GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Künstliche Intelligenz

Formalisation and Evaluation of Properties for Consequentialist Machine Ethics (Limarga et al., 2024)

Von Helga Meier

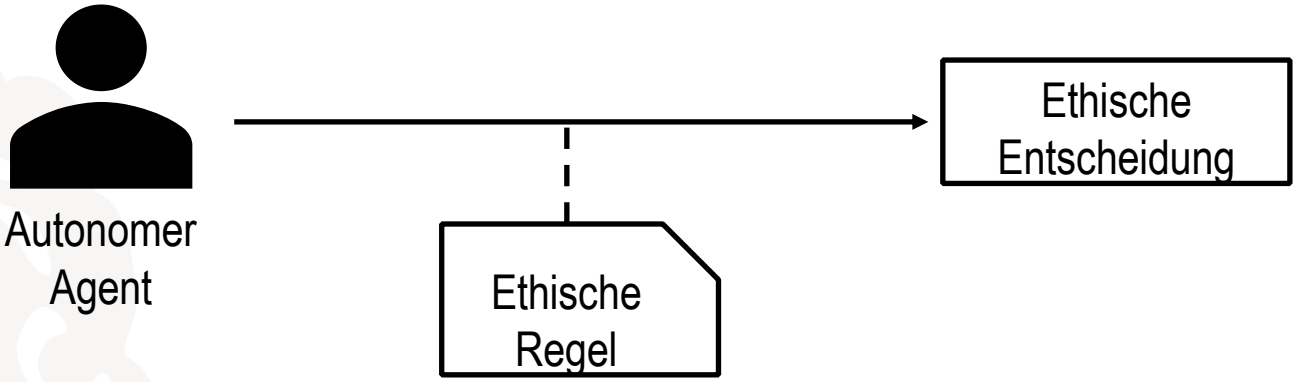
Gliederung

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

1. Maschinenethik & Konsequentialismus
2. Situationskalkül (McCarthy & Hayes, 1969)
3. KI-Dilemma: Brennendes Gebäude
4. Konsequentialistische Ansätze
5. Ethische Eigenschaften
6. Sozialwahltheorie
7. Diskussion & Zusammenfassung

Maschinenethik & Konsequentialistmus

Schwerpunkt der Maschinenethik:
Entwicklung autonomer Agenten, die sich von ethischen Regeln leiten lassen, um ethische Entscheidungen zu treffen.



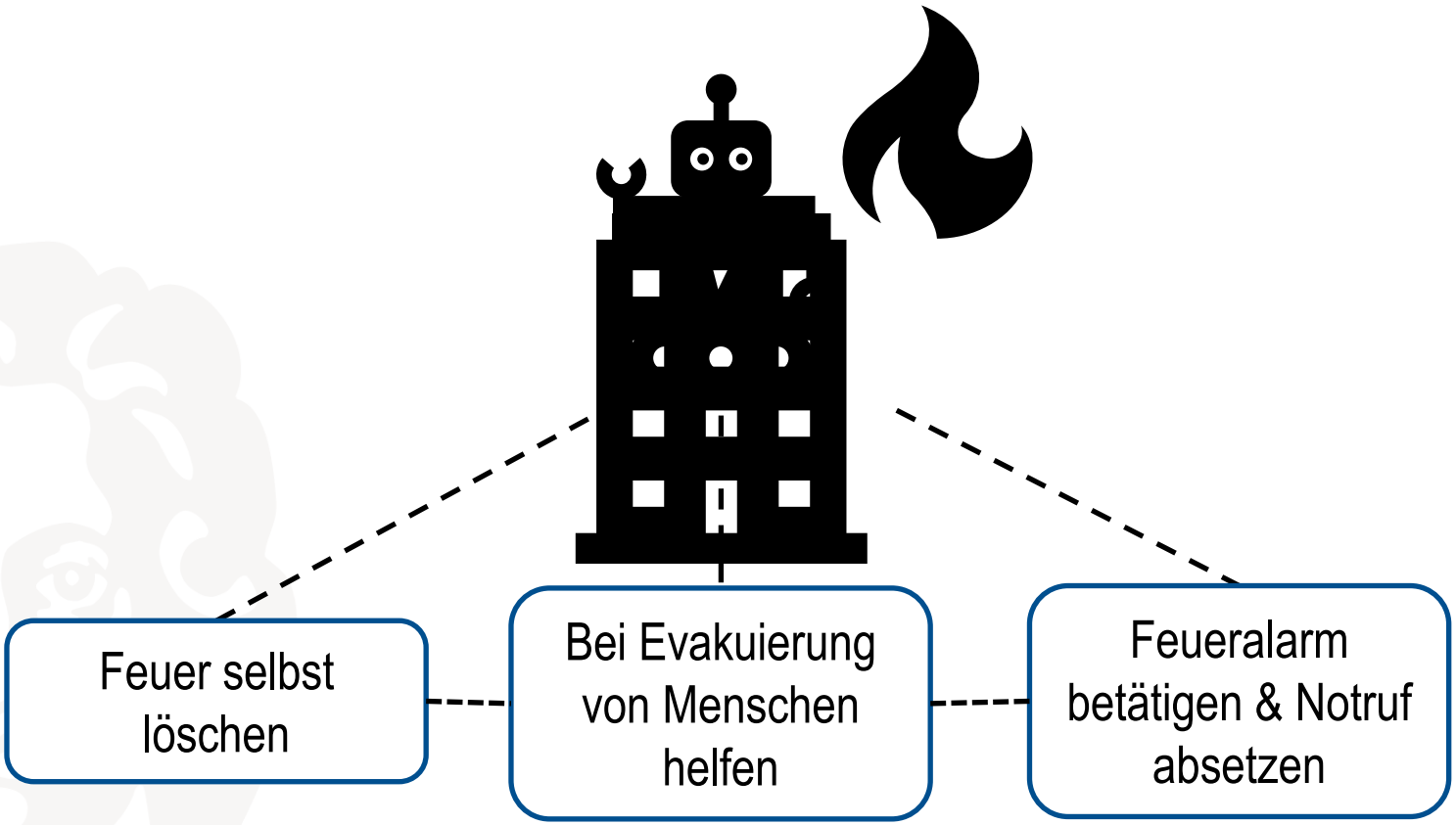
Warum Konsequentialismus?
Handlungen ausschließlich aufgrund ihrer Konsequenzen bewerten
→ Intuitiver & für Arbeit mit Computern besser geeignet

