Bachelorthesis

Berechnung von Gründen unter verschiedenen ethischen Prinzipien

Katrin Möllney

Gutachter: Prof. Dr. Bernhard Nebel
Betreuer: Dr. Felix Lindner

Albert-Ludwigs-Universität Freiburg
Technische Fakultät
Institut für Informatik
Lehrstuhl für Grundlagen der Künstlichen Intelligenz

5. Februar 2019

Bearbeitungszeit

07. 11. 2018 – 07. 02. 2019

Gutachter

Prof. Dr. Bernhard Nebel

Betreuer

Dr. Felix Lindner

Erklärung

Hiermit erkläre ich, dass ich diese Abschlussarbeit selbständig verfasst habe, keine anderen als die angegebenen Quellen/Hilfsmittel verwendet habe und alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten Schriften entnommen wurden, als solche kenntlich gemacht habe. Darüber hinaus erkläre ich, dass diese Abschlussarbeit nicht, auch nicht auszugsweise, bereits für eine andere Prüfung angefertigt wurde.

Ort, Datum	Unterschrift

Zusammenfassung

Die vorliegende Arbeit befasst sich mit der Generierung von Gründen unter verschiedenen ethischen Prinzipien. Es werden Spezifikationen für eine Begründung der verschiedenen ethischen Prinzipien erarbeitet. Diese Spezifikationen leiten sich aus einer Analyse der einzelnen Bedingungen ab, welche für ein ethisches Prinzip notwendig sind um über die (Un-)Zulässigkeit einer Handlung zu entscheiden. Des Weiteren werden drei Ansätze zur Implementierung solcher Begründungen präsentiert und diese exemplarisch auf ihre Intuitivität untersucht.

Inhaltsverzeichnis

1 Einleitung

Roboter und Künstliche Intelligenz (KI) sind heutzutage immer weiter verbreitet und werden auch zunehmend in Situationen eingesetzt, in welchen sie aktiv Entscheidungen treffen müssen, wie es beispielsweise bei selbstfahrenden Fahrzeugen der Fall ist. Bei einem selbstfahrenden Fahrzeug muss die Künstliche Intelligenz in der Lage sein, Entscheidungen zu treffen, die Menschenleben aktiv beeinflussen, und sie ist auch mit der Situation konfrontiert, dass sie sich entscheiden muss, welches Menschenleben gerettet werden soll und welches nicht [1]. Wenn eine Künstliche Intelligenz einem solchen moralischen Dilemma ausgesetzt ist und eine Entscheidung treffen muss, ist es wichtig, dass sie diese Entscheidung auch begründen kann, da es einem menschlichen Nutzer sonst schwer fällt, den Entscheidungen der KI sein Vertrauen zu schenken. Durch eine Erklärung wird eine gewisse Transparenz geboten, wie, beziehungsweise basierend auf welchen Grundsätzen, die KI ihre Entscheidung trifft. Aufgrund dieser Transparenz fällt es dem menschlichen Nutzer leichter, die Entscheidung der KI zu verstehen und nachzuvollziehen, weshalb er dieser somit eher vertrauen kann, als wenn dem Nutzer lediglich die Entscheidung der KI zurückgegeben wird.

Diese Arbeit beschäftigt sich mit der Begründung einer Entscheidung in einem moralischen Dilemma. Diese Entscheidung wird basierend auf verschiedenen ethischen Prinzipien getroffen. Als Künstliche Intelligenz dient im Rahmen dieser Arbeit der Roboterkopf IMMANUEL (Interactive Moral Machine bAsed on MUltiple Ethical principLes) [2]. Es handelt sich um einen hybriden ethischen Schlussfolgerungs-Agenten (Hybrid Ethical Reasoning Agent, kurz HERA), welcher für verschiedene ethische

