher offen, ob sie auch für spezifische, möglicherweise unvorhergesehene Situationen angemessen seien (Habermas, 2009, S. 202). Um die situationsspezifische Angemessenheit von Normen zu bewerten, würden sich Anwendungsdiskurse eignen (Habermas, 2009, S. 202f.). Anstelle der Argumentationsregel U trete in Anwendungsdiskursen das Prinzip der Angemessenheit, das entscheide, ob eine Norm im Hinblick auf die spezifischen Merkmale einer Situation passend sei (Habermas, 2009, S. 203). Der Fokus von Anwendungsdiskursen sei nicht die universelle Geltung einer Norm, stattdessen müsse die bereits für gültig befundene Norm „im Lichte der Situationsmerkmale konkretisiert und die Situation ihrerseits im Lichte der von der Norm vorgegebenen Bestimmungen beschrieben“ (Habermas, 2009, S. 203) werden. Im Fall von Normenkollisionen würden Anwendungsdiskurse dabei helfen, die für die Situation passende Norm zu bestimmen. Normen, die hinter einer im Spezialfall angemessenen Norm zurücktreten, verlören nicht ihre Gültigkeit, „sondern bilden zusammen mit allen anderen gültigen Regeln eine *kohärente Ordnung*“ (sei (Habermas, 2009, S. 204, Hervorhebung i. O.).

Durch die Einführung von Anwendungsdiskursen begrenzt sich die Diskursethik nicht mehr auf die Frage der Richtigkeit moralischer Normen, sondern beschäftigt sich auch mit der Richtigkeit moralischer Handlungen und Urteile (Seiler, 2014, S. 39).

### 4.2.2. Der pragmatische, ethische und moralische Gebrauch der praktischen Vernunft

Eine weitere Neuerung findet sich in Habermas Aufsatz „Der pragmatische, ethische und moralische Gebrauch der praktischen Vernunft“, in dem er erstmals auch *pragmatischen* und *ethischen* Fragen Diskursfähigkeit zuspricht (Habermas, 2009, S. 360ff.). Hier kommt er zu der Einsicht, dass selbst wenn sich der Bereich der Moral (bisher) auf Fragen der Richtigkeit und Gerechtigkeit begrenze, auch andere Probleme, etwa Probleme des guten Lebens, diskursiv erörtert werden könnten. Der Name der Diskursethik sei eventuell irreführend, da sich die Diskurs*theorie* „in je anderer Weise auf moralische, ethische und pragmatische Fragen“ beziehe (Habermas, 2009, S. 361).

Die klassische ethische Frage: Was soll ich tun? habe schließlich je nach Situation und Bewertungsaspekt einen anderen Sinn (Habermas, 2009, S. 361). Pragmatische Probleme kennzeichnen sich nach Habermas dadurch, dass sie gelöst werden müssen, um unangenehme Folgen zu vermeiden. Zu finden seien Gründen für die vernünftige Wahl zwischen unterschiedlichen Möglichkeiten in Bezug auf eine Aufgabe, die zur Erreichung eines Ziels erfüllt werden müsse (Habermas, 2009, S. 362). Es handle sich um eine „rationale Wahl der Mittel bei gegebenen Zwecken oder um die rationale Abwägung der Ziele bei bestehenden Präferenzen“ (Habermas, 2009, S. 362). Als Beispiel für pragmatische Probleme führt Habermas alltägliche Situationen an, wie etwa den Umgang mit gesundheitlichen Beschwerden oder einem kaputten Fahrrad. Alle hier angestellten Überlegungen seien zweckrational und beschränkten sich auf das Finden von geeigneten Methoden zur Erreichung eines Ziels oder zur Verwirklichung eines Wertes (Habermas, 2009, S. 363).

Stehe nun das Ziel oder der Wert selbst infrage, so handle es sich um ein ethisches Problem. Relevant würden hier eigene Prinzipien und das subjektive Selbstverständnis, also Fragen danach, wer man sei und welches Leben man führen wolle. Diese Thematiken seien eng mit der eigenen Identität verbunden und bezögen sich auf das gute Leben. Als Beispiele nennt Habermas hier die Wahl eines Berufs oder eines Lebenspartners (Habermas, 2009, S. 364). Das Habermas Werten nun Diskursfähigkeit zugesteht, ist interessant, da er sich in seinen vorherigen Schriften noch gegensätzlich geäußert hat (Habermas, 1981, S. 41f., S. 71). Er begründet dies, indem er anführt, dass auch die Argumentationsschritte zur Begründung von ethischen Wertentscheidungen „intersubjektiv nachvollziehbar bleiben müssen“ (Habermas, 2009, S. 373). Der einzelne könne sich in einem ethischen Diskurs nicht vertreten lassen, da es in diesem immer auch um seine Identität ginge. Der Kontext, indem er sich bewegt und der seinen Werthorizont bestimmt, sei für ethische Diskurse nach wie vor relevant. Allerdings ist der Fakt, dass Werte nur „im Kontext einer besonderen Lebensform plausibel gemacht“ (Habermas, 1981, S. 71) werden können, nicht mehr Grund für die fehlende Diskursfähigkeit von ethischen Werten (Habermas, 2009, S. 373). Schließlich könne Kritik solcher Werte auch von Menschen außerhalb des eigenen Kontextes kommen, es könnten also auch dem besonderen Kontext nicht angehörende Individuen am ethischen Diskurs teilnehmen. Ethische Werte gelten demnach nicht universell, können aber auch von Außenstehenden als richtig für die Betroffenen, die einer spezifischen gemeinsamen Lebensform angehören, nachvollzogen werden (Gottschalk-Mazouz, 2000, S. 141f.). Während

Normen nach Habermas den Geltungsanspruch der normativen Richtigkeit einlösen, könnten im ethischen Diskurs eingelöste Werte und Lebensentwürfe in Analogie zum Anspruch der Wahrhaftigkeit den Geltungsanspruch der Authentizität erheben (Habermas, 2009, S. 334).

Pragmatische und ethische Probleme sind nach Habermas in „den lebensweltlichen Kontext, also in der Tradition, in der wir leben, eingebettet" (Horster, 2006, S. 92) und außerdem selbstbezogen. Andere Personen würden selbst in ethischen Fragen nur soweit berücksichtigt, wie sie im Zusammenhang mit der eigenen Identität und den eigenen Interessen stünden (Habermas, 2009, S. 366). Erst im dritten Sinn der Frage: Was soll ich tun? werde dieser Kontext verlassen (Habermas, 2009, S. 366). Im Fall von *moralischen* Problemen würden die Handlungen des Einzelnen auch andere betreffen. Nun gelte es, unparteiische Entscheidungen zu treffen, die für alle akzeptabel und allgemein gültig sind. Es gehe darum zu prüfen, ob sich die leitenden Regeln, nach denen sich das eigene Handeln üblicherweise richtet und die Kant "Maximen" nennt, mit den Maximen anderer vertragen würden (Habermas, 2009, S. 367). Maximen könnten grundsätzlich auf zwei Arten geprüft werden: Einmal geleitet von der Frage, ob sie für das einzelne Individuum und dessen Situation angemessen sind, und einmal in Bezug auf die Frage, ob sich die Maxime als allgemeingültiges Gesetz eigne. Im ersten Fall handele es sich um eine ethische Prüfung, nur in Fragen letzterer Art könne von einer moralischen Prüfung gesprochen werden. (Habermas, 2009, S. 368). Um die Frage: Was soll ich tun? moralisch zu klären, müsse sie „mit Bezug auf das, was *man* tun sollte" (Habermas, 2009, S. 369, Hervorhebung i. O.) beantwortet werden.

Insgesamt richtet sich die praktische Vernunft folglich je nach Anwendung auf pragmatische, ethische, oder moralische Probleme „an die Willkür des zweckrational handelnden, an die Entschlusskraft des authentisch sich verwirklichenden oder an den freien Willen des moralisch urteilsfähigen Subjekts" (Habermas, 2009, S. 372).

### 4.2.3. Die Einbeziehung des Guten in den Bereich der Moral

Wie in Kapitel 4.1. dargestellt, beschränkte Habermas die Moral zunächst auf Fragen der Gerechtigkeit und der normativen Richtigkeit. Diese Konzeption hebt er in "Eine genealogische Betrachtung zum kognitiven Bereich der Moral" auf: Hier definiert er Gerechtigkeit als „das für alle gleichermaßen Gute" (Habermas, 2009, S. 337), nimmt also das Gute mit in den Bereich der Moral hinein.

Gerechtigkeit bzw. Moral bedeute nämlich auch Solidarität: Schließlich müssten Individuen ein gewisses Selbstverständnis als zugehörig zur moralischen Gemeinschaft besitzen, um ein moralisches Bewusstsein zu erlangen (Habermas, 2009, S. 337). Da sie ihre Identitäten aber immer nur in interpersonalen Beziehungen und gegenseitiger Anerkennung festigten, seien diese Identitäten verletzlich und bräuchten den Schutz einer höheren Instanz außerhalb des eigenen Kollektivs. Diesen biete die moralische Gesellschaft, in welcher „jede Person jede andere als >>einer von uns<< behandelt" (Habermas, 2009, S. 338). Das, was als Gutes im Gerechten enthalten ist, bezeichnet Habermas als „die Form eines intersubjektiv geteilten Ethos überhaupt und damit die Struktur der Zugehörigkeit zu einer Gemeinschaft, die freilich die ethischen Fesseln einer exklusiven Gemeinschaft abgelegt hat" (Habermas, 2009, S. 339). Das moralisch relevante Gute stehe nicht im Voraus fest, sondern müsse im Diskurs, genauer gesagt im moralischen Diskurs (vgl. Kapitel 4.2.2.), gefunden werden. Wenn ethische Inhalte oder Annahmen in Bezug auf das Gute im moralischen Diskurs eingelöst werden könnten, so zählten sie als moralisch. Auf diese Art nimmt Habermas jene Aspekte des guten Lebens, die verallgemeinerbar sind, mit in den Bereich der Moral auf (Habermas, 2009, S. 337ff.).