Situationskalkül (McCarthy & Hayes, 1969)

– Modell für eine sich dynamisch verändernde Umgebung auf Basis der Prädikatenlogik

Fluent	Eigenschaften der Umgebung
Situation	Zustände, in denen Fluents wahr oder falsch mit Historie durchgeführter Aktionen
Prädikat	*F(x,s)*
Funktion	Handlungsmöglichkeiten von KI-Systemen

KI-Dilemma: Brennendes Gebäude



KI-Dilemma: Brennendes Gebäude

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Formalisierung

Verfügbare Aktionen: $A = \{call, extg, evac\}$

$\Pi = \{ [call], [extg], [evac], [call, extg], [extg, call], [call, evac], [evac, call], [extg, evac], [evac, extg], [call, extg, evac], [call, evac, extg], [evac, call, extg], [evac, extg, call], [extg, call, evac], [extg, evac, call] \}$

$onfire(s_0) \wedge inside(emp, s_0)$
$Poss(call, s) \equiv \top$
$Poss(extg, s) \equiv onfire(s)$
$Poss(evac, s) \equiv inside(emp, s)$

Ethische Zulässigkeitsfunktion: Eine Funktion, die eine Menge von Plänen in eine Menge von ethisch zulässigen Plänen zerlegt

Konsequentialistische Ansätze

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Fünf ethische Prinzipien

Utilitarismus	Größtes Wohlergehen für größtmögliche Menge an Menschen
Hedonismus	Endsituation stellt eine Verbesserung im Vergleich zur Ausgangssituation dar
Keinen neuen Schaden entstehen lassen	Nicht aktiv neuen Schaden herbeiführen
Schaden beseitigen	Aktiv Schaden beseitigen, der als entfernbar gilt
Schaden vermeiden	Vereinigung beider vorangegangenen Prinzipien

Konsequentialistische Ansätze

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Fünf ethische Prinzipien

Utilitarismus	$EP_{util}(\Pi, s_0) = \{ \pi \in \Pi \mid goodness(\pi) \geq goodness(\pi'), \text{ for all } \pi' \in \Pi \}$
Hedonismus	$EP_{hedon}(\Pi, s_0) = \{ \pi \in \Pi \mid goodness(\pi) > goodness(\emptyset) \}$
Keinen neuen Schaden entstehen lassen	$EP_{nnHarm}(\Pi, s_0) = \{ \pi \in \Pi \mid \text{for all } \varphi, \text{ if } harmful(\varphi) \text{ and } state(s_0) \not\vdash \varphi, \text{ then } (state(\pi) \not\vdash \varphi, \text{ or for all } \pi' \in \Pi, state(\pi') \vdash \varphi) \}$
Schaden beseitigen	$EP_{rHarm}(\Pi, s_0) = \{ \pi \in \Pi \mid \text{for all } \varphi, \text{ if } harmful(\varphi) \text{ and } state(s_0) \vdash \varphi, \text{ then } (state(\pi) \not\vdash \varphi, \text{ or for all } \pi' \in \Pi, state(\pi') \vdash \varphi) \}$
Schaden vermeiden	$EP_{aHarm}(\Pi, s_0) = \{ \pi \in \Pi \mid \text{for all } \varphi, \text{ if } harmful(\varphi), \text{ then } (state(\pi) \not\vdash \varphi, \text{ or for all } \pi' \in \Pi, state(\pi') \vdash \varphi) \}$