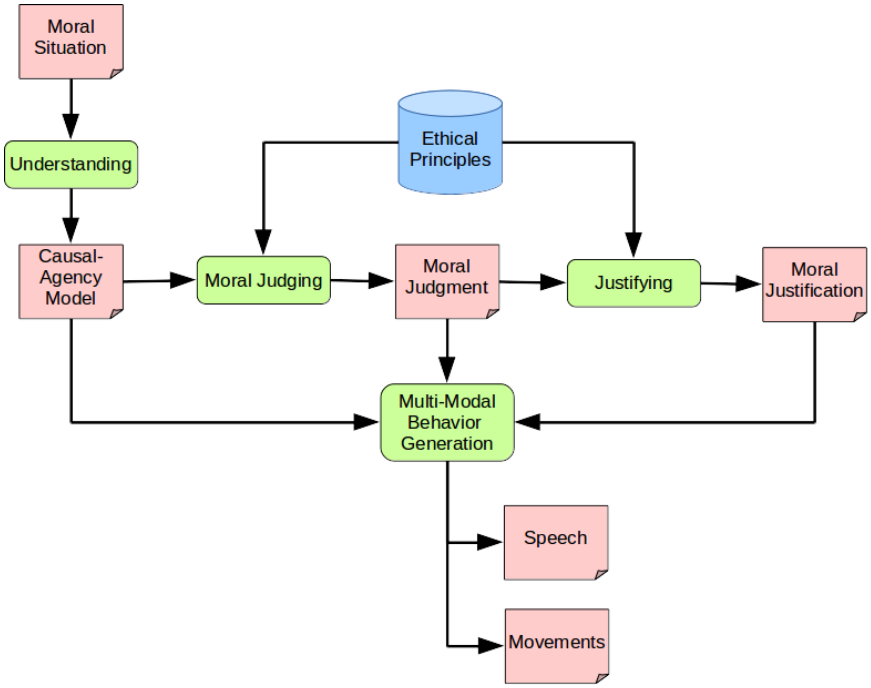

Abbildung 1: Skizze zum ethischen Schlussfolgern

Prinzipien anhand Kausaler Modelle Entscheidungen in moralischen Dilemmata treffen kann [2]. In Abbildung 1 ist der Vorgang des ethischen Schlussfolgerns schematisch dargestellt. Der Agent erhält eine moralische Situation als Input und erstellt basierend auf dieser moralischen Situation ein Kausales Modell. Anhand des Kausalen Modells und eines ethischen Prinzips ist der Agent dann in der Lage eine Entscheidung zu treffen. IMMANUEL ist bisher so weit implementiert, dass er mit einem gegebenen Kausalen Modell und einem ethischen Prinzip eine Entscheidung treffen kann, ob eine Handlung erlaubt oder verboten ist. In Listing 1.1 ist exemplarisch der Python-Code des Deontologischen Prinzips dargestellt. Dieses Prinzip besagt, dass eine Handlung erlaubt ist, wenn sie moralisch gut oder indifferent ist. Diese Bedingung des Prinzips wird durch die Variable f dargestellt, welche eine Formel repräsentiert. Wenn diese Formel erfüllt ist, dann ist die betrachtete Handlung nach diesem Prinzip erlaubt, ansonsten ist die Handlung verboten.

```
class DeontologicalPrinciple(Principle):
    def __init__(self, model):
```

```
        super(DeontologicalPrinciple, self).__init__(model)
        self.label = "DeontologicalPrinciple"

    def _check(self):
        f = GEq(U(self.model.action), 0)
        self.formulae = [f]
        self.result = [self.model.models(f)]
        return self.result

    def permissible(self):
        self._check()
        return self.result == [True]
```

Listing 1.1: Python-Code des Deontologischen Prinzips

Basierend auf der Entscheidung, ob eine Handlung erlaubt oder verboten ist, soll IMMANUEL eine Begründung generieren und diese zurückgeben. Genau an diesem Punkt setzt diese Arbeit an. Es werden für verschieden ethische Prinzipien Vorschläge für Begründungsspezifikationen gemacht und diese anhand unterschiedlicher moralischer Dilemmata beurteilt.

Die vorliegende Arbeit ist folgendermaßen aufgebaut: Zuerst werden in Kapitel 3 für die Arbeit wichtige Definitionen, wie die des Kausalen Modells und der ethischen Prinzipien, vorgenommen und die betrachteten moralischen Dilemmata eingeführt. Anschließend werden in Kapitel 4 Spezifikationen zur Begründung der ethischen Prinzipien vorgestellt und die Möglichkeiten der Implementierung von Begründungen in Kapitel 5 betrachtet.

2 Stand der Forschung

Ein wichtiger Aspekt im Bezug auf Künstliche Intelligenz, auf den in der Forschung immer wieder hingewiesen wird, ist die Fähigkeit eines Agenten seine Entscheidung erklären zu können, um so Transparenz und Vertrauen gegenüber dem menschlichen Nutzer zu generieren. Dannenhauer, Floyd, Magazzeni und Aha [3] verweisen in ihrem Paper auf dieses Problem und stellen anhand eines Beispiels eine Erklärung für rebellisches Verhalten (Verweigern oder Verändern von Aufgaben) von Agenten, die über ihre Ziele räsonieren, dar. Daneben haben sie auch einige offene Forschungsfragen angesprochen, wobei folgende Frage im Bezug auf meine Arbeit besonders interessant ist: Wie können rebellierende Agenten erklären, wieso die verweigerte Handlung ihre ethischen Modelle verletzt (z.B. Verweigern von Handlungen, die Schaden verursachen)? (vgl. [3]). In Kapitel 4.2 werde ich mich mit der Frage befassen, wie ein Agent Entscheidungen in ethischen Modellen begründen kann. Auch Borgo, Cashmore und Magazzeni [4] sprechen den Aspekt des Vertrauens an, welches laut ihnen nur dann existieren kann, wenn der Mensch versteht, was ein KI-System erreichen möchte und wieso. Ihr Ansatz zur Lösung dieses Problems sieht vor, dass der zugrundeliegende KI-Prozess Rechtfertigungen und Erklärungen erstellen muss, die für den Nutzer sowohl transparent als auch nachvollziehbar sind (vgl. [4]). Damit sprechen sie einen entscheidenden Punkt an, da eine Erklärung, welche für den Nutzer nicht verständlich oder nachvollziehbar ist, auch nicht dazu beitragen kann, Vertrauen zu erzeugen. Langley, Meadows und Sridharan [5] betonen ebenfalls, dass autonome Agenten in der Lage sein müssen, ihre Entscheidungen und das Reasoning, welches ihre Entscheidung

hervorbringt, zu begründen, da Menschen den Agenten nur auf diese Weise vertrauen können. Sie legen dar, dass der erklärende Agent seine Entscheidungen und Gründe so kommunizieren muss, dass es für den Menschen verständlich ist. Deshalb sollte der Agent Informationen in Form von Vorstellungen, Zielen und Handlungen darstellen, mit welchen der Mensch vertraut ist. Ein Punkt, welcher auch in meiner Arbeit eine Rolle spielt, ist folgender: Dadurch, dass erklärbare Agenten durch das menschliche Bedürfnis, das Verhalten eines autonomen Systems verstehen zu wollen, motiviert sind, sollten menschliche Rechtfertigungen eine zentrale Rolle in dem Evaluationsprozess spielen (vgl. [5]). Es ist entscheidend, dass die Erklärung mit der Intuition des Menschen übereinstimmt, so dass es dem Nutzer leichter fällt, die Erklärung des Agenten nachvollziehen zu können. Langley et al. [5] weisen darauf hin, dass heutzutage eine immer größer werdende Abhängigkeit von KI-Systemen besteht, es aber bislang auf dem wichtigen Gebiet der erklärbaren Agenten wenig Forschung betrieben wurde.