### 4.2.4. Die diskursive Rechtstheorie

Basierend auf der Diskursethik entwickelte Habermas zudem eine diskursive Rechtstheorie, die er vor allem 1992 in "Faktizität und Geltung" entfaltete. Diese kann hier nur grob zusammengefasst werden. Mit der diskursiven Rechtstheorie entwirft Habermas ein prozedurales Verfahren für die Entwicklung von Rechtsnormen, dass dem diskursiven Begründungsprinzip von Normen in der Diskursethik gleicht und ebenfalls auf dem diskursethischen Grundsatz D basiert (Horster, 2006, S. 98). Diesen überführt Habermas in ein allgemeines Diskursprinzip, das für die Moral und das Recht gleichermaßen anwendbar ist: „Gültig sind genau die Handlungsnormen, denen alle möglicherweise Betroffenen als Teilnehmer an rationalen Diskursen zustimmen könnten" (Habermas, 1992, S. 138). Im prozeduralen Verfahren zur Begründung von Rechtsnormen würden die normativ gehaltvollen Kommunikationsvoraussetzungen und das Diskursverfahren vom Staat institutionalisiert: „Die Existenzberechtigung des Staates liegt […] in der Gewährleistung eines inklusiven Meinungs- und Willensbildungsprozesses, worin sich freie und gleiche Bürger darüber verständigen, welche Ziele und Normen im gemeinsamen Interesse aller liegen" (Habermas, 1992, S. 329). Die Rechtstheorie lässt Habermas normative Orientierung deutlich hervortreten:

„Der Rechtsstaat [ist] ohne radikale Demokratie nicht zu haben und zu erhalten“ (Habermas, 1992, S. 13).

## 4.3. Kritische Stimmen zur Diskursethik

Habermas Diskursethik wurde von vielen Seiten aufgegriffen, kommentiert, und kritisiert. Eine vollständige Darstellung aller Kritiken und Auseinandersetzungen ist in dieser Arbeit nicht möglich. Sie beschränkt sich daher auf die Darstellung einiger relevanter Aspekte.

Häufig angezweifelt wird beispielsweise Habermas Anspruch, mit der Diskursethik ein rein formales Moralprinzip begründet zu haben (Horster, 2006, S. 106ff.). Detlef Horster merkte dazu an, dass sich Habermas Diskursverfahren nicht, wie von diesem beabsichtigt, als interkulturelle Verständigungsbasis eigne. Schließlich sei die Praxis der Argumentation mit ihrem Bezug zum Verfahren der formalen Logik ein Spezifika der abendländischen Kultur, das in anderen Kulturen hinter alternativen Interaktions- und Verständigungsformen zurücktrete (Horster, 2006, S. 106ff.). Ähnlich wie Horster merkt auch Charles Taylor an, dass es ein kontextloses, wertfreies Verfahren zur Findung von moralischen Normen nicht geben könne. Menschen bewegten sich immer im Horizont gesellschaftlicher Werte und Interpretationen, die ihre eigene Sicht auf die Dinge prägen und strukturieren würden. So basiere auch Habermas Verfahren und seine Bedingungen des rationalen Diskurses auf der hohen Wertschätzung der Vernunft in der westlichen Kultur und der christlichen Idee der universellen Gültigkeit von Normen. Es sei eben nicht, wie behauptet, rein formal, sondern treffe substanzielle, normative Vorannahmen (Horster, 2006, S. 125f.). Derselbe Kritikpunkt findet sich auch bei John Rawls (Horster, 2006, S. 128) und Seyla Benhabib (Horster, 2006, S. 131). Diese identifizieren noch weitere Werte hinter Habermas Diskursethik, wie etwa universelle moralische Achtung und egalitäre Reziprozität (Horster, 2006, S. 132).

Jean-Francois Lyotard kritisiert die Diskursethik aus einer anderen Perspektive: Seiner Meinung nach gibt es kein einheitliches Regelsystem, das auf alle Diskurse angewendet werden kann. Stattdessen existierten unterschiedliche Diskursarten, die verschiedenen Regeln folgen würden. Je nach der Art des Diskurses und der in ihm geltenden Regeln könnten in ein und derselben Situation unterschiedliche Entscheidungen, Aussagen, etc. richtig sein. Konsens sei ein veralteter Wert, wichtiger

sei die Gerechtigkeit: Es brauche daher eine Idee der Gerechtigkeit, die ohne Konsens auskommt (Horster, 2006, S. 138ff.).

Ein weiterer Kritikpunkt an der Diskursethik beinhaltet, dass diese die individuelle moralische Prioritätensetzung nicht genügend berücksichtige (Horster, 2006, S. 113f.). Habermas betone den Aspekt der Gerechtigkeit zu stark und vergesse dabei jene persönlichen moralischen Werte, die sich Individuen selbst setzen würden. Horster geht davon aus, dass die Gerechtigkeitsmoral wohl als gemeinsamer moralischer Rahmen diene, aber in Konfliktsituationen die individuellen Wertpräferenzen zur Entscheidung herangezogen würden. Diese individuelle Prioritätensetzung finde bei Habermas zu wenig Beachtung (Horster, 2006, S. 113f.).

Im Zusammenhang mit diesem Punkt steht Kritik an der rationalitätstheoretischen Ausrichtung der Diskursethik. Hier wird angemerkt, dass moralische Gefühle und deren Beitrag zur Moral in der Diskursethik nicht explizit behandelt würden (Habermas et al., 2016, S. 813f.). Der Mensch sei jedoch ein fühlendes Wesen, dessen Handlungen auch affektiv motiviert seien. Die Vernachlässigung der emotionalen Komponente der Moral wird als Schwäche der Diskursethik gesehen, da ohne Gefühle und emotionale Reaktionen nichts von der Moral zurückbleibe (Habermas et al., 2016, S. 813f.).

Bisher wurden generelle Kritikpunkte erläutert, die sich auf die Voraussetzungen, Annahmen und Ansprüche der Diskursethik beziehen. Neben diesen Kritikpunkten sind weitere zu nennen, die Schwächen innerhalb der Diskursethik aufzeigen, also solche Probleme, die entstehen, wenn ihre grundlegenden Voraussetzungen, Annahmen und Ansprüche akzeptiert werden. Einwände dieser Art diskutiert zum Beispiel Seiler: In der Realität würden niemals alle Betroffenen zu einem Diskurs zusammenkommen können und überhaupt seien nicht alle Betroffenen diskursfähig bzw. in der Lage, rational zu argumentieren. Zudem könnten in der Moderne nur sehr abstrakte Werte allgemeine Zustimmung finden, was zu Anwendungsproblemen führe (Seiler, 2014, S. 47ff.). Habermas führte in Reaktion auf diese Kritikpunkte das Modell des advokatorischen Diskurses ein und verwies darauf, dass eine bloß annähernde Erfüllung der rationalen Argumentationsvoraussetzungen genüge. Um Probleme bei der Anwendung von Normen in konkreten Situationen zu lösen, stellte er Anwendungsdiskurse vor (Seiler, 2014, S. 47ff.). Diese Erwiderungen bringen laut Seiler jedoch neue Probleme mit sich: Die Repräsentation von Individuen durch einen Advokaten widerspreche dem Prinzip der Inklusion aller Betroffenen und die Akzeptanz von Annäherungen lasse offen, wer wie

entscheidet, ob eine Annäherung genügend ist. Das Konzept der Anwendungsdiskurse hingegen arbeite Habermas nicht im Detail aus (Seiler, 2014, S. 47ff.).

Insgesamt sind die Voraussetzungen der Kommunikation und der Argumentation schwer zu erfüllen. Menschen lösen nicht immer alle Geltungsansprüche der Kommunikation ein und eine ideale Sprechsituation wird es in der Realität nie geben. Dem trägt Habermas selbst mit dem Konstrukt der nur kontrafaktischen Unterstellung Rechnung (Habermas, 1983, S. 102). Dieses führt jedoch, wie Seiler aufzeigt, zu einer „Unbestimmtheit darüber, unter welchen Voraussetzungen die im praktischen Diskurs zur Überprüfung gestellten moralischen Normen gültig sein sollen“ (Seiler, 2014, S. 41). Einerseits könne man argumentieren, dass moralische Normen gerechtfertigt sind, wenn die Realität der idealen Sprechsituation ähnelt. Andererseits könne daraus geschlossen werden, dass es gar keine gültigen Normen gibt, da die ideale Sprechsituation schließlich nie erreicht wird. In diesem Punkt ist die Diskursethik arbiträr (Seiler, 2014, S. 41).

## 4.4. Zusammenfassung und Zwischenfazit

Mit seiner Diskursethik entwickelt Habermas ausgehend von den Voraussetzungen der Kommunikation eine Konsenstheorie der normativen Richtigkeit (Seiler, 2014, S. 32ff.), die für die Beantwortung der Frage nach normativen Universalien relevant ist. Er geht davon aus, dass Menschen in ihrer Kommunikation stets Anspruch auf Wahrheit, Richtigkeit und Wahrhaftigkeit erheben und gleichzeitig unterstellen, dass ihre Gesprächspartner diese Ansprüche ebenfalls einlösen. Die Verständlichkeit des Gesagten ist dabei die Bedingung für das Gelingen von Kommunikation (Horster, 2006, S. 50ff.).

In jeder Art des kommunikativen Handelns sind diese Geltungsansprüche implizit. Wenn Zweifel über ihre Einlösung entstehen, können diese im Diskurs thematisiert werden (Gottschalk-Mazouz, 2000, S. 19f.). Diskurse lösen Probleme mithilfe rationaler Argumente und setzen auf logischer und dialektischer Ebene die Einhaltung bestimmter Regeln voraus. Auf der rhetorischen Ebene fordern Diskurse die Existenz einer idealen Sprechsituation (Gottschalk-Mazouz, 2000, S. 20). Diese ist in realen Diskursen nicht gegeben, aber wird von allen Diskursteilnehmern kontrafaktisch unterstellt (Seiler, 2014, S. 33). Während theoretische Diskurse den Gültigkeitsanspruch der Wahrheit prüfen, testen praktische Diskurse den Gültigkeitsanspruch der normativen Richtigkeit. Wird ein diskursiver Konsens erzielt, d. h. eine Übereinstimmung aller Teilnehmer in einem Diskurs, der die genannten Anforderungen erfüllt, so konnte der betroffene

Geltungseinspruch eingelöst werden. In praktischen Diskursen können auf diese Art die wahrheitsanaloge Richtigkeit von Normen begründet und universell gültige Normen gefunden werden (Seiler, 2014, S. 33). Die Kernprinzipien der Diskursethik sind im diskursethischen Grundsatz D und im Universalisierungsgrundsatz bzw. in der Argumentationsregel U enthalten (Gottschalk-Mazouz, 2000, S. 17).

