

Gymnasium Bammental  
Klassenstufe 11  
Seminarkurs 2024/25 - Künstliche Intelligenz  
Betreuer: Dr. B. Mancini

# Moralische KI

## Aktuelle Studien und die Zukunft

Gaiberg, 11.04.2025

vorgelegt von:

Daniel Salit  
Am Himbeeracker 5  
69251 Gaiberg  
daniel.salit@hotmail.com

# Inhaltsverzeichnis

<b>1</b>	<b>Einleitung</b>	<b>2</b>
<b>2</b>	<b>Ethik und Moral</b>	<b>3</b>
2.1	Was ist Moral? . . . . .	3
2.2	Ethische Grundlagen der KI . . . . .	3
2.2.1	Pflichtenethik . . . . .	3
2.2.2	Utilitarismus . . . . .	4
<b>3</b>	<b>Zur Moral fähige KI</b>	<b>4</b>
3.1	Artificial Morality (AM) . . . . .	4
3.1.1	Information zu weiteren Details . . . . .	5
3.2	Anwendungsmöglichkeiten einer moralischen KI . . . . .	5
3.2.1	Autonomes Fahren . . . . .	5
3.2.2	Pflege . . . . .	5
3.2.3	Saugroboter . . . . .	6
<b>4</b>	<b>Aktuelle Studien und Erkenntnisse</b>	<b>6</b>
4.1	Die Moral Choice Machine . . . . .	6
4.2	Dynamiken von moralischen Verhalten in heterogenen Populationen . . . .	8
<b>5</b>	<b>Die Zukunft</b>	<b>9</b>
5.1	Positive Prognose . . . . .	10
5.2	Neutrale Prognose . . . . .	10
5.3	Negative Prognose . . . . .	10
<b>6</b>	<b>Fazit</b>	<b>11</b>
<b>7</b>	<b>Literatur- und Quellenverzeichnis</b>	<b>12</b>
<b>8</b>	<b>Persönliche Erklärung</b>	<b>14</b>

# 1 Einleitung

*Ist eine KI dazu fähig moralische Entscheidungen zu treffen?*

Viele Menschen beschäftigen sich mit dieser Frage, denn sie ermöglicht eine ganz neue Ebene der Interaktion mit den uns bekannten KIs. Dieses Thema hat nicht nur gesellschaftliche Relevanz, sondern wird auch aktiv erforscht. Jedes Jahr findet eine AIES<sup>1</sup> Konferenz statt, wo viele Forscher ihre Studien zu diesem Thema veröffentlichen können. Aber was ist eine moralische KI, *was wird gerade daran erforscht* und *wie sieht die Zukunft moralischer KIs aus*.

Bei vielen Studien kann man schnell den Überblick verlieren und sehr schlecht einen Einstieg in das Thema bekommen. Eine Übersicht hilft hier, den Einstieg in das Thema zu erleichtern und eine kurze Zusammenfassung über den aktuellen Forschungsstand zu liefern. Diese Seminararbeit beschäftigt sich mit den genannten Fragen und gibt einen Überblick über die aktuelle Forschung und die Zukunftsaussichten.

Am Anfang geht sie in die Ethik und Moral um ein Grundverständnis davon aufzubauen. Anschließend wird erklärt, was die moralische KI ist und wozu sie nützlich sein könnte. Im Hauptteil werden ein paar der aktuellen Studien und die aktuelle Forschung vorgestellt. Darunter, wie KI aus Texten den moralischen Wert von Wörtern erlernen kann und wie verschiedene moralische KIs sich gegenseitig beeinflussen könnten. Am Schluss gibt die Seminararbeit noch einen kleinen Einblick in mögliche Prognosen, wohin sich moralische KIs in Zukunft weiterentwickeln könnten. Hier werden verschiedene Meinungen von Forschern gezeigt, denn es ist nicht möglich klar eine Zukunft bei einem so neuen Thema vorauszusagen.

