Ethische Grundbegriffe für eine Maschinenethik und ein Fallbeispiel zur Klärung von Grundlagenfragen

Philip Wälde, M.A. (4556354)
B.Sc. Student (Informatik)
FU Berlin
Fakultät für Mathematik und Informatik
Email: p.waelde@fu-berlin.de
Eingereicht am 20.04.2020 bei Prof. Dr. C. Benzmüller

Abstract

Grundlegende Begriffe werden im Rahmen der Maschinenethik eingeführt. Wichtige Kriterien für ethische Fragestellungen und insbesondere deontologischer Ansätze werden entwickelt. Für eine angewandte Ethik wird als Beispiel angeführt wie Leibniz' Unterscheidung von logischer vs. Realmöglichkeit bei ethischen Dilemmata wie "Sollen impliziert Können" oder der Dyadic Deontic Logic eine Präzisierung erfährt. Systematische Untersuchungen formaler Systeme ermöglichen ein tieferes Verständnis der Strukturen und tragen zur Implementierung automatisierten normativen Schließens in KI-Systemen bei.

1. Einleitung

Maschinen-Ethik als eine Form angewandter Ethik wirft verschiedene Fragestellungen auf, die gleich mehrere Dimensionen der Ethik als Disziplin betreffen. Sowohl theoretische Fragestellungen der Meta-Ethik und normativen Ethik als auch praktische Fragestellungen, die mit der Modellierung und Implementierung symbolischer KI-Systeme sowie deren rechtlicher Regulation einhergehen, kennzeichnen diesen vielschichtigen Forschungsgegenstand. Der Fokus soll auf der im Ausgang von Kant entwickelten Deontologie ethischer Sollenssätze (Geboten, Verboten, Erlaubt) und der Möglichkeit liegen, die Formalisierungen deontischer Logik(en) für eine Präzisierung philosophischer Fragestellungen einzusetzen. Zunächst werden die notwendigen philosophischen Grundbegriffe eingeführt, um schließlich einen Kritierenkatalog zu entwickeln, der einer Identifikation ethischer Fragestellung dient. Im Anschluss an die einleitenden Bemerkungen zu einer Maschinenethik soll anhand eines Beispiels ein Ausblick geboten werden, wie sich

theoretische und praktische Fragestellungen anhand der Formalisierung deontischer Ethik entwickeln lassen und es dadurch ermöglichen, die Mehrdimensionalität ethischer-rechtlicher Fragestellungen aus Sicht der Informatik zu vertiefen und zu präzisieren. Hierbei soll der Frage nachgegangen werden, inwieweit die Unterscheidung von logischer und realer Möglichkeit bei Leibniz eine Vertiefung erfährt durch Herausforderungen, die sich bei deontischen Logiken stellen. Paradoxien wie etwa die These „Ein Sollen impliziert ein Können“ oder die Dyadische Deontische Logik (DDL) können als Anwendungsfälle der leibnizschen Unterscheidung aufgefasst werden, und die Herausforderung für ein hinreichend mächtiges System besteht darin, fähig zu sein zwischen logischer und realer Möglichkeit zu diskriminieren.

2. Klärung einiger Grundbegriffe für eine Maschinenethik

Wie viele Grundbegriffe ist auch die Ethik zunächst mehrdeutig. Etymologisch bedeutet ethos 1. Sitte/Gewohneit oder Charakter/Denkweise und ethikos 2. Sitte/die Sitte betreffend.[1] Mos/Moralis ist die lateinischen Übersetzung [1]. Es gilt zu beachten, dass sich der moderne Gebrauch von Moral unterscheidet, insofern wir unter Moral 1. ein Normensystem für das Verhalten von Menschen mit Anspruch auf unbedingte Gültigkeit verstehen und 2. Moral im Plural deskriptiv und ohne Wertung Moralsysteme beschreibt. Das Prädikat moralisch hingegen wird oft wertend gebraucht, so handelt jemand moralisch, wenn er sittlich gut handeln, und handelt unmoralisch, wenn er unsittlich schlecht handelt. Ethik ist demnach die Wissenschaft von der Moral, Moral(en) und Sitten sind hingegen ihr Untersuchungsgegenstand [2]. Folgende Einteilung in Grunddisziplinen bietet sich an:

1. die deskriptive Ethik fragt: Welche Moralen gibt es?
2. die normative Ethik fragt: Wie lassen sich Moralen begründen und bewerten?
3. die Metaethik fragt: Welchen generellen Status haben moralische Begriffe, Propositionen und Argumentationsweisen?
4. fragt die Angewandte Ethik : Wie lassen sich Domänen ethisch beurteilen?