Die Diskursethik kann nicht in Abgrenzung von der Lebenswelt gesehen werden: Einerseits entnimmt sie ihre Inhalte aus den kulturellen Wertesystemen und dem moralischen Alltagswissen der Menschen, andererseits benötigt sie die Unterstützung gelungener Sozialisationsprozesse und regulierender gesellschaftlicher Institutionen, um Menschen zur Umsetzung von diskursiv eingelösten Normen zu motivieren (Habermas, 1983, S. 113f.; Habermas, 2009, S. 197ff.).

Trotz der diversen Kritikpunkte an der Diskursethik (vgl. Kapitel 4.3.), die im Hinterkopf behalten werden sollten, gibt diese auf die Frage nach universellen moralischen Werten eine klare Antwort: Werte, die sich auf Fragen des guten Lebens beziehen, sind kulturell variabel und gelten im Rahmen begrenzter Kollektive. Allerdings gilt unter Einbezug der späteren Erweiterungen der Diskursethik, dass Werte bzw. ethische Grundsätze universelle Geltung erhalten können, eben wenn sie sich als verallgemeinerungsfähig erweisen und in moralischen Diskursen eingelöst werden können (Habermas, 2009, S. 337ff.). Immer universell gültig und für alle Menschen verpflichtend sind laut Habermas Normen. Diese beziehen sich auf Fragen der normativen Richtigkeit und der Gerechtigkeit und können diskursiv eingelöst werden (Habermas, 1992, S. 311).

In der Einleitung dieser Arbeit wurde ausgehend von einer soziologischen Definition des Wertbegriffs vermutet, dass zunächst universelle Werte gefunden werden müssen, aus denen dann Normen abgeleitet werden können. Im Rahmen von Habermas Diskursethik gilt diese Annahme nicht, da Werte und Normen anders definiert werden (Habermas, 1992, S. 311). Mit dem von ihm beschriebenen Diskursverfahren können direkt universelle Normen und eben auch einige universelle Werte gefunden werden (Habermas, 2009, S. 337ff.). Die Diskursethik liefert jedoch keine Sammlung von konkreten universellen Normen. Stattdessen stellt sie ein Verfahren zur Prüfung der universellen Gültigkeit verschiedener Normenkandidaten bereit, das vor allem auf den von Habermas identifizierten normativ gehaltvollen Voraussetzungen der Kommunikation und des Diskurses basiert.

# 5. Diskursethik und künstliche Intelligenz zur Identifikation von normativen Universalien

Dieses Kapitel widmet sich dem zweiten Teil der Forschungsfrage, der Umsetzung der gewonnen Erkenntnisse in der künstlichen Intelligenz. Angesetzt wird hier bei dem Habermassen Diskursverfahren: Im moralischen Diskurs können über einen diskursiven Konsens aus einer Auswahl von Normenkandidaten universell gültige Normen gewonnen werden. Diskurse beruhen auf den Voraussetzungen der Kommunikation und der Argumentation, die in der Realität nie erfüllt werden. Daraus ergeben sich Probleme bei der Umsetzung der Diskursethik und bei der Entscheidung darüber, wann in moralischen Diskursen gefundene Normen tatsächlich gültig sind. Im Folgenden beschäftigt sich die Arbeit daher mit der Möglichkeit, Diskurse nach dem Habermassen Ideal mit den Methoden der künstlichen Intelligenz zu simulieren. Ziel der Simulation ist es, die in Habermas Theorie nur kontrafaktischen Voraussetzungen tatsächlich zu erfüllen und so einen Teil der weiter oben diskutierten Probleme innerhalb der Diskursethik zu lösen (vgl. Kapitel 3).

Dazu wird zunächst ein Konzept zur automatisierten ethischen Entscheidungsfindung vom Massachusetts Institute of Technology vorgestellt und auf Parallelen zur Diskursethik untersucht. In Kombination mit einem weiteren Teilgebiet der künstlichen Intelligenz, der verteilten künstlichen Intelligenz und ihren Multiagentensystemen, wird ein Vorschlag für eine Erweiterung dieses Konzepts dargelegt. Dieser kann möglicherweise bei der Entwicklung bzw. der Entdeckung von universellen Normen unterstützen. Beschränkungen und Probleme des Vorschlags werden im letzten Teil des Kapitels diskutiert.

## 5.1. Ein abstimmungsbasiertes System zur ethischen Entscheidungsfindung

Noothigattu und Kollegen stellten 2018 ein abstimmungsbasiertes System zur ethischen Entscheidungsfindung von KI-Systemen vor (Noothigattu et al., 2018). Sie sind der Meinung, dass sich die ethische Entscheidungsfindung von KI-Systemen automatisieren lässt, selbst wenn im Vorhinein keine konkreten ethischen Prinzipien feststehen. In ihrem vorgeschlagenen System nutzen sie Methoden des maschinellen Lernens und der computergestützten Sozialwahltheorie (Noothigattu et al., 2018, S. 1). Maschinelles Lernen wurde als Methode der künstlichen Intelligenz bereits im zweiten Kapitel

vorgestellt. Die computergestützte Sozialwahltheorie hingegen ist ein junges Teilgebiet der künstlichen Intelligenz, in dem computerbasierte Techniken eingesetzt werden, um Mechanismen der sozialen Entscheidungsfindung zu erforschen und zu modellieren und neue Mechanismen zu entwickeln (Brandt et al., 2016, S. 2). Ein Fokus der Disziplin liegt auf der Analyse und dem Design von Abstimmungsregeln, also der Aggregation individueller Meinungen zu kollektiven Entscheidungen (Brandt et al., 2016, S. 9). Das System von Noothigattu und Kollegen soll als Basis für die Einbeziehung zukünftiger ethischer und rechtlicher Prinzipien in KI-Systeme dienen, aber auch vorläufige Antworten auf einige ethische Fragen liefern, mit denen sich die künstliche Intelligenz konfrontiert sieht (Noothigattu et al., 2018, S. 2).

Es besteht aus vier Schritten, in denen ein Algorithmus entwickelt wird. Die technischen Hintergründe dieses Algorithmus werden an dieser Stelle nicht erläutert und sind für die Arbeit auch nicht relevant. Dargestellt werden soll lediglich die generelle Idee hinter dem Vorschlag: Im ersten Schritt werden Daten über die Entscheidungen verschiedener Menschen in moralischen Konfliktsituationen gesammelt. Diese Daten müssen in Form von Entscheidungen zwischen jeweils zwei möglichen Alternativen in verschiedenen moralischen Konfliktsituationen vorliegen. Jede mögliche Alternative wird dabei über ihre Eigenschaften charakterisiert, wie etwa die Anzahl negativ betroffener Menschen und deren Alter und Geschlecht (Noothigattu et al., 2018, S. 2). Mithilfe dieser Daten und Methoden des maschinellen Lernens wird im zweiten Schritt ein Modell der Präferenzen eines jeden Teilnehmers über alle möglichen Alternativen gelernt. Nach Abschluss des Lernprozesses existieren dann Repräsentationen der Präferenzen jedes einzelnen Befragten, die auf seinen Entscheidungen in den paarweisen Vergleichen basieren (Noothigattu et al., 2018, S. 2, S. 15). Die Präferenzmodelle der einzelnen Teilnehmer werden anschließend zu einem einzigen Modell zusammengefasst, welches die kollektiven Präferenzen aller Befragten über alle möglichen Alternativen enthält (Noothigattu et al., 2018, S. 15). Dieses Model wird eingesetzt, wenn das KI-System vor einer ethischen Konfliktsituation steht und zwischen verschiedenen Alternativen wählen muss. Aus dem zusammengefassten Model können die Präferenzen aller Befragten in Bezug auf das spezifische Problem, also die präferierte Alternative jedes Befragten in der vorliegenden Problemsituation, abgeleitet werden. Mithilfe der Präferenzen der Einzelnen kann unter Anwendung einer Abstimmungsregel eine kollektive Entscheidung über die zu wählende Alternative gefunden werden. Anders

ausgedrückt findet eine Abstimmung zwischen den Präferenzen der Individuen statt (Noothigattu et al., 2018, S. 15).

Noothigattu und Kollegen nutzten in ihrem Beispiel die Ergebnisse der Onlineplattform *The Moral Machine:* Ein Datensatz mit 18.254.285 paarweisen Vergleichen zwischen Alternativen in ethischen Konfliktsituationen aus dem Bereich des autonomen Fahrens von insgesamt 1.303.778 Menschen (Awad et al., 2018). So entwickelten sie ein System, dass glaubwürdige Entscheidungen in ethischen Dilemmas beim autonomen Fahren treffen kann (Noothigattu et al., 2018, S. 20).

Während es in der Diskursethik um die Prüfung von moralischen Normen geht, beschäftigt sich das System von Noothigattu und Kollegen mit dem Treffen von Entscheidungen in ethischen Konfliktsituationen. Bei genauerer Betrachtung zeigen sich einige Parallelen zwischen der Diskursethik und dem abstimmungsbasierten Entscheidungssystem von Noothigattu und Kollegen. Bevor diese erläutert werden, muss klargestellt werden, dass die Begriffe „Ethik“ und „ethische Entscheidungsfindung“ von Noothigattu und Kollegen anders verwendet werden als von Habermas. Sie spezifizieren ihre Begriffe nicht, da ihre Arbeit sich vor allem auf die Vorstellung und Erklärung des entworfenen Systems beschränkt. Allerdings lassen die im Text genutzten Beispiele darauf schließen, dass sich Noothigattu und Kollegen nicht ausschließlich auf Fragen des guten Lebens, sondern genauso auf Problemen der normativen Richtigkeit und der Gerechtigkeit beziehen, also jene Probleme, die bei Habermas von Anfang an in den Bereich der Moral fallen (Noothigattu et al., 2018, S. 1, S. 20).