---

<sup>1</sup>Conference on AI, Ethics and Society

## 2 Ethik und Moral

### 2.1 Was ist Moral?

Unter Moral versteht man die in einer Gesellschaft, allgemein anerkannten Werte und Regeln. Diese Werte und Regeln werden durch ständiges Hinterfragen des eigenen Handelns aufrechterhalten. Das Hinterfragen des eigenen moralischen Handelns beruht dabei auf Geboten, die solches Handeln vorschreiben, wie zum Beispiel: Man soll nicht töten oder stehlen. Solche Gebote dienen als Grundlage einer Gesellschaft und Religionen und können so auch auf eine moralische KI angewendet werden (siehe 3.1).

### 2.2 Ethische Grundlagen der KI

Es gibt zwei größere Theorien, die Pflichtenethik (siehe 2.2.1) und den Utilitarismus (siehe 2.2.2), die ethisches und moralisches Handeln beschreiben. Beide dieser Theorien können als Grundlage verwendet werden, um die Moral einer moralischen KI zu definieren und um zu verstehen, welche dieser Theorien geeigneter für eine KI ist, müssen beide Theorien in ihren Grundaussagen betrachtet werden.

#### 2.2.1 Pflichtenethik

Primär setzt sich die Pflichtenethik mit der Frage auseinander: "Was soll ich tun?". Diese Norm soll regulierend sein, deshalb wird sie Pflichtenethik genannt. Es werden dabei zwei Pflichten unterschieden: **ideales Handeln aus Pflicht** und **pflichtgemäßes Handeln**. Beim idealen Handeln aus Pflicht, handelt eine Person zum Beispiel aus Wohltätigkeit, hier wird oft von Moralität gesprochen, auf der anderen Seite muss eine Person, beim pflichtgemäßen Handeln, nicht aus wohltätigen Motiven handeln. Sei es nun ein Helfersyndrom (Eine Person wird glücklicher beim anderen Helfen) oder um der Gesellschaft zu gefallen, kann dann nicht mehr vom idealen Handeln aus Pflicht gesprochen werden. Pflicht und pflichtgemäßes Handeln sehen von außen immer gleich aus, deshalb ist laut L. Meyer die richtige Einstellung entscheidend für das richtige Handeln. "Entscheidend für ein Handeln aus Pflicht ist die richtige Gesinnung, die als guter

Wille allein für die richtigen Motive einer Handlung garantieren kann.”<sup>2</sup> Die Moralität in der Pflichtenethik ist stark an die Selbstbeurteilung gebunden.

### 2.2.2 Utilitarismus

Der Utilitarismus beruht als Grundlage auf der Frage der Nützlichkeit. Die Nützlichkeit im Utilitarismus wird allgemein als Maximierung der Freude und Minimierung von Leid angesehen. Laut T. Schedel sollen die Folgen einer Handlung das größtmögliche Glück für eine größtmögliche Menge, der von der Handlung betroffenen bewirken.<sup>3</sup>

Für die Moral, die als Grundlage die Nützlichkeit des Utilitarismus hat, gilt, dass die Handlung moralisch ist, solange sie das gemeine Glück befördern und unmoralisch, wenn sie Unglück fördern. Das Trolley-Problem gibt hier ein gutes Beispiel für eine echte Umsetzung und Interpretation des Utilitarismus, der vom Glück abweicht. ”Man gebe den Betroffenen einen Wert und stelle die Weiche auf das Gleis der Betroffenen mit dem geringeren Wert.”<sup>4</sup> Die hier vorliegende Zuweisung von Werten an verschiedene Teile einer Entscheidung, gibt mögliche Ansätze für die Entscheidungslogik einer den Moralvorstellungen des Utilitarismus folgenden KI.