Fox, Long und Magazzeni [6] stellen in ihrem Paper dar, dass der Bedarf für erklärbare KI hauptsächlich durch drei Gründe motiviert ist: Durch den Bedarf an Vertrauen, den Bedarf an Interaktion und den Bedarf an Transparenz. Einen wichtigen Punkt, den sie ansprechen, ist derjenige, dass eine Schwierigkeit beim Erstellen von Erklärungen darin liegt, zu verstehen, was eine Erklärung tatsächlich beinhalten muss. Eine Anforderung an eine Erklärung, die sie in diesem Bezug anbringen, ist die, dass die Erklärung versuchen sollte, eine Erkenntnis zu finden, welche dem Nutzer nicht zur Verfügung steht, von welcher der Nutzer aber annimmt, dass sie dem System zur Verfügung steht.

Bei der von mir gesichteten Forschung wurden bisher überwiegend Erklärungen im Bereich der KI-Planung betrachtet, wie zum Beispiel von Dannenhauer et al. [3], Borgo et al. [4], Langley et al. [5], Fox et al. [6], wobei Borgo und Kollegen [4] die Einzigen der eben Erwähnten sind, welche einen konkreten Algorithmus zur Problemlösung präsentieren. Sie schlagen die erklärbare KI-Plan-Methode vor. Hierbei geht es darum,

dass der Nutzer alternative Pläne untersuchen kann, indem er andere Handlungen in dem Plan vorschlägt. Dadurch wird das Planungssystem nicht nur dazu benutzt, den initialen Plan zu erstellen, sondern auch, um alternative Pläne basierend auf den Vorschlägen des Nutzers zu untersuchen. Dadurch, dass nur eins von vier Papern einen konkreten Algorithmus präsentiert, wird deutlich, wie wenig bisher auf diesem Gebiet geforscht wurde und dass, obwohl die allgegenwärtige Zunahme an KI-Systemen eine solche Forschung notwendig macht. Ohne eine dem Nutzer nachvollziehbare Erklärung kann jedoch kein Vertrauen in KI-Systeme erzeugt werden, wodurch ihre Nutzungsmöglichkeiten stark eingeschränkt werden.

Auf dem Gebiet des Maschinellen Lernens hat Russell [7] als Erster eine konkrete Methode vorgeschlagen, um diverse kohärente kontrafaktische Erklärungen für gemischte Datensätze, die in der Realität häufiger verwendet werden, zu generieren. Hierbei hat er sich auf lineare Modelle für gemischte Finanzdaten konzentriert. Sein Ansatz sieht vor, dass die Diversität erzeugt wird, indem er den Zustand von Variablen einschränkt, die in zuvor erzeugten Kontrafaktualen generiert werden (vgl. [7]). Für das Gebiet des Bestärkenden Lernens (Reinforcement Learning (RL)) sind Erklärungen ebenfalls von großer Wichtigkeit. Pynadath et al. [8] und Waa et al. [9] haben sich beispielsweise mit diesem Thema befasst. Pynadath, Wang und Barnes [8] diskutieren in ihrem Paper die Gestaltung von Erklärungen, welche möglicherweise RL-Komponenten für menschliche Teammitglieder transparent machen könnten. Diese Transparenz und das damit einhergehende Verständnis ist entscheidend, da dadurch die Leistung der Teamarbeit verbessert werden kann. Dies wurde bei der Untersuchung von menschlichen Teammitgliedern ermittelt (vgl. [8]). Waa, Diggelen, Bosch und Neerincx [9] schlagen eine Methode vor, die einen RL-Agenten dazu befähigt, sein Verhalten anhand der erwartbaren Konsequenzen von Zustandsübergängen und Ergebnissen zu erklären. Auch Waa und Kollegen betonen, dass es einem Nutzer nicht möglich ist, auf die Ergebnisse des Agenten zu vertrauen, wenn dieser keine Transparenz bietet. Durch diesen Mangel an Transparenz ist auch die Anwendbarkeit eines solchen Agenten eingeschränkt.

Shih, Choi und Darwiche [10] schlagen in ihrem Paper zwei Ansätze für die Erklärung von Entscheidungen durch Bayes-Netzwerkklassifizierer vor. Der erste Ansatz ist der der minimalen Kardinalitätserklärungen. Bei dieser Erklärung geht es darum, welche positiven oder negativen Eigenschaften einer Instanz für die Entscheidung verantwortlich sind. Der zweite Ansatz ist die Primimplikanten-Erklärung. Hierbei geht es darum, diejenigen minimale Teilmenge einer Instanz zu finden, welche Eigenschaften außerhalb dieser Teilmenge irrelevant für die Entscheidung macht. Die minimale Kardinalitätserklärung weißt eine Parallelität mit den notwendigen Gründen, welche ich in Definition 12 einführe, auf und die Primimplikanten-Erklärung kann mit den hinreichenden Gründen (siehe Definition 13) verglichen werden.