Eine Parallele zur Habermassen Diskursethik besteht darin, dass das Modell von Noothigattu und Kollegen annimmt, dass ethische Entscheidungen unter Einbeziehung von individuellen Meinungen kollektiv getroffen werden können (Noothigattu et al., 2018, S. 1f.). Dabei zählen die Präferenzen eines jeden Individuums gleich viel. Das zeigt sich besonders daran, dass Noothigattu und Kollegen in ihrer Arbeit von anonymen Präferenzprofilen der Individuen sprechen und explizit schreiben: „the identity of voters does not play a role“ (Noothigattu et al., 2018, S. 5). Durch ihre Anonymität haben alle Beteiligten, bzw. die Repräsentationen aller Beteiligten im Rahmen des Modells, die gleichen Rechte und Möglichkeiten und es existieren keine Machtgefälle, was Habermas idealer Sprechsituation entspricht (Habermas, 1983, S. 98). Nicht erfüllt wird jedoch die Voraussetzung der Integration aller Betroffenen: Noothigattu und Kollegen erwähnen in ihrem Text nicht, dass alle von der Entscheidung Betroffenen berücksichtigt und ins System mitaufgenommen werden müssen (Noothigattu et al., 2018, S. 1). Sie nutzen in

ihrem Beispiel für ihr Vorhaben geeignete Daten von einer Auswahl an Menschen, die sicherlich nicht alle Betroffenen repräsentieren (Noothigattu et al., 2018, S. 18).

Darüber hinaus ist im System von Noothigattu und Kollegen die Annahme impliziert, dass Menschen verschiedene Präferenzen besitzen und sich in unterschiedlichen Situationen nicht immer einig sind. Davon geht auch Habermas aus: Wenn er von der nur intersubjektiven Gültigkeit von Werten in gewissen kulturellen Kontexten spricht, erkennt er an, dass sich die Werte verschiedener Menschen unterscheiden können (Habermas, 1992, S. 311). Werte beziehen sich bei Habermas auf ethische Fragen des guten Lebens und fielen in seiner Grundkonzeption der Diskursethik nicht in den Bereich der Moral (Habermas, 1983, S. 113ff.; Habermas, 1992, S. 311). Habermas stellte später fest, dass gewisse Werte in den Bereich der Moral fallen und verallgemeinert werden können. Es hinterbleiben jedoch weiterhin solche Werte, die nur in gewissen kulturellen Kontexten Geltung besitzen (Habermas, 2009, S.337ff.). Um trotzdem zu einer kollektiven Entscheidung zu finden, nutzen Noothigattu und Kollegen ein Abstimmungsfahren. Die auf diese Art getroffene Entscheidung wird jedoch nicht von allen Betroffenen gleichermaßen unterstützt, sondern lediglich von der nach einer Abstimmungsregel gebildeten Mehrheit (Noothigattu et al., 2018, S. 2, S. 5f.). Das entspricht nicht dem von Habermas vorgeschlagenen Verfahren des moralischen Diskurses, in dem für gültige Normen eine von allen Betroffenen getragene Zustimmung erreicht werden muss (Habermas, 1983, S. 73ff.).

Das System von Noothigattu und Kollegen kann einen solchen Diskurs nicht leisten, da die in ihm enthaltenen Modelle individueller Präferenzen der Betroffenen nicht über die Fähigkeiten verfügen, die sie für das Eintreten in einen Diskurs benötigen würden. Sie können lediglich zahlenmäßig kombiniert und mit verschiedenen Abstimmungsregeln zusammengefasst werden, nicht aber agieren und kommunizieren (Noothigattu et al., 2018, S. 2f., S. 5ff.). Die Diskursethik hingegen arbeitet mit autonomen Individuen, die in subjektive und kulturelle Lebenswelten eingebettet sind und umfassende Handlungsfähigkeiten besitzen. Gerade die kommunikative Vernunft und die mit ihr verbundene Einlösung der Geltungsansprüche der Wahrheit, Richtigkeit und Wahrhaftigkeit ist für die Diskursethik und die Begründung der wahrheitsanalogen Richtigkeit von diskursiv eingelösten Normen zentral (Habermas, 1992, S. 17f.; Habermas, 1983, S. 68). Das abstimmungsbasierte System von Noothigattu und Kollegen repräsentiert Menschen vereinfacht anhand von nicht-diskursfähigen Modellen ihrer Präferenzen zwischen verschiedenen Handlungsalternativen. Darin besteht die größte

Differenz zwischen dem System von Noothigattu und Kollegen und der Diskursethik. Um diese Differenz zu überwinden und dabei einige Kernideen von Noothigattu und Kollegen beizubehalten, soll das abstimmungsbasierte System zur ethischen Entscheidungsfindung im Folgenden mit den Multiagentensystemen der verteilten künstlichen Intelligenz kombiniert werden.

## 5.2. Ein Modellvorschlag für die künstliche Intelligenz: Diskurse simulieren

Multiagentensysteme (MAS) stehen im Zentrum eines Teilgebiets der künstlichen Intelligenz, der verteilten künstliche Intelligenz (VKI). Diese kann wie folgt definiert werden: VKI „is the study, construction, and application of multiagent systems" (Weiss, 1999, S. 1). Während sich die künstliche Intelligenz mit der Entwicklung von KI-Systemen beschäftigt, die für sich gesehen intelligent sind, konzentriert sich die VKI auf Intelligenz als Eigenschaft von Systemen (Weiss, 1999, S. 4f.). MAS sind dann Systeme, in denen mehrere *interagierende, intelligente Agenten* (von nun an kurz: Agenten) ein gemeinsames Ziel verfolgen oder gemeinsam bestimmte Aufgaben erledigen (Weiss, 1999, S. 1). Der wichtigste Bestandteil eines MAS sind seine Agenten: Als autonome Recheneinheiten nehmen sie ihre Umgebung wahr und wirken auf diese ein. Sie verfolgen eigene Ziele und führen Aufgaben so aus, dass sie bestimmte Leistungsparameter optimieren. Sie sind nicht allwissend, allmächtig oder unfehlbar, können jedoch in unterschiedlichen Umgebungen flexibel und rational mit denen ihnen zur Verfügung stehenden Informationen arbeiten. Agenten können in ihrem Verhalten von anderen Agenten oder von Menschen beeinflusst werden, etwa indirekt durch die Beobachtung ihrer Umgebung oder direkt durch eine gemeinsame Sprache (Weiss, 1999, S. 2f.).

MAS bieten genau das, was im abstimmungsbasierten System von Noothigattu und Kollegen fehlt: Die Möglichkeit, autonome Agenten zu entwickeln, die mehr sind als nur mathematische Repräsentationen von Präferenzen in ethischen Konfliktsituationen. Die Agenten eines MAS sind wie eben erläutert autonom: Sie können selbstständig handeln, ihre Umwelt beobachten und sich mit den anderen Agenten austauschen (Weiss, 1999, S. 2f.). Insgesamt scheinen sie besser dafür geeignet, Diskursteilnehmer bzw. handelnde und kommunizierende Menschen mit subjektiven Zielen und Werten zu repräsentieren.

Aus diesem Grund wird vorgeschlagen, die Möglichkeiten von MAS zu nutzen. Basierend auf den Kernideen des abstimmungsbasierten Systems zur ethischen Entscheidungsfindung von Noothigattu und Kollegen und der Diskursethik soll ein MAS

zur Identifikation von Normen entwickelt werden. Ziel des Modells ist es, moralische Diskurse nach der Habermassen Vorstellung zu simulieren und auf diesem Weg gültige Normen zu finden. Das Modell stützt sich auf die Diskursethik, versucht aber, einen Teil ihrer Probleme zu umgehen: Es sorgt dafür, dass die Kommunikations- und Argumentationsvoraussetzungen tatsächlich erfüllt werden, erleichtert den Einbezug aller Betroffenen in den Diskurs und stellt deren Diskursfähigkeit sicher. So löst es den Großteil der bereits dargestellten Probleme innerhalb der Theorie (vgl. Kapitel 4.3.). Es simuliert einen moralischen Diskurs unter idealen Bedingungen, oder kommt diesen zumindest sehr nahe.

In dem MAS zur Prüfung von Normen würden die verschiedenen Agenten jeweils unterschiedliche Individuen mit ihren Wertvorstellungen und normativem Alltagswissen repräsentieren. Durch Interaktionen in einem moralischen Diskurs könnten die Agenten dann mithilfe einer gemeinsamen Sprache über potenzielle Normenkandidaten diskutieren und gültige Normen finden. Um dieses Konzept zu realisieren, müsste in drei Schritten vorgegangen werden:

1. *Ausstattung von Agenten mit kulturellen und persönlichen Werthorizonten*

Um ein Individuum zu repräsentieren, müssen die Agenten im MAS über ein Modell seiner kulturellen und persönlichen *Werthorizonte* verfügen. Darunter werden in der vorliegenden Arbeit die Wertvorstellungen des Individuums, sein normatives Alltagswissen, seine Ansichten und Meinungen und seine Einbettung in kulturelle und soziale Umgebungen verstanden. Im Sinne Habermas müsste dabei sichergestellt werden, dass die Agenten des MAS in der Kommunikation über ihre Werthorizonte die Geltungsansprüche der Wahrheit, Richtigkeit, und Wahrhaftigkeit einlösen (vgl. Kapitel 4.1.). Um Agenten Werthorizonten an die Hand zu geben, stehen verschiedene Methoden zur Verfügung, die jeweils unterschiedliche Vor- und Nachteile mit sich bringen. Werthorizonte können unter Anwendung von Methoden des maschinellen Lernens nach dem bottom-up Ansatz erlernt oder nach dem top-down Ansatz in Form von eindeutigen Vorgaben implementiert werden. Möglich ist auch eine Kombination der beiden Paradigmen in hybriden Ansätzen (vgl. Kapitel 2.3). Aktuell beschäftigen sich viele Wissenschaftler mit den verschiedenen Möglichkeiten (z. B. Soares, 2016; Hadfield-Menell et al., 2016; Arkoudas, 2005; Sotala, 2016), eine finale oder beste Lösung ist jedoch noch nicht in Sicht.

2. *Konstruktion einer Sprache für den moralischen Diskurs zwischen Agenten*

Damit die Agenten des MAS miteinander kommunizieren und diskutieren können, benötigen sie eine gemeinsame Sprache. Da in moralischen Diskursen nach dem Habermassen Ideal rationale Argumente ausgetauscht werden sollen (vgl. Kapitel 4.1.), müsste eine solche Sprache in der Lage sein, rationale Argumente logisch zu repräsentieren. Kommunikation zwischen Agenten ist in MAS jeder Art essenziell, sodass sich viele wissenschaftliche Arbeiten mit Kommunikations- und Argumentationsstrategien für MAS auseinandersetzen (z. B. Singh, 1991; Woolridge, 1998; Amgoud et al., 2002; Torreño et al., 2010; Huget, 2003).