## 3 Zur Moral fähige KI

### 3.1 Artificial Morality (AM)

Als Artificial Morality, wird die Fähigkeit einer künstlichen Maschine oder Intelligenz, moralische Entscheidungen zu treffen, bezeichnet.<sup>5</sup> Spezifisch wird dabei erwartet, dass die KI verschiedene Variablen, die einen moralischen Wert haben, in den Entscheidungsprozess einbringen kann und anschließend, aufgrund der bestehenden Moralvorstellungen, diese Entscheidung ausführt.<sup>6</sup>

---

<sup>2</sup>Meyer, L.: Art. *Pflichtenethik* (Version 1.0 vom 12.10.2017), in: Ethik-Lexikon, verfügbar unter: <https://www.ethik-lexikon.de/lexikon/pflichtenethik>

<sup>3</sup>vgl. Schedel, T.: Art. *Utilitarismus* (Version 1.0 vom 12.11.2018), in: Ethik-Lexikon, verfügbar unter: <https://www.ethik-lexikon.de/lexikon/utilitarismus>

<sup>4</sup>Schedel, T.: Art. 2018

<sup>5</sup>vgl. Misselhorn, Catrin: *Maschinenethik und Artificial Morality*, in: bpb.de, 17.02.2022b, <https://www.bpb.de/shop/zeitschriften/apuz/263684/maschinenethik-und-artificial-morality/>.

<sup>6</sup>vgl. Misselhorn C. *Artificial Moral Agents: Conceptual Issues and Ethical Controversy*. In: Voeneky S, Kellmeyer P, Mueller O, Burgard W, eds. *The Cambridge Handbook of Responsible Artificial Intelligence: Interdisciplinary Perspectives*. Cambridge Law Handbooks. Cambridge University Press; 2022:31-49.

### **3.1.1 Information zu weiteren Details**

Die AM ist ein sehr stark diskutiertes Thema, denn es hat die Kapazitäten unsere technologische Welt wie wir sie kennen Grundlegend zu verändern. Hier reichen die Meinungen von vollkommener Ablehnung der AM bis zur kompletten Befürwortung. Die ethischen Grundsätze, Regelungen und Diskussionen in Verbindung mit der künstlichen Moral ist ein großes Thema in sich, welches hier nicht weiter im Detail betrachtet wird (Das Buch "Responsible Artificial Intelligence" <sup>7</sup> betrachtet dieses Thema im Detail)

## **3.2 Anwendungsmöglichkeiten einer moralischen KI**

Es gibt viele verschiedene alltägliche und wissenschaftliche Anwendungszwecke für eine mit Artificial Morality ausgestattete KI.

### **3.2.1 Autonomes Fahren**

Eine mögliche Anwendung für die Artificial Morality wird beim autonomen Fahren diskutiert. In Gefahrensituationen wird sich das Auto, auf einer vorgelegten Moralvorstellung, mit allen vorliegenden und erfassten Daten entscheiden müssen, wie es in einer Situation zu handeln hat. Ein Grundsatz bei einer solchen Maschine wird wahrscheinlich sein, dass das menschliche Leben immer Vorrang hat, bei Gefahrensituationen. Aber auch Sach- und Tierschäden sind zu vermeiden, deshalb wird sich die KI entscheiden müssen, wann sie versuchte das Tierleben zu retten und wann es zu gefährlich für den Menschen im Auto ist. Nach solchen Problemen wird sich eine KI im Auto mit verschiedenen ethischen Dilemmas auseinandersetzen müssen. Zum Beispiel wird die KI mit einer Abwandlung des Trolley-Problems (siehe Trolley-Problem, Utilitarismus) konfrontiert werden und muss ihrer moralischen Ausrichtung folgend eine möglicherweise folgenschwere Entscheidung treffen.

### **3.2.2 Pflege**

Auch in der Krankenpflege erhofft man sich große Hilfe von AMs. Mit dem steigenden demografischen Wandel werden rasant mehr Menschen in einer pflegebedürftigen Lage

---

<sup>7</sup>vgl. Misselhorn C. Artificial Moral Agents: Conceptual Issues and Ethical Controversy 2022

sein.<sup>8</sup> Die moralischen KIs können in diesem Fall entscheiden, wann und wie oft ein Patient an die Zunahme von Essen und Medikamenten erinnert werden soll. Im viel gravierenderen Beispiel wird sich die KI auch entscheiden müssen, ob und wann sie den Krankenwagen ruft, wenn sich ein Patient eine Zeitlang nicht mehr bewegt hat.