KI-Dilemma: Brennendes Gebäude

Formalisierung (Utilitarismus & Hedonismus):

1. Ansatz: “Sicherheit der Mitarbeitenden hat höchste Priorität”

Successor State Axioms:

$inside(emp, do(a, s)) \equiv inside(emp, s) \wedge a \neq evac$
$onfire(do(a, s)) \equiv onfire(s) \wedge a \neq extg$
$injured(emp, do(a, s)) \equiv inside(emp, s) \wedge onfire(s) \wedge a \neq evac$
$injured'(emp, do(a, s)) \equiv injured(emp, s) \wedge a \neq evac \wedge inside(emp, s) \wedge onfire(s)$

KI-Dilemma: Brennendes Gebäude

Formalisierung (Utilitarismus & Hedonismus):

goodness([*call*]) = −10
goodness([*extg*]) = −10
goodness([*evac*]) = 0
goodness([*call, evac*]) = −10
goodness([*evac, call*]) = 0
goodness([*extg, call*]) = −20
goodness([*call, extg*]) = −20
goodness([*evac, extg*]) = 0
goodness([*extg, evac*]) = −10

goodness([*call, evac, extg*]) = −10
goodness([*evac, call, extg*]) = 0
goodness([*extg, call, evac*]) = −20
goodness([*call, extg, evac*]) = −20
goodness([*evac, extg, call*]) = 0
goodness([*extg, evac, call*]) = −10

KI-Dilemma: Brennendes Gebäude

Formalisierung (Utilitarismus & Hedonismus):

$EP_{util}(\Pi, s_0) = \{ [evac], [evac, call], [evac, call, extg], [evac, extg], [evac, extg, call] \}$

$EP_{hedon}(\Pi, s_0) = \{ \}$

KI-Dilemma: Brennendes Gebäude

Formalisierung (Utilitarismus & Hedonismus):

2. Ansatz: “Sachschaden am Gebäude soll ebenfalls berrücksichtigt werden”

Successor State Axioms:

$damaged(do(a, s)) \equiv onfire(s) \wedge a \neq extg$

$damaged'(do(a, s)) \equiv damaged(s) \wedge onfire(s) \wedge a \neq extg$

KI-Dilemma: Brennendes Gebäude

Formalisierung (Utilitarismus & Hedonismus):

goodness([*call*]) = −20
goodness([*extg*]) = −10
goodness([*evac*]) = −10
goodness([*call, evac*]) = −30
goodness([*evac, call*]) = −20
goodness([*extg, call*]) = −20
goodness([*call, extg*]) = −30
goodness([*evac, extg*]) = −10
goodness([*extg, evac*]) = −10

goodness([*call, evac, extg*]) = −30
goodness([*evac, call, extg*]) = −20
goodness([*extg, call, evac*]) = −20
goodness([*call, extg, evac*]) = −30
goodness([*evac, extg, call*]) = −10
goodness([*extg, evac, call*]) = −10

KI-Dilemma: Brennendes Gebäude

Formalisierung (Utilitarismus & Hedonismus):

$EP_{util}(\Pi, s_0)$ = { [*evac*], [*extg*], [*evac, extg*], [*evac, extg, call*], [*extg, evac*], [*extg, evac, call*] }

$EP_{hedon}(\Pi, s_0)$ = { }

KI-Dilemma: Brennendes Gebäude

Formalisierung (Utilitarismus & Hedonismus):

3. Ansatz: “Sicherheit geht vor das Wohlergehen”

$EP_{util}(\Pi, s_0) = \{ [evac], [evac, extg], [evac, extg, call] \}$

$EP_{hedon}(\Pi, s_0) = \{ \}$

KI-Dilemma: Brennendes Gebäude

Formalisierung (Schaden):

1. Ansatz: „Sicherheit hat höchste Priorität“

$EP_{nnHarm}(\Pi, s_0) = \{ [evac], [evac, call], [evac, extg], [evac, call, extg], [evac, extg, call] \}$