*3. Simulation eines moralischen Diskurses nach einem Diskursprotokoll*

Um die Agenten des MAS in einen moralischen Diskurs eintreten zu lassen, wird neben einer gemeinsamen Sprache auch eine Art Diskursprotokoll benötigt. Dieses müsste dafür sorgen, dass die simulierten Diskurse die von Habermas spezifizierten Voraussetzungen erfüllen. Besonders wichtig ist dabei, dass eine ideale Sprechsituation herrscht und alle Agenten gleichberechtigt am Diskurs teilnehmen können (vgl. Kapitel 4.1.). Darüber hinaus müsste der Verlauf des Diskurses modelliert werden. Beispielsweise sollten Diskurse mit dem Einführen eines Normenkandidaten beginnen. Im nächsten Schritt müssten Argumenten für bzw. gegen die Gültigkeit des Normenkandidaten ausgetauscht werden. Am Ende des Diskurses sollte die Zustimmung bzw. Ablehnung der Norm von allen Agenten geprüft werden. Normenkandidaten, die im Diskurs Zustimmung von allen Agenten erhalten, müssten gespeichert, jene, die keine allgemeine Zustimmung erzielen, verworfen werden. Mit unterschiedlichen Arten von leitenden Protokollen für MAS beschäftigen sich unter anderem Rahwan und Kollegen (2008), Poslad (2007) und Mineau (2003).

Sollten die drei vorgeschlagenen Schritte bewältigt werden, so ergäbe sich daraus ein computersimulierter Diskurs zwischen den verschiedenen Agenten im MAS. Der Vorteil des vorgeschlagenen Modells besteht darin, dass es den Großteil jener Kritikpunkte an der Diskursethik, die sich auf Probleme innerhalb der Theorie beziehen, umgehen kann. In realen Diskursen müssen Diskursteilnehmer laut Habermas sowohl die Einlösung der Geltungsansprüche der Kommunikation durch ihre Gesprächspartner als auch die Existenz einer idealen Sprechsituation kontrafaktisch unterstellen (Habermas, 1983, S. 68, S. 102). Das führt zu Unklarheiten darüber, unter welchen Bedingungen Normen überhaupt gültig sind (Seiler, 2014, S. 41). In einem simulierten Diskurs im MAS hingegen können diese Voraussetzungen tatsächlich erfüllt werden. Schließlich haben die rationalen Agenten eines MAS keinen Grund zu lügen, zu diskriminieren, einander zu

beherrschen oder sonstiges für einen Diskurs schädliches Verhalten zu zeigen, solange sie nicht entsprechend programmiert werden oder mit verzerrten Trainingsdaten lernen. Darüber hinaus kann ein entsprechend angelegtes Diskursprotokoll dafür sorgen, dass alle Agenten im Diskurs die gleichen Rechte und Möglichkeiten haben und eine ideale Sprechsituation besteht.

Gelöst werden außerdem Probleme der Umsetzung von Diskursen in der Realität: Ein simulierter moralischer Diskurs zwischen Betroffenen ist einfacher umzusetzen als ein realer Diskurs, dessen Zustandekommen zeitliche und räumliche Probleme im Weg stehen (Seiler, 2014, S. 47f.). Die Agenten im MAS haben unendlich viel Zeit für ihre Diskurse und müssen diese nicht, wie Menschen, neben andere Pflichten stellen. Zudem können sie Diskurse deutlich schneller führen als Menschen. Auch die sonst hinderliche räumliche Trennung von Menschen, die in unterschiedlichen Teilen dieser Welt leben und die nur schwer zu Diskursen zusammentreffen können, wird im Modell überwunden. Schließlich werden die Einzelnen von Agenten repräsentiert und müssen nicht selber vor Ort sein. Eine Repräsentation aller Betroffenen im Diskurs ist dabei selbst in einer Simulation kaum zu erreichen. Auch die Simulation müsste wahrscheinlich mit repräsentativen Stichproben arbeiten und etwa Gruppen ähnlicher Menschen durch einen Agenten repräsentieren. Im vorgeschlagenen Modell ist das Ziel der Beteiligung aller Betroffenen dennoch leichter zu erreichen als in einem realen Diskurs.

Des Weiteren löst das Modell das Problem der nicht gegebenen Diskursfähigkeit aller Betroffenen und der Akzeptanz von Annäherungen (Seiler, 2014, S. 47f.). Die Agenten des MAS werden nicht durch psychische, körperliche oder sonstige Faktoren daran gehindert, rational zu argumentieren. Sie sind, solange keine Fehler in ihrer Programmierung bestehen, jederzeit diskursfähig.

Ein nach den vorgeschlagenen Schritten realisiertes Modell würde einen großen Beitrag zur Identifikation von universellen Normen leisten. Diese könnten dann die Grundlage für die Ausrichtung von KI-Systemen an menschlichen Moralvorstellungen bilden. Allerdings ergeben sich für das beschriebene Modell einige Beschränkungen und Probleme, die im Anschluss diskutiert werden.

## 5.3. Limitationen

Der hier entwickelte Vorschlag ist rein theoretischer Natur und bedarf der Ergänzung praktischer sowie technischer Ansätze seitens der künstlichen Intelligenz. Herauszufinden ist unter anderem, welche Schwierigkeiten die praktische Entwicklung des vorgestellten MAS mit sich bringt und welche Einschränkungen aus diesen folgen würden. Unklar ist auch, inwieweit dieses aufwendige Modell mit den Rechenkapazitäten heutiger Computer bewältigt werden kann. Zum Beispiel ist zu prüfen, mit wie vielen Informationen die Agenten lernen bzw. programmiert werden können und wie tiefgehend und ausführlich diese Informationen sein dürfen. Geklärt werden muss darüber hinaus, wie viele Menschen oder Gruppen in einem MAS als Agenten repräsentiert werden können. Mit diesen Fragen sollten sich zukünftige Forschungsprojekte auseinandersetzen.

Im Vorschlag wurde nicht festgelegt, ob das konzipierte MAS lediglich zum Zweck der Generierung eines universell gültigen Normenkatalogs eingesetzt wird oder ob es gleichzeitig auch andere Aufgaben erfüllt. In ersterem Fall würde das MAS speziell für die Prüfung von Normenkandidaten bzw. die Entwicklung eines Normenkatalogs genutzt. Seine Ergebnisse müssten dann in andere KI-Systeme überführt werden. Im zweiten Szenario würde das MAS in Echtzeit Normenkandidaten prüfen, die für die Lösung von Problemen in seinem jeweiligen Aufgabenbereich relevant sind. Diese Entscheidung hängt sowohl von den technischen Möglichkeiten als auch von den Anforderungen der Einsatzbereiche von MAS ab. Beides sollte geprüft werden.

Zudem ist anzumerken, dass selbst eine erfolgreiche Simulation von Diskursen nach dem Habermassen Ideal im vorgeschlagenen Modell keinen nicht hinterfragbaren, unbedingt gültigen Normenkatalog garantieren würde. Schließlich handelt es sich bei dem dargelegten Vorschlag um die Umsetzung einer Theorie, die keine unbedingte Gültigkeit besitzt, die sich als fehlerhaft und unangemessen herausstellen kann und die schon häufig aus verschiedenen Gründen kritisiert und hinterfragt worden ist (vgl. Kapitel 4.3.). Während ein Großteil der Kritikpunkte an der Diskursethik, die auf Probleme innerhalb der Theorie zielen, auf den erfolgreich umgesetzten Vorschlag nicht länger zutreffen würden, greift die grundsätzliche Kritik an ihren Voraussetzungen, Annahmen und Ansprüchen weiterhin. Vorwürfen, wie jenen von Horster, Taylor, Benhabib und Rawls, die am Formalismus der Diskursethik zweifeln und ihr vorausgehende substanzielle, normative Annahmen identifizieren (Horster, 2006, S. 125ff.), muss sich

auch das Modell stellen. Weiterhin geht das Modell ebenfalls von einem einheitlichen Regelsystem für alle Diskurse aus und stellt Konsens in Verbindung mit moralischer Gerechtigkeit, sodass Lyotards Kritik relevant bleibt. Auch das vorgeschlagene Modell ist rationalitätstheoretisch und kann für einen mangelnden Einbezug individueller moralischer Prioritätensetzung (Horster, 2006, S. 113f.) und moralischer Gefühle (Habermas et al., 2016, S. 813f.) kritisiert werden.

Auf der anderen Seite würde das Versagen des vorgeschlagenen Modells nicht direkt die Gültigkeit der Habermassen Diskursethik widerlegen, da durchaus denkbar ist, dass gewisse Aspekte der menschlichen Kommunikation und Argumentation sich mit den gegenwärtigen Methoden der künstlichen Intelligenz nicht repräsentieren lassen oder sich sogar niemals repräsentieren lassen werden.

Fraglich ist weiterhin, ob sich im Rahmen des Modells gefundene Normen überhaupt für KI-Systeme eignen. Die Habermasse Diskursethik und das auf ihr basierende vorgestellte Modell beschäftigen sich schließlich mit Normen, die menschliches Zusammenleben regulieren, nicht mit Normen für die Kontrolle der Handlungen und Entscheidungen von KI-Systemen. Ob für KI-Systeme dieselben Regeln wie für Menschen gelten sollten, muss diskutiert werden. Relevant ist diese Frage vor allem vor dem Hintergrund, dass KI-Anwendungen im Fall der Entwicklung von genereller KI oder einer Superintelligenz (vgl. Kapitel 2.1.) in Zukunft große Macht besitzen und weitreichende Entscheidungen treffen könnten.