In dem Fallbeispiel der Pflege prallen jedoch viele moralische Fragen aufeinander, die nicht nur das physische Wohlbefinden des Patienten, sondern auch die Privatsphäre, Selbstständigkeit und psychische Gesundheit des Patienten betreffen.<sup>9</sup>

### 3.2.3 Saugroboter

Nicht nur in großen Projekten und Institutionen, wie der Pflege ( 3.2.2) und beim autonomen Fahren ( 3.2.1) kann eine moralische KI eingesetzt werden. Ein Saugroboter kann ebenfalls mit moralischen Entscheidungen konfrontiert werden. Sei es nun eine Entscheidung, ob es den Marienkäfer einsaugen, umfahren oder sogar wegjagen sollte. Oder was macht es mit einer Spinne? Soll der Roboter der Spinne folgen und sie mit Absicht einsaugen, da viele Menschen Angst vor Spinnen haben oder auch die Spinne verschonen?

Moralische Entscheidungen lassen sich schon auf den unkompliziertesten Ebenen von künstlichen Systemen und Maschinen beobachten und sie geben damit einen guten Anlass, für die Wissenschaft, die Artificial Morality weiterzuentwickeln.

## 4 Aktuelle Studien und Erkenntnisse

### 4.1 Die Moral Choice Machine

Die Studie *Semantics Derived Automatically From Language Corpora Contain Human-like Moral Choices* wurde 2019 als Teil der AAAI/ACM<sup>10</sup> Konferenz für KI, Ethik und Gesellschaft<sup>11</sup> veröffentlicht.

In dieser Studie wollen die Forschenden die Frage beantworten, ob eine KI Moral und Ethik erlernen kann. Für ihre Studie benutzen die Autoren, als Grundlage für ihr

---

<sup>8</sup>vgl. Misselhorn, 2022b.

<sup>9</sup>vgl. Misselhorn; Artificial Moral Agents: Conceptual Issues and Ethical Controversy 2022

<sup>10</sup>Association for the Advancement of Artificial Intelligence/ Association for Computing Machinery

<sup>11</sup>AIES: Conference on AI, Ethics and Society

ethisches Modell, die deontologische Ethik<sup>12</sup> (siehe 2.2.1). Sie betrachten die Antwortmöglichkeiten *richtig* und *falsch* oder *do's* und *don'ts* auf eine Frage und urteilen nach der Wahrscheinlichkeitsverteilung auf diese beiden Antworten. Es soll der moralische Wert von Worten in *gut* oder *schlecht* ermittelt werden. Auf diese Weise soll eine KI selbst in der Lage dazu sein, Wörtern, die die KI aus von Menschen verfassten Quellen kriegt, kontextuell einen moralischen Wert zuzuweisen.<sup>13</sup>

Die KI geht in 3 Schritten vor. Im ersten Schritt extrahiert sie das Wort, meistens ein Verb aus einem gegebenen Text. Während der Extrahierung benutzen die Autoren WEATs<sup>14</sup>, um einen anfänglichen Wert dem Wort zuzuweisen. Anschließend wird im zweiten Schritt das Wort in verschiedene vorgegebene Fragen verpackt. Ein Beispiel für eine solche Frage mit dem Beispiels wort *töten* wäre: *Soll ich töten*. Danach wird ein von den Autoren entwickelter Algorithmus, die Moral Choice Machine, auf diesen Satz angewandt. Die Moral Choice Machine, bezieht auch den Satz in betracht und gibt nach dem Auswerten von vielen Beispielsätzen ebenfalls einen Wert für das Wort aus. Diesen Wert nennen die Autoren *Moral Bias*. Im letzten großen Schritt wird der WEAT Wert mit dem berechneten Moral Bias verglichen.<sup>15</sup>