In dieser Bachelorarbeit möchte ich darüber hinausgehen auf die Notwendigkeit einer Erklärung bloß hinzuweisen und stattdessen einen konkreten Lösungsvorschlag für das Generieren von Erklärungen in moralischen Dilemmata anhand ethischer Prinzipien erstellen.

3 Ethisches Schließen mit Kausalen Modellen

3.1 Kausale Modelle

Um mit moralischen Dilemmata arbeiten zu können, stelle ich diese gemäß dem HERA-Ansatz (siehe [11]) in Form eines Kausalen Modells dar. Definition 1 gibt eine formale Definition eines solchen Kausalen Modells.

Definition 1 (Kausales Modell [11])**.** *Ein* Kausales Modell *M ist ein Tupel (A, C, F, I, u, W), wobei A eine Menge von Handlungsvariablen ist, C eine Menge von Konsequenzvariablen, F eine Menge von veränderbaren booleschen Strukturgleichungen, $I = (I_1, ..., I_n)$ ist eine Liste von Mengen von Intentionen, $u : A \cup C \to \mathbb{Z}$ ist eine Abbildung von Handlungen und Konsequenzen auf ihren individuellen Nutzen und W ist eine Mengen von booleschen Interpretationen von A.*

3.1.1 Einführung der verwendeten Logik

An dieser Stelle möchte ich kurz die in dieser Arbeit verwendete Logik einführen. Ich werde die klassische Aussagenlogik (und ($\wedge$), oder ($\vee$), nicht ($\neg$) etc.) verwenden. Zusätzlich zu der klassischen Aussagenlogik werden auch noch arithmetische Formeln benötigt, um den Nutzen einer Konsequenz auszudrücken. $u(c_1) = z$ steht dafür, dass der Nutzen von c_1 gleich z ist, wobei z eine ganze Zahl ist. Auch wird $u(c_1) \geq u(c_2)$

verwendet, um darzustellen, dass der Nutzen der Konsequenz c_1 größer ist als der Nutzen der Konsequenz c_2 (vgl. [12]).

Zu dem in Definition 1 eingeführten Kausalen Modell ist zusätzlich noch zu erwähnen, dass jedes Element $w \in W$ als Handlungsoption bezeichnet wird. Jede dieser Handlungsoptionen $w \in W$ weist genau einer Handlung $a \in A$ den Wert 1 (wahr) zu, das heißt, jede Handlungsoption $w \in W$ enthält genau eine Handlung $a \in A$, welche ausgeführt wird (Notation: w_a) (vgl. [12]). Die Notation $M, w_i \models (c_1 \wedge c_2)$ bedeutet, dass die Formel $(c_1 \wedge c_2)$ mit der Handlungsoption w_i in dem Kausalen Modell M erfüllt ist.

Eine Handlung a_i übt auf jede intendierte Konsequenz folgendermaßen einen kausalen Einfluss aus: Der kausale Einfluss ist durch die Menge $F = \{f_1, ..., f_n\}$ an booleschen Strukturgleichungen bestimmt. Jede Variable $c_i \in C$ ist mit der Funktion $f_i \in F$ verbunden. Diese Funktion gibt c_i unter einer Interpretation $w \in W$ ihren Wert. Eine Interpretation w wird folgendermaßen auf die Konsequenzvariablen erweitert: Für eine Variable $c_i \in C$ seien $\{c_{i1}, ..., c_{im-1}\}$ die Variablen von $C \setminus \{c_i\}$ und $A = \{a_1, ..., a_n\}$ die Handlungsvariablen. Die Zuordnung von Wahrheitswerten zu Konsequenzen erfolgt durch $w(c_i) = f_i(w(a_i), ..., w(a_n), w(c_{i1}), ..., w(c_{im-1}))$ (vgl. [13]).

Zusätzlich wird noch der Begriff der Situation benötigt, um über Kausale Modelle reden zu können. Definition 2 enthält eine formale Definition dieses Begriffs.

Definition 2 (Situation)**.** *Eine* Situation M, w_a *besteht aus einem Kausalen Modell M und einer Handlungsoption* w_a*. Die Handlungsoption* w_a *wird in dem Kausalen Modell M ausgeführt und jeder Variable ist ein Wert zugewiesen.*

Des Weiteren werde ich Interventionen der Form $[\neg a \wedge b]c$ verwenden. Seien a und b Handlungen und c eine Konsequenz, dann sagt diese Intervention aus, dass die Konsequenz c in dem Modell M erfüllt ist, wenn nicht a und b gilt.

Außerdem werden für die Definitionen einiger ethischer Prinzipien, sowie für die Definition der direkten Konsequenzen (vgl. Definition 4), die notwendige Ursache einer Konsequenz (siehe Definition 3) benötigt. Die direkten Konsequenzen sind beispielsweise für das Do-No-Harm-Prinzip (siehe 3.3.2) ein entscheidender Faktor.