Weitere Probleme ergeben sich, wenn bedacht wird, dass in den simulierten Diskursen wahrscheinlich ausschließlich generalisierte Normen gefunden werden. Schließlich erkennt Habermas selbst an, dass die Normen, die in einem moralischen Diskurs mit allen Individuen eingelöst werden können, in der pluralisierten modernen Gesellschaft immer allgemeiner und abstrakter werden (Habermas, 2009, S. 131). Abstrakte Normen müssen nach Habermas in Anwendungsdiskursen auf ihre Angemessenheit in konkreten Situationen geprüft werden (Habermas, 2009, S. 201ff.). Gerade für die mangelnde Erläuterung dieses Konzepts der Anwendungsdiskurse wurde Habermas, wie oben beschrieben, kritisiert (Seiler, 2014, S. 47ff.). Das ist ein Kritikpunkt in Bezug auf die Probleme innerhalb der Theorie, den das vorgeschlagene Modell nicht lösen kann. Schließlich spezifiziert es nicht, wie mit den gefundenen Normen umgegangen werden soll und wie KI-Systeme diese anwenden können. Wenn schon in Bezug auf die Anwendung diskursiv eingelöster Normen in konkreten Situationen durch Menschen Unklarheiten bestehen, ergeben sich wahrscheinlich mindestens genauso

große Probleme für die Anwendung solcher Normen durch KI-Systeme. Eine mögliche Lösung könnte im Einsatz von hybriden Ansätzen liegen, in denen KI-Systeme vorgegebene abstrakte Prinzipien durch verschiedene Lernstrategien an verschiedene Kontexte anpassen und top-down und bottom-up Methoden miteinander verbinden (vgl. Kapitel 2.3.).

Unabhängig von der Anwendung abstrakter Normen wächst nach Habermas „der Umfang regelungsbedürftiger Materien, die nur noch partikulare Interessen berühren und daher auf die Aushandlung von Kompromissen, nicht auf diskursiv erzielte Konsense, angewiesen sind“ (Habermas, 2009, S. 132, Hervorhebung i. O.). Rahmenbedingungen für solche Kompromissbildungen schlägt Habermas nicht vor, er merkt lediglich an, dass diese moralisch gerechtfertigt werden müssten (Habermas, 2009, S. 132). Voraussichtlich ergeben sich auch im Bereich der künstlichen Intelligenz viele moralische Fragen, die mit Kompromissen zu regeln sind. Hier bleiben Probleme offen, die das vorgeschlagene Modell mit seinen simulierten Diskursen nicht lösen können wird.

Trotz einiger Einschränkungen des Modells ist davon auszugehen, dass seine Umsetzung die Diskussion um normative Universalien vorantreiben wird, die vor dem Hintergrund der Fortschritte in der künstlichen Intelligenz stetig an Relevanz gewinnt. In einem simulierten Diskurs bestätigte Normen können nicht ohne weiteres als universell gültig übernommen werden, jedoch sind sie richtungsweisend und sollten näher untersucht werden.

# 6. Fazit

Ziel dieser Arbeit ist es, aus einer soziologischen Perspektive neue Erkenntnisse über mögliche normative Universalien für die KI-Systeme der Zukunft zu gewinnen. Die Auseinandersetzung mit Habermas Diskursethik hat sich dafür als fruchtbar herausgestellt: Mit ihr lässt sich basierend auf den normativ gehaltvollen Voraussetzungen der Kommunikation für die Existenz von universellen Normen argumentieren. Gleichzeitig stellt sie ein Verfahren für die Entdeckung von gültigen Normen zur Verfügung.

Den ersten Teil der Forschungsfrage, die Frage nach der Existenz universeller Werte, beantwortet die Grundkonzeption der Diskursethik, wie bereits erläutert, noch negativ (vgl. Kapitel 4.4.). Jedoch gibt es in der Erweiterung der Diskursethik durch den Einbezug des Guten in den Bereich der Moral grundsätzlich die Möglichkeit universeller Werte. Weitaus wichtiger ist aber, dass mit der Diskursethik für die Existenz universeller Normen argumentiert werden kann. Da Habermas Normen und Werte in Abweichung von den soziologischen Definitionen der Begriffe bestimmt, können mit der Diskursethik direkt universelle Normen gefunden werden. Die anfangs aufgestellte Vermutung, dass zunächst universelle Werte gefunden werden müssen, aus denen dann Normen abgeleitet werden können, konnte aus der Perspektive der Diskursethik verworfen werden. Nachträglich wäre es sinnvoller gewesen, in der Forschungsfrage direkt nach universellen Normen zu fragen. Diese Einsicht hat sich jedoch erst aus der Auseinandersetzung mit der Diskursethik ergeben und kann daher als Erkenntnis der Arbeit gelten.

Das Ergebnis in Bezug auf den ersten Teil der Forschungsfrage lautet dann wie folgt: Erstens gibt es nach der Diskursethik universelle Werte, jedoch sind nicht alle Werte universell. Zweitens argumentiert die Diskursethik für die Existenz universeller und verpflichtender Normen. Konkrete universelle Normen erwähnt die Diskursethik aber nicht, sie stellt ausschließlich ein formales Verfahren zur Prüfung von Normen zur Verfügung. Wenn man sich der Frage aus diskursethischer Perspektive widmet, müssen drittens nicht erst universelle Werte gefunden werden, um aus diesen universelle Normen zu gewinnen. Die Diskursethik bietet einen Weg zur Identifikation von universellen Normen ohne den Umweg über universelle Werte.

Im zweiten Teil der Forschungsfrage hat sich die Arbeit mit einer Möglichkeit auseinandergesetzt, dass Diskursverfahren aus der Diskursethik und die Methoden der künstlichen Intelligenz zu nutzen, um universelle Normen für die künstliche Intelligenz zu identifizieren. Dazu wurde ein Modell zur Simulation von Diskursen in MAS vorgeschlagen, dass einige Schwächen der Diskursethik überwinden kann. Die Probleme durch die nur kontrafaktische Unterstellung der Voraussetzungen der Kommunikation und der Argumentation können gelöst werden, indem die Agenten des MAS eindeutig zur Einhaltung der Geltungsansprüche der Kommunikation programmiert und im Diskursprotokoll die Merkmale der idealen Sprechsituation berücksichtigt werden. Zudem ist es leichter, viele Agenten in einem MAS miteinander interagieren zu lassen, als alle von einer Norm Betroffenen zu einem realen Diskurs zu versammeln, der immer räumlichen und zeitlichen Beschränkungen unterliegt. Darüber hinaus kann das Modell

die Diskursfähigkeit aller Agenten sicherstellen und muss nicht auf Annäherungen zurückgreifen. Das Modell bringt einige Probleme mit sich, die es zu klären gilt (vgl. Kapitel 5.3.). Nichtsdestotrotz liefert es einen Weg, soziologische Erkenntnisse für die Weiterentwicklung der künstlichen Intelligenz heranzuziehen und kommt somit den Forderungen der künstlichen Intelligenz nach der Beteiligung anderer Disziplinen nach (Irving & Askell, 2019; Russel et al., 2015).

Die vorliegende Arbeit stellt lediglich einen ersten Versuch dar, die Möglichkeiten der Verwendung sozialwissenschaftlicher Erkenntnisse in der künstlichen Intelligenz aufzuzeigen. Sie muss dahingehend eingegrenzt werden, dass sie ihre Forschungsfrage nach der Existenz normativer Universalien nur aus Sicht einer einzigen soziologischen Theorie, der Diskursethik nach Habermas, behandelt. Von einer finalen Beantwortung ihrer Forschungsfrage ist sie noch weit entfernt. Um diese zu finden, müssen weitere soziologische Position, die hier nur kurz angerissen wurden (vgl. Kapitel 3), detailliert auf ihre Beiträge zum Thema Moral untersucht werden. Darüber hinaus sind die Perspektiven anderer Disziplinen, die bereits zum Forschungsgebiet der künstlichen Intelligenz gezählt werden, wie etwa die Neurowissenschaft und die Psychologie (Russel & Norvig, 2010, S. 5ff.), für eine umfassende Analyse der Möglichkeit normativer Universalien und der Moral an sich höchst relevant. Erst eine Kombination der diversen Herangehensweisen und Erkenntnisse der unterschiedlichen Disziplinen lässt Antworten erhoffen, die für eine sichere und verantwortungsvolle Weiterentwicklung der künstlichen Intelligenz eingesetzt werden können. Eine weitere Einschränkung besteht darin, dass die Umsetzbarkeit des vorgeschlagenen Modells zur Simulation von Diskursen noch nicht geprüft worden ist.

Die bereits angeklungenen Hinweise auf zukünftige Fragestellungen und Probleme, die es im Anschluss an diese Arbeit zu erforschen gilt, sollen im Folgenden spezifiziert werden: Erstens muss das vorgeschlagenen Modells zur Simulation von Diskursen ausgestaltet und praktisch umgesetzt werden, sodass eine kritische Bewertung und Weiterentwicklung möglich wird. Zweitens sollten andere soziologische Ansätze zum Thema Moral ausführlich analysiert und auf mögliche Beiträge zur Entwicklung von an menschlichen Moralvorstellungen ausgerichteten KI-Systeme geprüft werden. Drittens ist es ratsam, sich mit den für den Bereich der Moral relevanten Beiträgen aus Disziplinen wie den Neurowissenschaften und der Psychologie zu beschäftigen. Nur so ergibt sich ein umfassendes Bild menschlicher Moral und möglicher normativer Universalien, das in der Konstruktion von KI-Systemen berücksichtigt werden kann. Um neu gewonnene

Erkenntnisse in der künstlichen Intelligenz überhaupt verwenden zu können, müssen viertens die Möglichkeiten der Implementierung von Moral weiter erforscht und erprobt werden. Insgesamt sind die Kooperation und der Austausch zwischen der künstlichen Intelligenz und den Sozialwissenschaften extrem sinnvoll. Das gilt besonders für die Bewältigung der ethischen Probleme der künstlichen Intelligenz. Allerdings können sozialwissenschaftliche Theorien auch in anderen Fragen der künstlichen Intelligenz hilfreich sein: Etwa können Erkenntnisse zur Dynamik und Koordination von sozialen Gruppen auf die Konstruktion von effektiv zusammenarbeitenden Agenten in MAS übertragen werden.