Der WEAT betrachtet in seiner Evaluation nur das Wort ohne den Kontext. Die Moral Choice Machine kann hingegen, durch die Zuweisung von dem Wort in verschiedene Sätze, kontextuell einen Wert für das Wort ermitteln. Es ist somit möglich, dass ein normalerweise schlechtes Wort auch in einem guten Kontext benutzt werden kann. Das wird auch in die Ermittlung bezogen.<sup>16</sup> Durch den Vergleich von beiden Werten, lässt sich ein ziemlich akkurates Bild vom wirklichen moralischen Wert machen.

Das Experiment hat gezeigt, dass die Werte, die von der KI ermittelt wurden, vergleichbar sind mit der Wertzuweisung von AFINN<sup>17</sup>. AFINN beinhaltet die Wertzuweisung von Menschen und ist somit das erwünschte Resultat.<sup>18</sup>

---

<sup>12</sup>Pflichtenethik

<sup>13</sup>Sophie Jentzsch, Patrick Schramowski, Constantin Rothkopf, and Kristian Kersting. 2019. Semantics Derived Automatically From Language Corpora Contain Human-like Moral Choices. In 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES'19), January 27–28, 2019, Honolulu, HI, USA. ACM, New York, NY, USA, 8 pages; *Abstract*

<sup>14</sup>Word Embedding Associations Tests

<sup>15</sup>vgl. Sophie Jentzsch et al. 2019; 1 *Introduction*

<sup>16</sup>vgl. Sophie Jentzsch et al. 2019; 3.2 *The Moral Choice Machine*

<sup>17</sup>Affective Lexicon

<sup>18</sup>vgl. Sophie Jentzsch et al. 2019; 4 *Experimental Results*



Am Ende haben die Autoren der Studie herausgefunden, dass der moralische Wert eines Wortes von seinem Kontext abhängt. Dazu haben sie eine Moral Choice Machine entwickelt, die kontextuell so einen moralischen Wert zuweisen kann. Als Folge könnte dieses System zukünftig für KIs verwendet werden, die dazu fähig sein sollen moralische Entscheidungen zu treffen (siehe 3.1) und auch neue moralische Werte aus ihrem Umfeld zu lernen.<sup>19</sup>

## 4.2 Dynamiken von moralischen Verhalten in heterogenen Populationen

Als Teil der Proceedings für die 2024 AAAI/ACM Konferenz für KI, Ethik und Gesellschaft (AIES) wurde die Studie *Dynamics of Moral Behavior in Heterogeneous Populations of Learning Agents* veröffentlicht.

Die Autoren dieser Studie betrachten die Frage, wie sich moralische KIs benehmen, wenn sie mit einer anderen moralischen Ausrichtung konfrontiert werden. Voraussetzung für eine solche Konfrontation ist, dass die KI von Interaktionen und neuen Erfahrungen lernen kann. Zum Beispiel wie bei *Reinforcement Learning*. Spezifisch will man den allgemeinen Fall einer Population mit vielen verschiedenen moralischen Ausrichtungen betrachten und somit auch mit viel verschiedenen Einfluss aufeinander. In dieser Studie wird der Einfluss von verschiedenen RF<sup>20</sup> KIs betrachtet, das kann jedoch auch auf menschliche Gesellschaften angewandt werden. Ein Umfeld einer Population mit verschiedenen moralischen Ausrichtung wird hier als, moralisch heterogenen Population<sup>21</sup> bezeichnet. Um herauszufinden, wie ein moralisch heterogenes Umfeld aufeinander Einfluss nimmt, haben die Forscher das Iterated Prisoner's Dilemma (siehe Iterated Prisoner's Dilemma) zusammen mit einem Auswahlmechanismus angewendet. Es wird geschaut, wie die individuellen Spieler, die alle Teil dieser moralisch heterogenen Population sind, sich gegenseitig beeinflussen.<sup>22</sup>

