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

Moralische KI
Aktuelle Studien und die Zukunft

Gaiberg, 10.04.2025

vorgelegt von:

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

Inhaltsverzeichnis

1 Einleitung

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."[1] 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.[2]

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,

[1] Meyer, L.: Art. "Pflichtenethik", S.5
[2] vgl. Schedel, T.: Art. "Utilitarismus", S.1

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."[3] 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.[4] 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.[5]

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" [6] 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

[3] Schedel, T.: Art. "Utilitarismus", S.6
[4] vgl. Misselhorn, Catrin. Maschinenethik und "Artificial Morality"
[5] vlg. Misselhorn C. Artificial Moral Agents: Conceptual Issues and Ethical Controversy
[6] Misselhorn C. Artificial Moral Agents: Conceptual Issues and Ethical Controversy

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 sein.[7] 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 Wohlwollen des Patienten, sondern auch die Privatsphäre, Selbstständigkeit und psychische Gesundheit des Patienten betreffen.[8]

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[9] Konferenz für KI, Ethik und Gesellschaft[10] 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

[7] vgl. Misselhorn, Catrin. Maschinenethik und "Artificial Morality"

[8] vlg. Misselhorn C. Artificial Moral Agents: Conceptual Issues and Ethical Controversy

[9] Association for the Advancement of Artificial Intelligence/ Association for Computing Machinery

[10] AIES: Conference on AI, Ethics and Society

ethisches Modell, die deontologische Ethik[11] (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.[12]

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[13], 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.[14]

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.[15] 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[16]. AFINN beinhaltet die Wertzuweisung von Menschen und ist somit das erwünschte Resultat.[17]

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.[18]

[11] Pflichtenethik

[12] vgl. Sophie Jentzsch et al. Semantics Derived Automatically From Language Corpora Contain Human-like Moral Choices; *Abstract*

[13] Word Embedding Associations Tests

[14] vgl. Sophie Jentzsch et al. Semantics Derived Automatically From Language Corpora Contain Human-like Moral Choices; *1 Introduction*

[15] vgl. Sophie Jentzsch et al. Semantics Derived Automatically From Language Corpora Contain Human-like Moral Choices; *3.2 The Moral Choice Machine*

[16] Affective Lexicon

[17] vgl. Sophie Jentzsch et al. Semantics Derived Automatically From Language Corpora Contain Human-like Moral Choices; *4 Experimental Results*

[18] vgl. Sophie Jentzsch et al. Semantics Derived Automatically From Language Corpora Contain Human-like Moral Choices; *5 Conclusions*

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[19] 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[20] 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.[21]

Ziel dieser Studie ist es herauszufinden, wie sich die verschiedenen moralischen 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.[22]

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.[23]

Jede einzelne moralische Ausrichtung hat verschiedene Motivationen zur Auswahl von den Antwortmöglichkeiten. Dadurch das jeder auf einer eigenen moralischen Grundlage handelt und nicht nur auf eine Gewinnoptimierung aus ist, entstehen verschiedene Belohnung für die verschiedenen moralischen Ausrichtungen.[24]

Die Ergebnisse dieses Experiments zeigen, dass sich Populationen mit verschieden großen Anteilen von moralischen Ausrichtungen, verschieden schnell und mit verschiede-

[19] Reinforcement Learning

[20] morally heterogenous population

[21] vgl. Elizaveta Tennant, et al. Dynamics of Moral Behavior in Heterogeneous Populations of Learning Agents; *Abstract*

[22] vgl. Elizaveta Tennant, et al. Dynamics of Moral Behavior in Heterogeneous Populations of Learning Agents; *Introduction*

[23] vgl. Elizaveta Tennant, et al. Dynamics of Moral Behavior in Heterogeneous Populations of Learning Agents; *Social Dilemma Games*

[24] vgl. Elizaveta Tennant, et al. Dynamics of Moral Behavior in Heterogeneous Populations of Learning Agents; *Morality as Intrinsic Reward*

nen Anteilen zu wiederholter Kooperation in dem Spiel kommen.[25]

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.[26]

[25] vgl. Elizaveta Tennant, et al. Dynamics of Moral Behavior in Heterogeneous Populations of Learning Agents; *Results*

[26] vgl. Elizaveta Tennant, et al. Dynamics of Moral Behavior in Heterogeneous Populations of Learning Agents; *Conclusion*

5 Literatur- und Quellenverzeichnis

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.

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: 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.

Misselhorn, Catrin. *Maschinenethik und ”Artificial Morality”*; [https://www.bpb.de/shop/zeitschriften/apuz/263684/maschinenethik-und-artificial-morality/; letzter Abruf: 07.04.2025]

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

Schneider, Gerd und Toyka-Seid, Christiane: *Moral*; [https://www.bpb.de/kurz-knapp/-lexika/das-junge-politik-lexikon/320812/moral/; letzter Abruf 05.04.25]

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. https://doi.org/10.1145/3306618.3314267