Abschließend muss festgehalten werden, dass die Thematik im Zentrum dieser Arbeit hohe gesellschaftliche Relevanz besitzt und ihr daher in zukünftigen wissenschaftlichen Auseinandersetzungen mehr Aufmerksamkeit geschenkt werden sollte. Wie eingangs schon erwähnt, kommt die rasante Weiterentwicklung der künstlichen Intelligenz mit Risiken, die nicht ignoriert werden dürfen. Diese Arbeit möchte dazu anstoßen, mehr Energie in die Konstruktion von sicheren, verantwortungsvollen und moralisch akzeptablen KI-Systemen zu investieren, also die Chancen der neuen Technologien zu maximieren und ihre Risiken frühzeitig zu identifizieren und zu vermeiden. Nur so kann eine „friendly AI“ (Yudkowsky, 2001) entwickelt werden, die die gesellschaftliche Entwicklung positiv vorantreibt und menschliches Leben verbessert.

**Literaturverzeichnis**


Amgoud, L. & Parsons, S. (2002). Agent Dialogues with Conflicting Preferences. In J.-J. C. Meyer & M. Tambe (Hrsg.), *Intelligent Agents VIII* (S. 190–205). Berlin, Heidelberg: Springer VS.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J. & Mané, D. (2016). *Concrete Problems in AI Safety*. arXiv:1606.06565 [cs.AI].

Arkoudas, K., Bringsjord, S. & Bello, P. (2005). Toward ethical robots via mechanized deontic logic. In M. Anderson, S. L. Anderson & C. Armen (Hrsg.), *Machine Ethics. Papers from the 2004 AAAI Fall Symposium. Technical Report FS-05-06* (S. 17–23). Menlo Park, California: AAAI Press.

Artificial General Intelligence Society (o. D). *Journal of General Artificial Intelligence: Overview*. Abgerufen am 14. Juli 2020, von https://content.sciendo.com/view/journals/jagi/jagi-overview.xml?language=en.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F. & Rahwan, I. (2018). The Moral Machine experiment. *Nature, 563*(7729), 59–64. DOI: 10.1038/s41586-018-0637-6.

Beetz, M. (2009). Was können Soziologen von Moral verstehen? *Berliner Journal für Soziologie, 19*(2), 248–267. DOI: 10.1007/s11609-009-0060-9.

Bergmann, J. (1998). Über den lokalen Charakter der Moral in der gegenwärtigen Gesellschaft. *Mitteilungen des Instituts für Sozialforschung an der Johann Wolfgang Goethe-Universität Frankfurt am Main, 9*, 70–91.

Bergmann, J. & Luckmann, T. (1999). *Kommunikative Konstruktion von Moral: Von der Moral zu den Moralen* (Bd. 2). Wiesbaden: VS Verlag für Sozialwissenschaften.

Bohman, J. & Rehg, W. (2017). Jürgen Habermas. In E. N. Zalta (Hrsg.), *The Stanford Encyclopedia of Philosophy*. Abgerufen am 14. Juli 2020, von https://plato.stanford.edu/archives/fall2017/entries/habermas/.

Bostrom, N. (2003). Ethical Issues in Advanced Artificial Intelligence. In I. Smit, W. Wallach & G. E. Lasker (Hrsg.), *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence* (S. 12–17). Windsor, Ontario: International Institute for Advanced Studies in Systems Research and Cybernetics.

Bostrom, N. (2016). *Superintelligence. Paths, dangers, strategies*. Oxford: Oxford University Press.

Bostrom, N. & Yudkowsky, E. (2014). The ethics of artificial intelligence. In K. Frankish & W. Ramsey (Hrsg.), *The Cambridge Handbook of Artificial Intelligence* (S. 316-334). Cambridge: Cambridge University Press. DOI: 10.1017/CBO9781139046855.020.

Brandt, F., Conitzer, V., Endriss, U., Lang, J., Procaccia, A. D. & Moulin, H. (2016). *Handbook of Computational Social Choice*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9781107446984.

Bringsjord, S. & Govindarajulu, N. (2020). Artificial Intelligence. In E. N. Zalta (Hrsg.), *The Stanford Encyclopedia of Philosophy*. Abgerufen am 14. Juli 2020, von https://plato.stanford.edu/archives/sum2020/entries/artificial-intelligence/.

Brock, O. (2018). Künstliche Intelligenz und Robotik. Begriffsdifferenzierung und Forschungsperspektiven. *Analysen und Argumente, 327*, Konrad-Adenauer-Stiftung e.V. Abgerufen am 14. Juli 2020, von https://www.kas.de/documents/252038/3346186/K%C3%BCnstliche+Intelligenz+und+Robotik.pdf/7d7cab64-4a52-8868-885b-c154aeb79147?version=1.1&t=1544430005315.

Buckner, C. & Garson, J. (2019). Connectionism. In E. N. Zalta (Hrsg.), *The Stanford Encyclopedia of Philosophy*. Abgerufen am 14. Juli 2020, von https://plato.stanford.edu/archives/fall2019/entries/connectionism/.

Bughin, J., Seong, J., Manyika, J., Chui, M. & Joshi, R. (2018). *Notes from the frontier: Modeling the impact of AI on the world economy*. McKinsey & Company. Abgerufen am 10. Juli 2020, von https://www.mckinsey.de/~/media/McKinsey/Locations/Europe%20and%20Middle%20East/Deutschland/News/Presse/2018/2018-09-05%20-%20MGI%20AI-Studie%20Dampfmaschine/MGI-Studie_Notes_from_the_Frontier_2018.ashx.

Burton, E., Goldsmith, J., Koenig, S., Kuipers, B. Mattei, N. & Walsh, T. (2017). Ethical Considerations in Artificial Intelligence Courses. *AI Magazine, 38*(2), 22-34. DOI: 10.1609/aimag.v38i2.2731.

Buxmann, P. & Schmidt, H. (2019). *Künstliche Intelligenz: Mit Algorithmen zum wirtschaftlichen Erfolg*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Conn, A. (2017). *How Do We Align Artificial Intelligence with Human Values?* Abgerufen am 14. Juli 2020, von https://futureoflife.org/2017/02/03/align-artificial-intelligence-with-human-values/.

Cycorp (2020). *CYC'S KNOWLEDGE BASE*. Abgerufen am 14. Juli 2020, von https://www.cyc.com/archives/service/cyc-knowledge-base.

Dafoe, A. (2018). *AI Governance: A Research Agenda*. Abgerufen am 14. Juli 2020, von https://www.fhi.ox.ac.uk/wp-content/uploads/GovAI-Agenda.pdf.

Dallinger, U. (2009). *Die Solidarität der modernen Gesellschaft. Der Diskurs um rationale oder normative Ordnung in Sozialtheorie und Soziologie des Wohlfahrtsstaats.* Wiesbaden: VS Verlag für Sozialwissenschaften, GWV Fachverlage GmbH.

DeepAI (2020). *Logic Programming*. Abgerufen am 14. Juli 2020, von https://deepai.org/machine-learning-glossary-and-terms/logic-programming.

Döbel, I., Leis, M., Vogelsang, M. M., Neustroev, D., Petzka, H. Riemer, A., Dr. Püping, S., Voss, A. & Wegele, M. (2018). *Maschinelles Lernen: Eine Analyse zu Kompetenzen, Forschung und Anwendung*. München: Frauenhofer-Gesellschaft. Abgerufen am 14. Juli 2020 unter https://www.bigdata.fraunhofer.de/content/dam/bigdata/de/documents/Publikationen/Fraunhofer_Studie_ML_201809.pdf.

Durkheim, É. & Adorno, T. W. (1996). *Soziologie und Philosophie* (3. Auflage).

Frankfurt am Main: Suhrkamp.

Düwell, M., Hübenthal, C. & Werner, M. H. (2011). *Handbuch Ethik.* Stuttgart: J.B. Metzler.

Eitner, J., Berkler, K., Köhler, H. Möhlmann, R. & Tumescheid, A.-M. (2017). *Trends für die künstliche Intelligenz.* München: Frauenhofer-Gesellschaft. Abgerufen am 14. Juli 2020, von https://www.fraunhofer.de/content/dam/zv/de/publikationen/broschueren/Trends-fuer-die-kuenstliche-Intelligenz.pdf.

Engisch, K. (1930). *Untersuchungen über Vorsatz und Fahrlässigkeit im Strafrecht*. Berlin: Liebmann.

Gabriel, I. (2020). *Artificial Intelligence, Values and Alignment.* arXiv:2001.09768 [cs.CY]. Abgerufen am 29. Juni 2020, von https://deepmind.com/research/publications/Artificial-Intelligence-Values-and-Alignment.

Gottschalk-Mazouz, N. (2000). *Diskursethik. Theorien, Entwicklungen, Perspektiven*. Berlin, Boston: De Gruyter.

Großmaß, R. & Anhorn, R. (2013). Perspektiven kritischer Sozialer Arbeit: *Kritik der Moralisierung: Theoretische Grundlagen, Diskurskritik, Klärungsvorschläge für die berufliche Praxis* (Bd. 15*)*. Wiesbaden: Springer VS.

Habermas, J. (1973). *Legitimationsprobleme im Spätkapitalismus*. Frankfurt am Main: Suhrkamp Verlag. Abgerufen am 15. Juli 2020, von https://soth.alexanderstreet.com/cgi-bin/SOTH/hub.py?type=source_details&browse=full&sourceid=S10023108.

Habermas, J. (1981). *Theorie des kommunikativen Handelns: Handlungsrationalität und gesellschaftliche Rationalisierung* (Bd. 1). Frankfurt am Main: Suhrkamp Verlag. Abgerufen am 15. Juli 2020, von https://soth.alexanderstreet.com/cgi-bin/SOTH/hub.py?browse=full&showfullrecord=on&sourceid=S10023119&type=source_details.

Habermas, J. (1983). *Moralbewußtsein und kommunikatives Handeln*. Frankfurt am Main: Suhrkamp Verlag.

Habermas, J. (1984). *Vorstudien und Ergänzungen zur Theorie des kommunikativen Handelns*. Frankfurt am Main: Suhrkamp Verlag.

Habermas, J. (1992). *Faktizität und Geltung. Beiträge zur Diskurstheorie des Rechts und des demokratischen Rechtsstaats*. Frankfurt am Main: Suhrkamp Verlag.

Habermas, J. (2009). *Philosophische Texte: Diskursethik* (Bd. 3). Frankfurt am Main: Suhrkamp Verlag.

Habermas, J., Demmerling, C. & Krüger, H. (2016). Kommunikative Vernunft. *Deutsche Zeitschrift für Philosophie, 64*(5), 806-827. DOI: https://doi.org/10.1515/dzph-2016-0061

Hadfield-Menell, D., Russell, S. J., Abbeel, P. & Dragan, A. (2016). Cooperative Inverse

Reinforcement Learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon & R. Garnett (Hrsg.), *Advances in Neural Information Processing Systems, 29*, 3909–3917.