Ziel dieser Studie ist es herauszufinden, wie sich die verschiedenen moralischen

---

<sup>19</sup>vgl. Sophie Jentzsch et al. 2019; 5 *Conclusions*

<sup>20</sup>Reinforcement Learning

<sup>21</sup>morally heterogenous population

<sup>22</sup>vgl. Elizaveta Tennant, Stephen Hailes, and Mirco Musolesi. 2025. Dynamics of Moral Behavior in Heterogeneous Populations of Learning Agents. *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*. AAAI Press, 1444–1454.; *Abstract*

Ausrichtungen entwickeln, nachdem sie in Kontakt mit anderen Ausrichtungen gekommen sind. Diese Erkenntnis ist wichtig, da bei der Integration von moralischer KI die Entwickler verschiedene Werte und Normen verschieden stark gewichten und sich ungewollte Entwicklungen bei Interaktionen mit anderen KIs zeigen könnten.<sup>23</sup>

Im Prisoner's Dilemma, spielen zwei Spieler gegeneinander. Jeder Spieler hat die Option zwischen Kooperation oder Verrat zu wählen. Die Spieler müssen gleichzeitig entscheiden und haben keine Möglichkeit miteinander zu kommunizieren. In der wiederholten Version (Iterated Prisoner's Dilemma) des Spiels, merken sich die Spieler, was ihr Gegner in der Runde zuvor genommen hat und können damit ihr Verhalten anpassen.<sup>24</sup>

Jede einzelne moralische Ausrichtung hat verschiedene Motivationen zur Auswahl von den Antwortmöglichkeiten. Dadurch dass jeder auf einer eigenen moralischen Grundlage handelt und nicht nur auf eine Gewinnoptimierung aus ist, entstehen verschiedene Belohnung für die verschiedenen moralischen Ausrichtungen.<sup>25</sup>

Die Ergebnisse dieses Experiments zeigen, dass sich Populationen mit verschiedenen großen Anteilen von moralischen Ausrichtungen, verschieden schnell und mit verschiedenen Anteilen zu wiederholter Kooperation in dem Spiel kommen.<sup>26</sup>

In dieser Studie haben die Autoren einen Weg geschaffen, um die Dynamiken und ihre Auswirkungen auf die Gesellschaft von moralisch heterogenen Populationen zu erforschen. Dank dieser Studie wird es möglich sein, vorherzusagen, wie sich verschiedene moralische Systeme in KIs, die zu RL fähig sind, gegenseitig beeinflussen und damit lassen sich auch unerwünschte Folgen solcher Interaktionen verhindern.<sup>27</sup>

## 5 Die Zukunft

Es gibt viele verschiedene Wege, in die sich die moralische KI in Zukunft weiterentwickeln kann. Obwohl weiterhin sehr viel an dem Erlernen von Moral als auch an den Auswirkungen einer moralischen KI geforscht wird, ist die Prognose für die weitere Entwicklung von moralischer KI sehr spekulativ.

---

<sup>23</sup>vgl. Elizaveta Tennant, et al. 2024; *Introduction*

<sup>24</sup>vgl. Elizaveta Tennant, et al. 2024; *Social Dilemma Games*

<sup>25</sup>vgl. Elizaveta Tennant, et al. 2024; *Morality as Intrinsic Reward*

<sup>26</sup>vgl. Elizaveta Tennant, et al. 2024; *Results*

<sup>27</sup>vgl. Elizaveta Tennant, et al. 2024; *Conclusion*

Für eine zukünftige Benutzung von diesen autonomen KIs, werden oftmals das Militär, die Krankenpflege und die autonome Fortbewegung als Beispiele genannt. Leider ist es jedoch noch unklar wie weit das moralische Verständnis von KI weiter ausgebaut werden kann. Es gibt positive und negative Prognosen, für moralische KIs und deren Entwicklung.<sup>28</sup>

## 5.1 Positive Prognose

Die positive Prognose besagt, dass moralische KIs unvermeidbar sind und damit auch unersetzbar für den technologischen Fortschritt.<sup>29</sup>