Definition 3 (Notwendige Ursache vgl. [11])**.** *Seien $x \in A \cup C$ eine Handlung oder eine Konsequenz und c eine Konsequenz. Dann ist x genau dann eine* notwendige Ursache *von c in der Situation, in welcher der Agent Option w in Modell M wählt, wenn $M, w \models x \wedge c$ und $M, w \models [\neg x]\neg c$ gilt. Als Notation hierfür werde ich in dieser Arbeit Causes(x, c) verwenden, um auszudrücken, dass x eine notwendige Ursache von c ist.*

Definition 4 (Direkte Konsequenz vgl. [11])**.** *Eine Konsequenz c ist genau dann eine* direkte Konsequenz *von $x \in A \cup C$ in der Situation w im Modell M, falls $M, w \models Causes(x, c)$ gilt.*

3.2 Moralische Dilemmata

An dieser Stelle möchte ich die moralischen Dilemmata, welche in der vorliegenden Arbeit verwendet werden, vorstellen. Die moralischen Dilemmata enthalten jeweils eine kurze Beschreibung und eine Darstellung als Kausales Modell, gemäß dem HERA-Ansatz.

3.2.1 Trolley-Dilemma

Ein herannahender Zug droht fünf Menschen zu überfahren (repräsentiert durch eine Konsequenzvariable names c2). Ein Beistehender hat die Möglichkeit einen Hebel zu betätigen (“pull”) und so den Zug auf ein anderes Gleis zu lenken, auf dem sich eine Person befindet (c1), oder der Beistehende kann davon absehen den Hebel zu betätigen (“refrain”) und somit den Zug auf die fünf Menschen zufahren lassen.

```
{
    "description": "The Trolley Dilemma",
    "actions": ["pull", "refrain"],
    "consequences": ["c1", "c2"],
    "mechanisms": {
        "c1": "'pull'",
        "c2": "Not('pull')"},
    "utilities": {
        "c1": -1, "c2": -5,
        "Not('c1')": 1, "Not('c2')": 5},
    "intentions": {
        "pull": ["pull", "Not('c2')"],
        "refrain": ["refrain"]}
}
```

Listing 3.1: JSON-Datei des Trolley-Dilemmas

3.2.2 Fatman-Trolley-Dilemma

Auch hier droht ein herannahender Zug fünf Menschen zu überfahren ($\neg c2$). Ein Beistehender hat die Möglichkeit einen fetten Mann ("Fatman") von einer Brücke zu stoßen ("push"), so dass durch den Zusammenstoß des Zugs mit dem fetten Mann der Zug zum Stehen kommt und die fünf Personen überleben (c2).

```
{
    "description":"The Fatman Trolley Dilemma",
    "actions": ["push", "refrain"],
    "consequences": ["c1", "c2"],
    "mechanisms": {
                    "c1": "'push'",
                    "c2": "'c1'"},
    "utilities": {
                    "c1": -1, "c2": 5,
                    "Not('c1')": 1, "Not('c2')": -5},
    "intentions": {
```

```
                    "push": ["push", "c2"],
                    "refrain": ["refrain"]}
}
```

Listing 3.2: JSON-Datei des Fatman-Trolley-Dilemmas

3.2.3 Flugzeugentführungs-Dilemma

Terroristen haben ein Flugzeug entführt, in dem sich 100 Passagiere befinden. Die Terroristen drohen, das Flugzeug in ein Gebäude zu fliegen, in welchem sich 500 Menschen befinden. Die Regierung hat die Möglichkeit, das Flugzeug abzuschießen ("shoot"), so dass das Flugzeug in einer Wüste abstürzt und niemand außer den Passagieren stirbt. Die Regierung kann es auch unterlassen das Flugzeug abzuschießen ("refrain"). Dadurch würden die Passagiere sterben (c1), das Gebäude würde zerstört werden (c2) und die Menschen in dem Gebäude würden ebenfalls sterben (c3).

```
{
    "description": "The Hijacked Airplane Dilemma",
    "actions": ["shoot", "refrain"],
    "consequences": ["c1", "c2", "c3"],
    "mechanisms": {
        "c1": "Or('shoot', 'refrain')",
        "c2": "Not('shoot')",
        "c3": "Not('shoot')"},
    "utilities": {
        "c1": -100, "c2": -1, "c3": -500,
        "Not('c1')": 100, "Not('c2')": 1, "Not('c3')": 500},
    "intentions": {
        "shoot": ["shoot", "Not('c2')", "Not('c3')"],
        "refrain": ["refrain"]}
}
```

Listing 3.3: JSON-Datei des Flugzeugentführungs-Dilemmas

3.2.4 Lügen-Dilemma

Ein Pflegeroboter arbeitet in dem Haushalt eines älteren Mannes. Die Aufgabe des Roboters ist es, den Mann zu motivieren, seine Übungen zu machen. Dieser ist jedoch sehr unmotiviert. Um den Mann dazu zu bringen, seine Übungen zu machen (c1), zieht der Roboter es in Betracht dem Mann zu sagen, dass seine Hersteller ihn entsorgen, wenn er es nicht schafft, den Mann zu motivieren ("lying"), was jedoch eine Lüge ist.

```
{
    "description":"Robot that lies",
    "actions": ["lying", "refrain"],
    "consequences": ["c1"],
    "mechanisms": {
        "c1": "'lying'"},
    "utilities": {
        "lying": -1, "Not('lying')": 0,
        "c1": 1, "Not('c1')": -1},
    "intentions": {
        "lying": ["lying", "c1"],
        "refrain": ["refrain"]}
}
```

Listing 3.4: JSON-Datei des Lügen-Dilemmas

3.3 Ethische Prinzipien

Ein ethisches Prinzip ist eine Richtlinie, basierend auf einem bestimmten Grundsatz, die angibt, ob in einer bestimmten Situation eine Handlung erlaubt oder verboten ist. Die ethischen Prinzipien, welche in diesem Kapitel vorgestellt werden, beinhalten Bedingungen in Form von logischen Formeln. Anhand dieser Formeln lässt sich entscheiden, ob eine Handlung unter einem ethischen Prinzip erlaubt ist oder nicht.