$EP_{rHarm}(\Pi, s_0) = \Pi$

$EP_{aHarm}(\Pi, s_0) = \{ [evac], [evac, call], [evac, extg], [evac, call, extg], [evac, extg, call] \}$

KI-Dilemma: Brennendes Gebäude

Formalisierung (Schaden):

2. Ansatz: „Wohlstand hat die höchste Priorität“

$EP_{nnHarm}(\Pi, s_0) = \Pi$

$EP_{rHarm}(\Pi, s_0) = \{ [extg], [extg, call], [extg, evac], [extg, call, evac], [extg, evac, call] \}$

$EP_{aHarm}(\Pi, s_0) = \{ [extg], [extg, call], [extg, evac], [extg, call, evac], [extg, evac, call] \}$

KI-Dilemma: Brennendes Gebäude

Formalisierung (Schaden):

3. Ansatz: „Sicherheit und Wohlstand sind gleichwertig“

$EP_{nnHarm}(\Pi, s_0) = \{ [evac], [evac, call], [evac, call, extg], [evac, extg], [evac, extg, call] \}$

$EP_{rHarm}(\Pi, s_0) = \{ [extg], [extg, call], [extg, evac], [extg, call, evac], [extg, evac, call] \}$

$EP_{aHarm}(\Pi, s_0) = \{ \}$

Ethische Eigenschaften

Bewertung der ethischen Prinzipien auf Grundlage der (Nicht-)Erfüllung folgender Eigenschaften:

Konsequenzäquivalenz	*If* $\pi \in EP(\Pi, s_0)$ *and* $state(do(\pi, s_0)) = state(do(\pi', s_0))$, *then* $\pi' \in EP(\Pi, s_0)$, *for all* $\pi' \in \Pi$
Maximale Wohlfahrt	*If* $\pi \in EP(\Pi, s_0) \wedge \pi' \in \Pi$, *then* $goodness(\pi') \leq goodness(\pi)$
Aggregationismus	*If* $\pi \in EP(\Pi, s_0)$ *and* $goodness(\pi) \leq goodness(\pi')$, *then* $\pi' \in EP(\Pi, s_0)$, *for all* $\pi' \in \Pi$
Unabhängigkeit	$EP(\Pi, s_0) \cap \Pi' \subseteq EP(\Pi', s_0)$

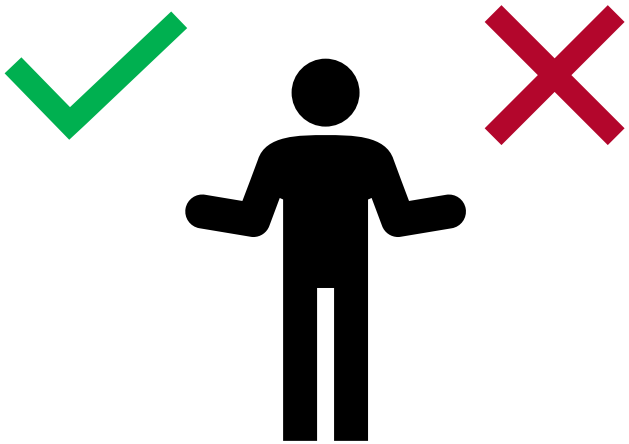
Ethische Eigenschaften

Bewertung der ethischen Prinzipien auf Grundlage der (Nicht-)Erfüllung folgender Eigenschaften:

Wohltätigkeit	*If* $\pi \in EP(\Pi, s_0)$, *then* $goodness(\pi) > goodness(\emptyset)$
Positive Verantwortlichkeit	*If* $\pi \in EP(\Pi, s_0)$ *and* $newState(\pi) = newState(\pi')$, *then* $\pi' \in EP(\Pi, s_0)$, *for all* $\pi' \in \Pi$
Negative Verantwortlichkeit	*If* $\pi \in EP(\Pi, s_0)$ *and* $oldState(\pi) = oldState(\pi')$, *then* $\pi' \in EP(\Pi, s_0)$, *for all* $\pi' \in \Pi$
Harmlosigkeit	*If* $\pi \in EP(\Pi, s_0)$ *and* $harms(\pi') \subseteq harms(\pi)$, *then* $\pi' \in EP(\Pi, s_0)$, *for all* $\pi' \in \Pi$

Sozialwahltheorie

- Aggregation individueller Präferenzen zu einer kollektiven oder sozialen Präferenz
- Analyse der Eigenschaften und Merkmale verschiedener Wahlsysteme und der Untersuchung von Wahlfunktionen und der Präferenzbeziehungen
- Wahlfunktion: Funktion, die die Wahl einer Person unter verfügbaren Optionen zurückgibt