[1] Der Ethos bedeutet zunächst den gewohnten Ort des Lebens und verweist somit bereits auf die Ethik als einer Disziplin die neben der Frage nach dem richtigen Handeln auch die Frage nach einem gelungenen Leben - der Eudaimonie stellt. Dieser Aspekt soll hier nicht weiter verfolgt werden.

In unserem Fall haben wir es mit der Domäne von Maschinen und insbesondere der Automatisierung normativen Schließens in autonomen Systemen zutun, die wiederum an die Domäne ethisch-rechtlicher Fragestellungen und Verfahrensweisen gebunden ist. Eine solche Bereichsethik setzt also sowohl Wissen um die allgemeine Ethik voraus als auch ein spezifisches Sachwissen der jeweiligen Domäne. Nur in dieser, notwendig interdisziplinären, Betrachtung lässt sich der Gegenstand folglich adäquat beurteilen. Eine besondere Herausforderung der Maschinenethik (und jeder Technikethik) besteht darin mittels bekannter und anerkannter Grundsätze[2] gänzlich neue - insbesondere technische - Handlungsmöglichkeiten zu beurteilen [3]. Somit ist die angewandte Ethik im Zwischenbereich allgemeiner Ethik und der Erörterung konkreter Fälle angesiedelt. Im Idealfall leistet sie hierbei sowohl einen Beitrag zu unserem theoretischen Verständnis ethischer Grundbegriffe und Theoriebildungen als auch konkrete und verbindliche Handlungsanweisungen für ihre jeweils spezifische Domäne. Die Maschinenethik befasst sich insbesondere mit der Problemstellung, die sich ergibt, wenn wir Maschinen mit der Fähigkeit zum moralischen Handeln und Entscheiden auszustatten versuchen.[3]

Innerhalb der normativen Ethik lassen sich nun drei Grundtypen unterscheiden. Die Tugendethik fokusiert 1. die Motivation einer Handlung, die deontologische Ethik reflektiert 2. auf die Handlung als solche und die utilitarische Ethik fragt 3. nach den Konsequenzen einer Handlung [4]. Jede dieser Positionen Versucht die Schwerpunkte der anderen Positionen aufzugreifen. Es bleibt oft strittig wie Motivation, Handlung und Konsequenz voneinander abzugrenzen sind, dennoch bietet diese Einteilung eine gute Orientierung über die jeweiligen Schwerpunkte. Insbesondere resultieren nämlich Differenzen in moralischen Urteilen aus der unterschiedlichen Betonung der Komponenten Motivation, Handlung und Konsequenz. Die drei Positionen eint wiederum ein generalistisches Verständnis, d. h. Moral ist eine Frage allgemeiner Prinzipien und konkrete Fälle werden diesen entweder induktiv oder deduktiv untergeordnet.

Unser Schwerpunkt liegt auf der deontologischen Ethik, sie beantwortet die Frage „Was soll ich tun?“ durch universalisierbare Maximen, die für jeden gelten sollen [5]. Immanuel Kants

[2] Zentral sind hier Grundsätze und Mechanismen aus der rechtlichen, sozialen, staatlichen und personalen Praxis.