Herbermann, M. (2002). Soziologische Argumentation und Ethik. *Sozialwissenschaften und Berufspraxis, 25*(3), 281–290.

Hochrangigen Expertengruppe für künstliche Intelligenz der Europäischen Kommission (HEG-KI). (2018). *Ethik-Leitlinien für eine vertrauenswürdige KI*. Abgerufen am 14. Juli 2020, von https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

Horster, D. (2006). Zur Einführung: *Jürgen Habermas zur* Einführung (3. überarbeitete Auflage, Bd. 185). Hamburg: Junius.

Huget, M.-P. (2003). Lecture Notes in Computer Science: *Communication in Multiagent Systems: Agent Communication Languages and Conversation Policies* (Bd. 2650). Berlin, Heidelberg: Springer VS.

IBM (2020). *Deep Blue*. Abgerufen am 14. Juli 2020, von https://www.ibm.com/ibm/history/ibm100/us/en/icons/deepblue/.

Irving, G. & Askell, A. (2019). AI Safety Needs Social Scientists. *Distill, 4*(2). DOI: 10.23915/distill.00014.

Kant, I. (2000). Werke: Akademie Textausgabe. *Kritik der reinen Vernunft (1. Aufl. 1781), Prolegomena; Grundlegung zur Metaphysik der Sitten; Metaphysische Anfangsgründe der Naturwissenschaften* (Bd. 4). Berlin: de Gruyter.

Kopp, J. & Steinbach, A. (2018). *Grundbegriffe der Soziologie*. Wiesbaden: Springer Fachmedien.

Kuhn, S. (2019). Prisoner's Dilemma. In E. N. Zalta (Hrsg.), *The Stanford Encyclopedia of Philosophy*. Abgerufen am 14. Juli 2020, von https://plato.stanford.edu/archives/win2019/entries/prisoner-dilemma/.

Legg, S. & Hutter, M. (2007). Universal Intelligence: A Definition of Machine Intelligence. *Minds & Machines, 17*(4), 391-444. arXiv:0712.3329 [cs.AI].

Leverhume Centre for the Future of Intelligence (2020). *The value alignment problem*. Abgerufen am 14. Juli 2020, von http://lcfi.ac.uk/projects/ai-futures-and-responsibility/value-alignment-problem/#:~:text=The%20Value%20Alignment%20Project%20seeks,domains%20in%20the%20real%20world., checked on 6/25/2020.

Liebig, S. (2007). Duisburger Beiträge zur soziologischen Forschung: *Theoretische Grundlagen und methodische Zugänge einer erklärenden Soziologie der Moral* (Bd. 6/2007). Duisburg: Institut für Soziologie der Universität Duisburg Essen.

Loh, J. (2018). Maschinenethik und Roboterethik. In O. Bendel (Hrsg.), *Handbuch Maschinenethik* (S. 1-19). Wiesbaden: Springer Fachmedien.

Luhmann, N. & Horster, D. (2008). *Die Moral der Gesellschaft*. Frankfurt am Main: Suhrkamp.

Luhmann, N. & Spaemann, R. (1990). *Paradigm lost, Über die ethische Reflexion der*

*Moral.* Frankfurt am Main: Suhrkamp.

Luhmann, N. & Pfürtner, S. H. (1978). *Theorietechnik und Moral*. Frankfurt am Main, Berlin: Suhrkamp.

Mannino, A., Althaus, D., Erhardt, J., Gloor, L., Hutter, A. & Metzinger, T. (2015). Künstliche Intelligenz: Chancen und Risiken. *Diskussionspapiere der Stiftung für Effektiven Altruismus, 2,* 1-17.

McCarthy, J., Minsky, M. L., Rochester, N. & Shannon, C. E. (2006). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Magazine*, *27*(4), 12-14.

Mineau, G. W. (2003). Representing and enforcing interaction protocols in multi-agent systems: an approach based on conceptual graphs. *Proceedings IEEE/WIC International Conference on Intelligent Agent Technology, IAT 2003,* 261-267. DOI: 10.1109/IAT.2003.1241077.

Monett, D., Lewis, C. W. P., Thórisson, K. R., Bach, J., Baldassarre, G., Granato, G., Berkeley, I. S. N., Chollet, F., Crosby, M., Shevlin, H., Fox, J., Laird, J. E., Legg, S., Lindes, P., Mikolov, T., Rapaport, W. J., Rojas, R., Rosa, M., Stone, P., . . . Winfield, A. (2020). Special Issue "On Defining Artificial Intelligence"—Commentaries and Author's Response. *Journal of Artificial General Intelligence, 11*(2), 1–100.

Müller, V. C. (2020). Ethics of Artificial Intelligence and Robotics. In E. N. Zalta (Hrsg.), *The Stanford Encyclopedia of Philosophy*. Abgerufen am 14. Juli 2020, von https://plato.stanford.edu/archives/fall2020/entries/ethics-ai/.

Noothigattu, R., Gaikwad, S. 'N.' S., Awad, E., Dsouza, S., Rahwan, I., Ravikumar, P. & Procaccia, A. D. (2018). *A Voting-Based System for Ethical Decision Making*. arXiv:1709.06692v2 [cs.AI].

Poslad, Stefan (2007): Specifying protocols for multi-agent systems interaction. *ACM Transactions on Autonomous and Adaptive Systems, 2*(4), Artikel 15. DOI: 10.1145/1293731.1293735.

Rahwan, I., Parsons, S. & Reed, C. (2008). Argumentation in Multi-Agent Systems: Forth International Workshop, ArgMAS 2007, Honolulu, HI, USA, May 2007. *Revised Selected and Invited Papers. Lecture Notes in Artificial Intelligence* (Bd. 4946). Berlin: Springer VS.

Reinhold, Gerd (2000). *Soziologie-Lexikon*. München: Oldenbourg Wissenschaftsverlag.

Russell, Stuart J. (2019). *Human compatible. AI and the problem of control.* London: Penguin Books.

Russell, S. J., Dewey, D. & Tegmark, M. (2015). Research Priorities for Robust and Beneficial Artificial Intelligence. *AI Magazine, 36*(4), 105-114.

Russell, S. J. & Norvig, P. (2010). *Artificial intelligence. A modern approach* (3. Auflage). Boston, Massachusetts: Pearson.

Schäfers, B. (1992). Lektion II: Die Grundlagen des Handelns: Sinn, Normen, Werte. In

H. Korte (Hrsg.), *Einführung in die Hauptbegriffe der Soziologie.* Wiesbaden: Springer Fachmedien.

Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(3), 417-424. Doi:10.1017/S0140525X00005756.

Seiler, C. (2014). *Die Diskursethik im Spannungsfeld von Systemtheorie und Differenzphilosophie*. Wiesbaden: Springer Fachmedien.

Simanowski, R. (2017). *Der Todesalgorithmus*. Zeit Online. Abgerufen am 14. Juli 2020, von https://www.zeit.de/kultur/2017-09/kuenstliche-intelligenz-algorithmus-spam-autonomes-fahren/komplettansicht.

Singh, M. P. (1991). Towards a Formal Theory of Communication for Multiagent Systems. *Proceedings of the 12th International Joint Conference on Artificial Intelligence (IJCAI), 1991*, 69–74. San Francisco, California: Morgan Kaufmann Publishers Inc.

Sinnott-Armstrong, W. (2019). Consequentialism. In E. N. Zalta (Hrsg.), *The Stanford Encyclopedia of Philosophy*. Abgerufen am 14. Juli 2020, von https://plato.stanford.edu/archives/sum2019/entries/consequentialism/.

Soares, N. & Fallenstein, B. (2014). Aligning Superintelligence with Human Interests: A Technical Research Agenda. *Technical report 2014–8*. Berkeley, CA: Machine Intelligence Research Institute.

Soares, N. (2015). Aligning Superintelligence with Human Interests: An Annotated Bibliography. *Intelligence, 17*(4), 391–444.

Sotala, K. (2016). Defining Human Values for Value Learners. *The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence: Technical Reports WS-16-01 – WS-16-15*, 113-123. Palo Alto, California: The AAAI Press.

Thorisson, K. R. (2020). Discretionarily Constrained Adaptation Under Insufficient Knowledge & Resources. *Journal of Artificial General Intelligence, 11*(2), 7–12.

Torreño, A., Onaindia, E. & Sapena, O. (2010). Reaching a Common Agreement Discourse Universe on Multi-Agent Planning. In: E. Corchado, M. Graña Romay, A. Manhaes Savio (Hrsg.), *Hybrid Artificial Intelligence Systems. HAIS 2010. Lecture Notes in Computer Science, 6077*, 185-192. Berlin, Heidelberg: Springer VS.

Von Rimscha, M. (2014). *Algorithmen kompakt und verständlich*. Wiesbaden: Springer Fachmedien.

Wallach, W. & Allen, C. (2009). *Moral machines. Teaching robots right from wrong*. Oxford: Oxford University Press.

Wang, P. (2019). On Defining Artificial Intelligence. *Journal of Artificial General Intelligence, 10*(2), 1–37. DOI: 10.2478/jagi-2019-0002.

Weiss, G. (1999). *Multiagent systems. A modern approach to distributed artificial intelligence*. Cambridge, Massachusetts, London: MIT Press.

Wooldridge, M. (1998). Verifiable semantics for agent communication languages.

*Proceedings International Conference on Multi Agent Systems (Cat. No.98EX160), Paris, France*, 349-356. DOI: 10.1109/ICMAS.1998.699219.

Yudkowsky E. (2011). Complex Value Systems in Friendly AI. In: J. Schmidhuber, K. R. Thórisson, M. Looks (Hrsg.), *Artificial General Intelligence. AGI 2011. Lecture Notes in Computer Science, LNAI 6830*, 388–393. Berlin, Heidelberg: Springer VS.

Yudkowsky, E. (2001). *Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures.* San Francisco: The Singularity Institute.

# Eidesstattliche Erklärung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und ausschließlich unter Zuhilfenahme der ausgewiesenen Hilfsmittel angefertigt habe.

Münster, den 29.07.2020 _____

Lara Lawniczak