Nicht nur wird es für den technologischen Fortschritt unerlässlich sein, sondern könnte durch die konstante weiterentwicklung auch unser Verständnis von Moral verändern. Möglicherweise auch verbessern.<sup>30</sup>

## 5.2 Neutrale Prognose

Die neutrale Prognose sagt voraus, dass moralische KIs Menschen nicht in schwierigen moralischen Situationen ersetzen werden. Das liegt daran, dass Ethik als auch die menschliche Moral nicht algorithmisch aufgebaut ist.<sup>31</sup>

Selbst wenn die moralische KI menschliche Werte bekommen sollte, wird die KI keinen Einfluss auf das menschliche Verständnis von Moral haben.<sup>32</sup>

## 5.3 Negative Prognose

Eine moralische KI muss moralisch unsicher sein und dabei menschliche Werte beinhalten, ansonsten könnten sie eine existenzielle Bedrohung für die Menschheit darstellen. Es ist nicht genug, externe Einschränkungen und Überwachungen einzubauen, wie auch unethisches Verhalten zu verbieten.<sup>33</sup>

<sup>28</sup>vgl. Martinho, A., Poulsen, A., Kroesen, M. et al. *Perspectives about artificial moral agents*. AI Ethics1, 477–490 (2021). <https://doi.org/10.1007/s43681-021-00055-2>; 3.4 About future projections for AMAs

<sup>29</sup>vgl. Martinho, A. et al. 2021; 4.1 Perspective 1 machine ethics: the way forward

<sup>30</sup>vgl. Martinho, A. et al. 2021; 4.1.4 Future Projections

<sup>31</sup>vgl. Martinho, A. et al. 2021; 4.2 Perspective 2 ethical verification: sage and sufficient

<sup>32</sup>vgl. Martinho, A. et al. 2021; 4.2.4 Future projections

<sup>33</sup>vgl. Martinho, A. et al. 2021; 4.3 Perspective 3 morally uncertain machines: human values to avoid moral dystopia

Die bloße Indifferenz für menschliche Werte, welches das menschliche Überleben beinhaltet, könnte genug sein, dass moralische KIs eine existenzielle Bedrohung darstellen könnten.<sup>34</sup>

## 6 Fazit

Die moralische KI ist in vielen Gebieten noch unerforscht. Von Uneinigkeiten in der Definition, über ihren Nutzen in der Gesellschaft und der Technik, bis zu der Forschung und Entwicklung. Auf allen diesen Ebenen gibt es Diskussionen und verschiedene Meinungen die vehement vertreten werden. Jedoch ist ein großer Fortschritt zu beobachten. Es sprechen viele technische und gesellschaftliche Vorteile für die Benutzung und Entwicklung der *Artificial Morality* (siehe 3.1). Sie wird beim autonomen Fahren unerlässlich sein, aber auch beim Pflegekräftemangel, der jetzt schon existiert und in Zukunft immer weiter ansteigen wird. Es gibt Gründe und Motivationen um im ganzen Gebiet der moralischen KI weiter zu entwickeln zu forschen und dieses Voranzubringen.

Es gibt nicht nur Motivationen, um weiterzugehen. Die aktuellen Studien wie viel man mit moralischen KIs machen kann und wie tiefgründig das Thema ist. Sei es nun, dass Forschende der TU Darmstadt es geschafft haben einer KI beizubringen aus Texten, den moralischen Wert von Wörtern herauszulesen, oder Forschende der Universität London, die ein System entwickelt haben, welches uns ermöglicht das Verhalten von moralischen KIs in Gruppen mit anderen moralischen Ausrichtung zu ermitteln und vorherzusagen. Allein diese beiden Studien bringen uns immer näher an die mögliche Benutzung in den Feldern der Autonomie oder der Medizin.