[3] Im Folgenden werden personenbezogene Fähigkeiten bzw. Vermögen nur durch funktionale Äquivalenzen bei Maschinen betrachtet. Der Funktionalismus legt sich zunächst nicht ontologisch fest, welche Beziehung zwischen menschlichen und maschinellen Fertigkeiten besteht.

kategorischer Imperativ (KI) ist der paradigmatische Fall eines solchen Prinzips. Er gibt keinen Maßstab für die Beurteilung von Moral ab, sondern versucht die Moral allgemein gültig seiner Form nach festzuhalten. Dabei wird nicht instrumentell eine Mittel-Zweck-Relation, sondern ein Gesetz aufgestellt: „Handle nur nach derjenigen Maxime, durch die du zugleich wollen kannst, dass sie ein allgemeines Gesetz werde." (Kant: AA IV, 421). Zentral ist die Maxime als ein Prinzip des Willens, die eine Universalisierbarkeit (und folglich ein Gesetz) fordert, um den Handlungstyp festzulegen, nämlich als einer, der sich entweder durch seine Maxime bestimmen (universalisieren) lässt oder nicht. Die große Herausforderung besteht hier im richtigen Abstraktionsgrad. Wenn eine Handlung durch eine Maxime charakterisiert wird, die dem KI entspricht, darf sie weder zu konkret sein, denn dann läuft sie Gefahr mit anderen Maximen in Konflikt zu geraten, noch zu abstrakt, denn dann entfernt sie sich zu weit von der Handlung.

Die hier skizzierten Grundbegriffe ethischer Theorien adressieren Menschen als Handlungssubjekte und insbesondere bei der kantischen Position ist die notwendige Bedingung der Willensfreiheit für die Realisierung des kategorischen Imperativs eine notwendige Voraussetzung.[4] Wir haben es also allgemein mit menschlichem Verhalten zu tun. Deshalb ist es wichtig aus der Perspektive einer Maschinenethik zu beachten, dass wir zunächst funkionale Äquivalente untersuchen, die einen Pseudoakteur, eine Pseudohandlung etc. in einer Pseudoethik [6] modellieren. Hierbei werden normalsprachliche Begriffe der Ethik in formale Strukturen mit mathematisch-logischen Methoden überführt. Trotz dieser notwendigen Übersetzung und Reduktion lassen sich folgende Eigenschaften deontischer Ethik festhalten:

1. Positiv: Gesetzescharakter von Moralität
    1.1. Speziell bei Kant muss die Handlungsmaxime als Gesetz kategorisch d. h. unbedingt gelten
    1.2. Das unbedingte Gesetz ist folglich die deontologische Ausrichtung
2. Negativ: Nicht die Natur des Menschen darf entscheiden (Tugendethiker)
3. Negativ: Nicht die Umstände in der Welt dürfen ausschlaggebend sein (Utilitaristen)
4. Der Moral werden kontingente Bedingungen auferlegt, wodurch sie potentiell relativiert wird
    4.1. Stufen der Verbindlichkeit für komplexe Situationen

[4] So schreibt Kant in der Grundlegung der Metaphysik der Sitten: „Denn die reine und mit keinem fremden Zusatz von empirischen Anreizen vermischte Pflicht und überhaupt des sittlichen Gesetzes hat auf das menschliche Herz durch den Weg der Vernunft […] einen so viel mächtigeren Einfluss, als alle anderen Triebfedern.", (GMS: 410) [17].

Abschließend sei noch ein heuristischer Kriterienkatalog [7] für die Identifikation und Beschreibung ethischer Phänomene erwähnt. Dieser soll insbesondere dazu dienen die disziplinübergreifenden Diskussionen anzuleiten:

1. **Normen und Werte**: keine Ist-Sätze, sondern Sollens-Sätze.
2. **Allgemeingültigkeit**
3. **Unparteilichkeit**
4. **Universalisierbarkeit**: Ist etwas für eine Person moralisch geboten/verboten/erlaubt, dann ist es unter ähnlichen Umständen für jede hinreichend ähnliche Person geboten/verboten/erlaubt.
5. **Unbedingtheit**: Moralische Normen gelten unabhängig von anderen Bedingungen.
6. **Vorrangigkeit**: Priorität bestimmter moralischer Erwägungen vor anderen.
7. **Kontextunabhängigkeit**: Stabilität in der Gültigkeit moralischer Normen trotz Wandels.
8. **Sanktionen**: Innere und äußere bzw. positive und negative Rechte.