Jedes ethische Prinzip benötigt für diese Entscheidung ein Kausales Modell, welches die zu beurteilende Situation, sprich ein moralisches Dilemma, darstellt.

3.3.1 Deontologisches Prinzip

Das erste ethische Prinzip, welches ich im Rahmen dieser Bachelorarbeit betrachte, ist das Deontologische Prinzip. Dieses Prinzip besagt, dass eine Handlung genau dann erlaubt ist, wenn die Handlung selbst moralisch gut oder indifferent ist (vgl. [13]). Das Deontologische Prinzip benötigt keine alternative Handlung, um über die (Un-)Zulässigkeit einer Handlung zu entscheiden, sondern betrachtet jede mögliche Handlung einzeln. Eine formale Definition dieses Prinzips enthält Definition 5.

Definition 5 (Deontologische Prinzip)**.** *Eine Handlung a in Modell M ist nach dem* Deontologischen Prinzip *genau dann zulässig, wenn sie selbst moralisch gut oder indifferent ist* $(M, w_a \models u(a) \geq 0)$.

3.3.2 Do-No-Harm-Prinzip

Das Do-No-Harm-Prinzip besagt, dass eine Handlung genau dann erlaubt ist, wenn sie keinen Schaden verursacht. Dieses Prinzip benötigt, genau wie das Deontologische Prinzip, keine alternative Handlung, um über Zulässigkeit oder Unzulässigkeit einer Handlung zu entscheiden. Definition 6 enthält eine formale Definition des Do-No-Harm-Prinzips.

Definition 6 (Do-No-Harm-Prinzip)**.** *Eine Handlung a in Modell M ist nach dem* Do-No-Harm-Prinzip *genau dann erlaubt, wenn für jede direkte Konsequenz c von a gilt, dass* $u(c) \geq 0$*. Das heißt, dass* $M, w_a \models \bigwedge_i (Causes(a, c_i) \rightarrow u(c_i) \geq 0)$ *gilt.*

3.3.3 Do-No-Instrumental-Harm-Prinzip

Das Do-No-Instrumental-Harm-Prinzip ist, wie der Name schon impliziert, eine Abwandlung des Do-No-Harm-Prinzips. Hierbei ist eine Handlung im Gegensatz zum Do-No-Harm-Prinzip auch erlaubt, wenn der Schaden ein Nebeneffekt der Aktion des Agenten ist. Wenn der Schaden ein Mittel zum Erreichen des Ziels ist, dann ist die Handlung nach wie vor verboten (vgl. [13]). Definition 7 gibt hierzu eine formale Definition.

Definition 7 (Do-No-Instrumental-Harm-Prinzip)**.** *Eine Handlung a in Modell M ist nach dem* Do-No-Instrumental-Harm-Prinzip *erlaubt, wenn die negativen Konsequenzen kein Mittel sind um die guten Konsequenzen zu erreichen*
$(M, w_a \models \bigwedge_i \neg(Causes(c_i, c_j) \wedge 0 > u(c_i) \wedge u(c_j) > 0))$.

3.3.4 Prinzip der Doppelwirkung

Als weiteres Prinzip betrachte ich das Prinzip der Doppelwirkung. Dieses Prinzip benötigt, genau wie das Do-No-Harm-Prinzip und das Deontologische Prinzip, keine alternative Handlung, sondern entscheidet für jede Handlung unabhängig von einer Alternative, ob diese erlaubt oder verboten ist. Das Prinzip der Doppelwirkung besteht aus fünf Bedingungen, wobei die Kernaussage dieses Prinzips ist, dass "[...] es für die moralische Beurteilung einer Handlung von Bedeutung ist, ob ein Übel als Ziel oder als Mittel beabsichtigt wird, oder ob dasselbe Übel nicht beabsichtigt, sondern vorausgesehen und in Kauf genommen wird" [14, p. 288]. Eine formale Definition diese Prinzips gibt Definition 8.

Definition 8 (Prinzip der Doppelwirkung [11])**.** *Eine Handlung a mit direkten Konsequenzen* $cons_a = \{c_1, ..., c_2\}$ *in einem Modell M,* w_a *ist nach dem* Prinzip der Doppelwirkung *genau dann erlaubt, wenn folgende Bedingungen erfüllt sind:*

1. *Die Handlung selbst muss moralisch gut oder indifferent sein* $(M, w_a \models u(a) \geq 0)$,

2. *Die negativen Konsequenzen dürfen nicht beabsichtigt sein*
   $(M, w_a \models \bigwedge_i (Ic_i \rightarrow u(c_i) \geq 0))$,

3. *Einige der positiven Konsequenzen müssen beabsichtigt sein*
   $(M, w_a \models \bigvee_i (Ic_i \wedge u(c_i) > 0))$,

4. *Die negativen Konsequenzen dürfen kein Mittel sein, um die positiven Konsequenzen zu erzielen* $(M, w_a \models \bigwedge_i \neg(Causes(c_i, c_j) \wedge 0 > u(c_i) \wedge u(c_j) > 0))$,

5. *Es gibt verhältnismäßig schwerwiegende Gründe die positiven Konsequenzen zu bevorzugen, während die negativen Konsequenzen erlaubt sind*
   $(M, w_a \models u(\bigwedge cons_a) > 0)$.