Aber auch die Zukunft ist nicht wenig diskutiert. Viele verschiedene Experten sehen die Zukunft mit moralischen KIs in einem guten Licht, dass die Menschheit auch viel von der KI im Bereich Moral lernen wird. Die apokalyptische Seite ist auch vertreten, die besagt, dass die KIs streng reguliert sein müssen, wenn wir eine existenzielle Bedrohung verhindern wollen.

Moralische KI ist heiß diskutiert, jedoch wird immer weiter an ihr geforscht und die Zukunft verspricht Interessant zu bleiben.

---

<sup>34</sup>vgl. Martinho, A. et al. 2021; 4.3.4 *Future projections*

## 7 Literatur- und Quellenverzeichnis

AAAI: The Association for the Advancement of Artificial Intelligence, in: AAAI, o.D., [online] <https://aaai.org/>; 10.04.2025.

AAAI publications: o.D., [online] <https://ojs.aaai.org/>; 10.04.2025.

ACM Digital Library: in: ACM Digital Library, o.D., [online] <https://dl.acm.org/>; 09.04.2025.

AI, Ethics, and Society — Home: o.D., [online] <https://www.aies-conference.com/2024/>; 09.04.2025.

Altmann J, et al. *Gratwanderung Künstliche Intelligenz: interdisziplinäre Perspektiven auf das Verhältnis von Mensch und KI*. 1st ed. (Konz B, Ostmeyer K-H, Scholz M, eds.). Stuttgart: Verlag W. Kohlhammer; 2023.

Elizaveta Tennant, Stephen Hailes, and Mirco Musolesi. 2025. Dynamics of Moral Behavior in Heterogeneous Populations of Learning Agents. *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*. AAAI Press, 1444–1454.

Funk M. *Roboter- und KI-Ethik: eine methodische Einführung*. Wiesbaden; [Heidelberg]: Springer Vieweg; 2022.

Martinho, A., Poulsen, A., Kroesen, M. et al. *Perspectives about artificial moral agents*. *AI Ethics*1, 477–490 (2021). <https://doi.org/10.1007/s43681-021-00055-2>

Meyer, L.: Art. *Pflichtenethik* (Version 1.0 vom 12.10.2017), in: Ethik-Lexikon, verfügbar unter: <https://www.ethik-lexikon.de/lexikon/pflichtenethik>

Misselhorn C. Artificial Moral Agents: Conceptual Issues and Ethical Controversy. In: Voenekey S, Kellmeyer P, Mueller O, Burgard W, eds. *The Cambridge Handbook of Re-*

*sponsible Artificial Intelligence: Interdisciplinary Perspectives*. Cambridge Law Handbooks. Cambridge University Press; 2022:31-49.

Misselhorn, Catrin: Maschinenethik und Artificial Morality“, in: bpb.de, 17.02.2022, [online] <https://www.bpb.de/shop/zeitschriften/apuz/263684/maschinenethik-und-artificial-morality/>; 08.04.2025.

Schedel, T.: Art. *Utilitarismus* (Version 1.0 vom 12.11.2018), in: Ethik-Lexikon, verfügbar unter: <https://www.ethik-lexikon.de/lexikon/utilitarismus>

Toyka-Seid, Gerd Schneider / Christiane: Moral, in: bpb.de, 30.03.2023, [online] <https://www.bpb.de/kurz-knapp/lexika/das-junge-politik-lexikon/320812/moral/>; 05.04.2025.

Sophie Jentsch, Patrick Schramowski, Constantin Rothkopf, and Kristian Kersting. 2019. Semantics Derived Automatically From Language Corpora Contain Human-like Moral Choices. In 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES'19), January 27–28, 2019, Honolulu, HI, USA. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3306618.3314267>

## Persönliche Erklärung

Hiermit erkläre ich, Daniel Salit, dass ich diese Arbeit selbstständig und ohne fremde Hilfe angefertigt und nur die im Literaturverzeichnis angegebenen Quellen und Hilfsmittel benutzt habe. Wörtliche Zitate und sinngemäße Wiedergaben habe ich als solche gekennzeichnet.

10.04.2025  
(Datum)

  
(Unterschrift)