3. Ein Anwendungsbeispiel

3.1 Ethische Paradoxien und metaphysische Herausforderungen

Wenn wir über moralische Gebote nachdenken scheint ein intuitiver Sachverhalt zu bestehen, nämlich dass eine moralische Pflicht nur dann erfüllbar ist, wenn die Möglichkeit für den Akteur besteht diese zu erfüllen. Kurz: Jedes Sollen impliziert ein Können [8]. Wenn ein Akteur eine Handlung nicht ausführen kann, die geboten ist, dann kann diese Gebot nicht die Handlung des Akteurs anleiten. Und so wird geschlossen, das eine vermeintliche, aber nicht erfüllbare Pflicht nicht den Status einer Pflicht haben kann. Ob nicht erfüllbare Pflichten immer noch den Status eines absoluten Gebots haben, trotz der Unmöglichkeit in einer bestimmten Situation erfüllt zu werden, ist eine grundsätzliche metaethische Fragestellung in diesem Kontext. Es handelt sich dabei um eines von verschiedenen moralischen Dilemmata, die in normalsprachlichen und formalen Ethiksystemen auftreten können [9]. Die Rolle der Maschinenethik aus Sicht der Informatik besteht nun darin, dass sie einen wichtig Beitrag zur Klärung kontrovers diskutierter moralischer Dilemmata liefern kann, weil es durch ein metalogisches Framework wie z. B. Higher Order Logics (HOL) möglich ist, verschiedene Formalisierungen moralischer Dilemmata und die korrespondieren Logiktypen praktisch zu testen [10].

Eine wichtige Beobachtung für unser gewähltes Beispiel besteht darin, dass Prinzipien wie der KI in der Realisation durch eine Handlung kontingenten Widerfahrnissen bzw. Umständen ausgeliefert sind, die die Möglichkeit etwas zu tun oder zu unterlassen einschränken. Eine Handlung ist also nur eingeschränkt unter bestimmten Bedingungen möglich. Leibniz spricht vom konstitutiven Unterschied[5] zwischen logischer Möglichkeit und Realmöglichkeit [11]. Für die Operatoren Möglichkeit bzw. Notwendigkeit gilt, dass sie dadurch konjugiert sind, dass Aussagen notwendig sind, wenn ihre Negation nicht möglich ist. Syntaktisch bedeutet Möglichkeit erst einmal Widerspruchsfreiheit. Für die semantische Belegung kann man mit möglichen Welten operieren [12].

Für eine Handlung können wir subjektive und objektive Erfüllungsbedingungen formulieren, unter der ihr Vollzug gelingen oder misslingen kann. Zu den subjektiven Bedingungen zählen 1. Freiheitsbewusstsein und 2. Abwägungsfähigkeit in Bezug auf Gründe. Zu den objektiven Bedingungen zählt 1. die Abwesenheit von Zwang, 2. die Gelegenheit zwischen Alternativen entscheiden zu können und 3. eine Bestimmbarkeit ohne Notwendigkeit; also das Mögliche als Kontingentes (=Nichtnotwendiges). Diese zusammengenommen zentralen Bedingungen artikulieren nun einen Möglichkeitsbegriff, der eine ontologische Anbindung (und damit eine Einschränkung) an Wirkliches bildet [13].

Die logischen Möglichkeiten können im Kontrast zu realen Möglichkeiten dadurch charakterisiert werden, dass sie eine generelle Abhängigkeitsrelation etablieren im Gegensatz zu einer spezifizierten Möglichkeit [14]. Eine spezifizierte Möglichkeit zeichnet sich durch einen Wirklichkeitskontext aus, der die betreffende Möglichkeit bietet. Anders formuliert hängt das logisch Mögliche als possibile logicum an der Widerspruchsfreiheit innerhalb von Formulierungssystemen, während sich für das real Mögliche eine spezifisch ontologische Anbindung durch einen jeweiligen Wirklichkeitskontext angeben lässt.[6] In unserem Beispiel sind das Wirklichkeitskontexte die ein Gebot ermöglichen oder verhindern. Außerdem müssen einem Akteur echte Alternativen zur Verfügung stehen - das Errichten von Normen impliziert also die Möglichkeit zur Abweichung von der Norm. Ein strikter Determinismus,

[5] Freilich geht es Leibniz hierbei noch um die übergeordnete Frage nach dem Verhältnis von Kontingenz und Determiniertheit des Naturverlaufs bzw. die Frage nach Determiniertheit und menschlicher Freiheit. Dieser große Bogen kann hier allerdings nicht geschlagen werden, bildet aber den notwendigen Kontext für Leibniz‘ Philosophie.