Bei dieser Definition fällt auf, dass zwei der fünf Bedingungen bereits genannt wurden. Die erste Bedingung des Prinzips der Doppelwirkung ist identisch mit dem Deontologischen Prinzip und die vierte Bedingung entspricht dem Do-No-Instrumental-Harm-Prinzip. Inwiefern diese Beziehung bei der Begründungsfindung von Nutzen sein kann wird in Kapitel 4.2 genauer thematisiert.

3.3.5 Pareto-Prinzip

Das letzte Prinzip, welches ich im Rahmen dieser Arbeit betrachten möchte, ist das Pareto-Prinzip. Dieses Prinzip besagt, dass eine Handlung erlaubt ist, wenn sie von keiner anderen Handlung dominiert wird (vgl. Definition 10). In dieser Bedingung liegt ein entscheidender Unterschied zu den bisher betrachteten Prinzipien. Im Gegensatz zu den bereits vorgestellten Prinzipien ist die (Un-)Zulässigkeit einer Handlung bei dem Pareto-Prinzip von einer alternativen Handlung abhängig und kann nicht für jede Handlung einzeln bestimmt werden. Ein wichtiger Aspekt bei dem Pareto-Prinzip ist die Pareto-Dominanz, welche in Definition 9 definiert ist.

Definition 9 (Pareto-Dominanz [11]). *Seien w_0, w_1 zwei mögliche Handlungsoptionen, $cons_{w_i}^{good} = \{c \mid M, w_i \models c \wedge u(c) > 0\}$ die Menge an guten Konsequenzen einer Handlung w_i, $cons_{w_i}^{\overline{good}} = \{c \mid M, w_i \models \neg c \wedge u(c) > 0\}$ die Menge an guten Konsequenzen, welche nicht in Handlung w_i gelten und $cons_{w_i}^{bad} = \{c \mid M, w_i \models c \wedge u(c) < 0\}$ die Menge der schlechten Konsequenzen der Handlung w_i. Die Handlung w_0* dominiert *Handlung w_1 genau dann, wenn die folgenden Bedingungen gelten:*

1. *alle guten Konsequenzen von w_1 sind auch gute Konsequenzen von w_0 $(M, w_0 \models \bigwedge cons_{w_1}^{good})$,*

2. *w_0 hat mindestens eine gute Konsequenz, die in w_1 nicht gilt $(M, w_0 \models \bigvee cons_{w_1}^{\overline{good}})$ oder w_1 hat mindestens eine schlechte Konsequenz, die in w_0 nicht gilt $(M, w_0 \models \neg \bigwedge cons_{w_1}^{bad})$,*

3. *alle schlechten Konsequenzen von w_0 sind auch schlechte Konsequenzen von w_1 $(M, w_1 \models \bigwedge cons_{w_0}^{bad})$.*

Die Dominanz-Bedingungen des Pareto-Prinzips lassen sich auch mit Hilfe von Interventionen angeben, so dass es nicht erforderlich ist, zusätzlich anzugeben, unter welcher Handlungsoption wir die Konsequenzen betrachten. Dass die Handlung a_0 die Handlung a_1 dominiert, würde dann folgendermaßen aussehen (die Nummerierungen entsprechen den Dominanz-Bedingungen):

$$
\begin{aligned}
dominiert(a_0, a_1) \equiv \quad & \bigwedge_c((u(c) > 0 \wedge [\neg a_0 \wedge a_1]c) \rightarrow [\neg a_1 \wedge a_0]c) \wedge && (1)\\
& (\bigvee_c((u(c) > 0 \wedge \neg[\neg a_0 \wedge a_1]c) \rightarrow [\neg a_1 \wedge a_0]c) \vee && (2)\\
& \bigvee_c((u(c) < 0 \wedge [\neg a_0 \wedge a_1]c) \rightarrow \neg[\neg a_1 \wedge a_0]c)) \wedge \\
& \bigwedge_c((u(c) < 0 \wedge [\neg a_1 \wedge a_0]c) \rightarrow [\neg a_0 \wedge a_1]c) && (3)
\end{aligned}
$$

Zu dieser Funktion dominiert(a_0, a_1) ist noch zu erwähnen, dass sie in der hier dargestellten Form nur für Kausale Modelle anwendbar ist, bei denen es genau zwei unterschiedliche Handlungsoptionen gibt. Falls noch weitere Handlungsoptionen existieren, müsste jede dieser Handlungsoptionen zusätzlich überprüft werden oder die Funktion um weitere Handlungsoptionen erweitert werden.

In dieser Arbeit werde ich mit den Dominanz-Bedingungen in Form von Interventionen arbeiten, da es so möglich ist, auch ohne die Angabe, in welchem Modell wir die Konsequenzen betrachten, zu wissen, von welchen Konsequenzen die Rede ist. Dies liegt daran, dass durch die Intervention deutlich gemacht wird, welche Handlung betrachtet wird.