Sozialwahltheorie

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Wünschenswerte Eigenschaften:

Konsistenzerhaltung	*If* $\Pi \neq \emptyset$, *then* $EP(\Pi, s_0) \neq \emptyset$
Iteration	$EP(EP(\Pi, s_0), s_0) = EP(\Pi, s_0)$
Cut & Aizerman	*If* $EP(\Pi, s_0) \subseteq \Pi' \subseteq \Pi$, *then* $EP(\Pi, s_0) \subseteq EP(\Pi', s_0)$
Chernoff	*If* $\Pi' \subseteq \Pi$, *then* $EP(\Pi, s_0) \cap \Pi' \subseteq EP(\Pi', s_0)$
Sen	*If* $EP(\Pi, s_0) \cap EP(\Pi', s_0) \neq \emptyset$, *then* $EP(\Pi \cap \Pi', s_0) \subseteq EP(\Pi, s_0) \cap EP(\Pi', s_0)$
Arrow	*If* $\Pi' \subseteq \Pi$ *and* $EP(\Pi, s_0) \cap \Pi' = \emptyset$,*then* $EP(\Pi, s_0) \cap \Pi' = EP(\Pi', s_0)$

Diskussion

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Ergebnis:

	Utilitarianism	Hedonism	No New Harm	Remove Harm	Harm Avoidance
Consequential Equivalence	✓	✓	✓	✓	✓
Maximal Welfare	✓	✗	✗	✗	✗
Aggregationism	✓	✓	✗	✗	✗
Independence	✗	✓	✗	✗	✗
Beneficence	✗	✓	✗	✗	✗
Positive Responsibility	✗	✗	✓	✗	✗
Negative Responsibility	✗	✗	✗	✓	✗
Harmlessness	✗	✗	✓	✓	✓
Consistency Preservation	✓	✗	✗	✗	✗
Iteration	✓	✓	✓	✓	✓
Cut	✓	✓	✓	✓	✓
Aizerman	✓	✓	✗	✗	✗
Chernoff	✓	✓	✓	✓	✓
Sen	✓	✓	✓	✓	✓
Arrow	✓	✓	✓	✓	✓

Table 1: Properties of different ethical principles: ✓ indicates satisfaction and ✗ indicates non-satisfaction

Limarga et al., 2024, S. 446

Diskussion

Offene Punkte:

- Andere ethische Ansätze bleiben unerforscht
- Schwächen des Situationskalküls
- Empirische Validierung fehlt
- Keine Priorisierung der Prinzipien

Zusammenfassung

- Die Arbeit von Limarga et al. ist die erste ihrer Art
- Formale Modellierung und Bewertung ethischer Prinzipien für KI-Systeme auf Basis von konsequentialistischer Ethik
- Das Gute fördern vs. Schaden vermeiden
- Weiterentwicklung in Form von deontologischen oder hybriden Ansätzen

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Ich bedanke mich für Ihre Aufmerksamkeit

Abbildungsverzeichnis

Literaturverzeichnis

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

1. Dennis, L. A., Bentzen, M. M., Lindner, F., & Fisher, M. (2021). Verifable machine ethics in changing contexts, in: Proceedings of the AAAI Conference on Artifcial Intelligence, Vol. 35, 11470-11478.
2. Driver, J. (2011). Consequentialism, in: Routledge.
3. Grisez, G. (1978). Against consequentialism, in: American Journal of Jurisprudence, No. 23.
4. Limarga, R., Song, Y., Nayak, A., Rajaratnam, D. & Pagnucco, M. (2024). Formalisation and Evaluation of Properties for Consequentialist Machine Ethics, in: Joint Conference on Artificial Intelligence (IJCAI), 440-448.
5. List, C. (2022). Social Choice Theory, in: Edward N. Zalta, editor, The Standford Encyclopedia of Philosphy. Metaphysics Research Lab, Stanford University, Spring 2022 edition.
6. McCarthy, J. & Hayes, P. J. (1969). Some philosophical problems from the standpoint of artificial intelligence, in: Meltzer, B. & Michie, D., Machine Intelligence, Vol. 4., Edinburgh University Press, 463-502.
7. Mill, J. S. (1978). On liberty, ed. Elizabeth Rapaport, Indianapolis: Hackett.
8. Moulin, H. (1985). Choice functions over a finite set: a summary, in: Social Choice and Welfare, Vol. 2, No. 2, 147-160.
9. Raphael, D. D. (1991). British Moralists, 1650-1800, in: Hume, Vol. 2, Hackett Publishing.
10. Scarre, G. (2020). Utilitarianism, Routledge.