[6] Leibniz entwickelt ein Kalkül ausgehend von der Annahme was widerspruchsfrei ist, ist möglich und entwickelt eine Semantik von Wahrheitheitswerten [12]. In der heutigen Form der Modallogik möglicher Welten ist eine Aussage notwendig, wenn sie in allen möglichen Welten gilt, möglich wenn sie in mindestens einer Welt wahr ist bzw. ihre Negation nicht für alle Welten gilt [14].

der sich von der logischen Notwendigkeit leiten ließe würde in letzter Konsequenz diese Alternativen von Realmöglichkeiten nivellieren.[7] Weil die im Ausgang von einem Handlungssubjekt skizzierte Unterscheidung von logischer und Realmöglichkeit auf ein frei handelndes moralisches Subjekt inklusive Determination in einem umfassenden System abzielt, wie es Leibniz entworfen hat. Doch lässt sich auch ohne die anspruchsvolle Konstruktion eines freien Handlungssubjekts ein zentraler Aspekte für pseudoethische Agenten herausarbeiten. Ein deontisch verfasster Pseudoagent bzw. ein „ethical reasoner" sollte, sofern er aktiv in seine Umwelt eingreift und mit ihr interagiert, in der Lage sein nicht nur eine Semantik möglicher Welten zu etablieren, sondern darüber hinaus den für kausal wirksame Handlungen entscheidenden Faktor der Realmöglichkeit miteinbeziehen. Auch ohne einen strittigen Begriff von Willensfreiheit und das damit einhergehende Vermögen einer Person, seine Handlung entweder zum Wollen oder nicht Wollen zu bestimmen, benötigt ein Pseudoagent funktionale Äquivalente. Dazu gehört die Möglichkeit normativ gebotene, verbotene oder erlaubte Maximen im Sinne des KI formal korrekt zu erfassen. Die Maximen müssen hierbei auf ein Weltformat von Handlungskandidaten X in einer Situation Y übersetzt werden. Darin bestünde die Fähigkeit eines KI-Systems aus der Menge logischer Möglichkeiten die Untermenge von Realmöglichkeiten zu repräsentieren.

3.2 Dyadic Deontic Logic und Ausblick

Abschließend soll ein Beispiel für die zunächst im Kontext der Willensfreiheit bei Leibniz diskutierte Frage nach logischen vs. realen Möglichkeiten innerhalb formaler normativer Systeme kurz vorgestellt werden. Neben der bereits gut etablierten Standard Deontic Logic (SDL) ist die Dyadic Deontic Logic (DDL) ein wichtiges modallogisches Paradigma, um normatives Schließen zu studieren. DDL zeichnet sich durch einen Operator aus, der Konditionalität in der Notation O (p/q) einführt, gelesen als „es sollte der Fall sein das p, gegeben q". Dadurch ist es innerhalb von DDL möglich durch Konditionale Normverletzungen zu beschreiben [9]. Hierbei werden sog. contrary-to-duty Szenarien beschrieben, die Fälle durchspielen in denen eine Norm sagt was getan werden soll, wenn etwas Verbotenes wahr ist: Wenn Person X den Mord begangen hat, dann sollte Person X ein

[7] Wobei hier der Unterschied zwischen Kausalität und Notwendigkeit zu beachten ist [15]. Insofern wird hier eine kompatibilistische These vertreten hinsichtlich Determination und der Möglichkeit freier Handlung. Determination ist dann keine notwendige Determination.

Geständnis ablegen. Diese Szenarien sind für den deontologischen Charakter des Rechts von Bedeutung, weil Rechtspflichten einklagbare gebotene Pflichten sind, die ein positives Tun oder negatives Unterlassen regeln und dabei auch Pflichtenkollisionen berücksichtigen müssen, wenn z. B. eine Menge von Pflichten zum gleichen Zeitpunkt oder in einer bestimmten Reihenfolge nicht erfüllbar sind. In diesen Fällen müssen verschiedene Normtypen unterschieden und mit unterschiedlicher Dringlichkeit versehen werden, um Normen durch Verbindlichkeitsstufen zu regeln. DDL schafft es auch hier absolute Pflichten von relativen Pflichten (mit Ausnahmefällen) zu unterscheiden. Somit leisten ethische Theorien und deren Formalisierung einen wichtigen Beitrag zu einer metaphysischen Unterscheidung zwischen logischer Möglichkeit und Realmöglichkeit, weil sie eine Anwendungsdomäne bieten, um sich mit den Herausforderungen einer korrekten Repräsentation verschiedener Möglichkeitsklassen auseinandersetzen und sind darüber hinaus für die Anwenderperspektive relevant, weil hier insbesondere Rechtssysteme wichtige Regelwerke im Umgang mit solchen Situationen bieten. Die Chance der Maschinenethik neben der Dringlichkeit automatisierte Regulationsmechanismen zu entwickeln besteht in einem grundlegenden Verständnis für metaphysische und logische Fragestellungen samt ihrer praktischen Implikation. Maschinenethik ist also beides - Theorie und Praxis.

Literatur:

[1] O. Höffe. Ethik. Eine Einführung. C.H. Beck, 2013, 9-11.

[2] D. Hübner. Einführung in die philosophische Ethik. Vandenhoeck & Ruprecht, 2018, 17-21.

[3] J. Loh. Roboterethik. Eine Einführung. Suhrkamp, 2019, 19-35.

[4] D. Hübner. Einführung in die philosophische Ethik. Vandenhoeck & Ruprecht, 2018, 17-21. 88-99.

[5] L. Alexander, M. Moore, Deontological Ethics, in: E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy, 2016 Edition, Metaphysics Research Lab, Stanford University, 2020, URL https://plato.stanford.edu/entries/ethics-deontological/

[6] C. Misselhorn. Grundfragen der Maschinenethik. Reclam, 2018, 29-33.

[7] C. Misselhorn. Grundfragen der Maschinenethik. Reclam, 2018, 49-51.

[8] T. McConnell, Moral Dilemmas, in: E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy, 2018 Edition, Metaphysics Research Lab, Stanford University, 2020, URL https://plato.stanford.edu/entries/moral-dilemmas/#DilCon.

[9] J. Carmo, A. J. I. Jones, Deontic logic and contrary-to-duties, in: D. M. Gabbay, F. Guenthner (Eds.), Handbook of Philosophical Logic: Volume 8, Springer Netherlands, Dordrecht, 2002, pp. 265–343.

[10] Designing Normative Theories of Ethical and Legal Reasoning: LogiKEy Framework, Methodology, and Tool Support (Christoph Benzmüller, Xavier Parent, Leendert van der Torre), In Artificial Intelligence (to appear, minor revisions), Elsevier, pp. 1--50, 2020.

[11] G. W. Leibniz. Monadologie und andere metaphysische Schriften. Meiner, 2002, 129 (§43-44).

[12] Malink, M., Va, Meinersudevan, A.: The logic of Leibnizs Generales Inquisitiones de analysis notionum et veritatum. Rev. Symbolic Logic 9, 686–751 (2016)

[13] T. Buchheim, Freiheit und Determination, in: Freiheit. Stuttgarter Hegel-Kongress 2011, hrsg. von Gunnar Hindrichs und Axel Honneth, Frankfurt a.M. (Klostermann), 99-117.

[14] S. Kripke. Naming and Necessity. Blackwell, 1980, 1-22.

[15] B. Russel. On the Notion of Cause. Longmans Green, 1917, 190.

[17] I. Kant. Grundlegung zur Metaphysik der Sitten. Meiner, 1999, AA IV, 421 / GMS, BA 52.