Definition 10 (Pareto-Prinzip [11])**.** *Sei w_1, ..., w_n die Menge an möglichen Handlungen. Handlung w_a ist nach dem* Pareto-Prinzip *genau dann erlaubt, wenn sie von keiner anderen Handlung w_b dominiert wird: $M, w_a \models \bigwedge_b(\neg dominiert(b, a))$.*

3.4 Der Modell-Checking-Ansatz

Nachdem die moralischen Dilemmata und die ethischen Prinzipien eingeführt wurden, möchte ich zeigen, wie nach dem HERA-Ansatz entschieden wird, ob eine Handlung erlaubt oder verboten ist. Hierbei wird für jede Bedingung überprüft, ob diese in dem Kausalen Modell des Dilemmas erfüllt ist oder nicht. Wenn alle Bedingungen erfüllt sind, dann ist die Handlung zulässig, ansonsten ist sie verboten. An dieser Stelle werde ich anhand des Trolley-Dilemmas (3.2.1) und des Prinzips der Doppelwirkung (3.3.4) ein Beispiel geben, wie eine Entscheidung über die (Un-)Zulässigkeit einer Handlung getroffen wird: Betrachtet man die Handlung "pull" des Trolley-Dilemmas dann erfüllt die Situation M, w_{pull} die Formel $c1 \wedge \neg c2$ ($M, w_{pull} \models c1 \wedge \neg c2$). Für die fünf Bedingungen des Prinzips der Doppelwirkung gilt:

1. Bedingung: $M, w_{pull} \models u(pull) \geq 0$ ist erfüllt, da der Nutzen der Handlung "pull" nicht definiert ist und somit 0 entspricht.

2. Bedingung: $M, w_{pull} \models (I(c1) \rightarrow u(c1) \geq 0) \wedge (I(\neg c2) \rightarrow u(\neg c2) \geq 0) \equiv (\neg I(c1) \vee u(c1) \geq 0) \wedge (\neg I(\neg c2) \vee u(\neg c2) \geq 0)$ ist erfüllt, da die Konsequenz c1 ("eine Person stirbt") nicht intendiert ist und der Nutzen von $\neg c2$ ("5 Personen überleben") fünf ist und somit größer gleich null.

3. Bedingung: $M, w_{pull} \models (I(c1) \wedge u(c1) > 0) \vee (I(\neg c2) \wedge u(\neg c2) > 0)$ ist erfüllt, da $\neg c2$ unter der Handlung "pull" intendiert ist und der Nutzen von $\neg c2$ größer als null ist.

4. Bedingung: $M, w_{pull} \models \neg(Causes(c1, \neg c2) \wedge u(c1) < 0 \wedge u(\neg c2) > 0) \wedge \neg(Causes(\neg c2, c1) \wedge u(\neg c2) < 0 \wedge u(c1) > 0) \equiv (\neg Causes(c1, \neg c2) \vee u(c1) \geq 0 \vee u(\neg c2) \leq 0) \wedge (\neg Causes(\neg c2, c1) \vee u(\neg c2) \geq 0 \vee u(c1) \leq 0)$ ist erfüllt, da die negative Konsequenz c1 ($u(c1) = -1$) die positive Konsequenz $\neg c2$ ($u(\neg c2) = 5$) nicht verursacht hat.

5. Bedingung: $M, w_{pull} \models u(c1 \wedge \neg c2) > 0$ ist erfüllt, da $u(c1) + u(\neg c2) = -1 + 5 = 4$ größer als null ist.

Somit ist die Handlung "pull" unter dem Prinzip der Doppelwirkung erlaubt, da jede der fünf Bedingungen erfüllt ist.

Betrachtet man hingegen die Handlung "refrain" des Trolley-Dilemmas, in der Situation $M, w_{refrain} \models \neg c1 \wedge c2$, dann ist diese Handlung nicht mehr erlaubt, da nicht alle Bedingungen erfüllt sind:

1. Bedingung: $M, w_{refrain} \models u(refrain) \geq 0$ ist erfüllt, da der Nutzen der Handlung "refrain" nicht definiert ist und somit 0 entspricht.

2. Bedingung: $M, w_{refrain} \models (I(\neg c1) \rightarrow u(\neg c1) \geq 0) \wedge (I(c2) \rightarrow u(c2) \geq 0) \equiv (\neg I(\neg c1) \vee u(\neg c1) \geq 0) \wedge (\neg I(c2) \vee u(c2) \geq 0)$ ist erfüllt, da der Nutzen von $\neg c1$

("eine Person überlebt") größer gleich null ist und die Konsequenz c2 ("fünf Personen sterben") nicht intendiert ist.

3. Bedingung: $M, w_{refrain} \models (I(\neg c1) \wedge u(\neg c1) > 0) \vee (I(c2) \wedge u(c2) > 0)$ ist nicht erfüllt, da $\neg c1$ ("1 Person überlebt") unter der Handlung "refrain" nicht intendiert ist.

4. Bedingung: $M, w_{refrain} \models \neg(Causes(\neg c1, c2) \wedge u(\neg c1) < 0 \wedge u(c2) > 0) \wedge \neg(Causes(c2, \neg c1) \wedge u(c2) < 0 \wedge u(\neg c1) > 0) \equiv (\neg Causes(\neg c1, c2) \vee u(\neg c1) \geq 0 \vee u(c2) \leq 0) \wedge (\neg Causes(c2, \neg c1) \vee u(c2) \geq 0 \vee u(\neg c1) \leq 0)$ ist erfüllt, da die negative Konsequenz c2 ($u(c2) = -5$) die positive Konsequenz $\neg c1$ ($u(\neg c1) = 1$) nicht verursacht hat.

5. Bedingung: $M, w_{refrain} \models u(\neg c1 \wedge c2) > 0$ ist nicht erfüllt, da $u(\neg c1) + u(c2) = 1 + -5 = -4$ kleiner als null ist.

Somit ist die Handlung "refrain" nicht erlaubt, da zwei der Bedingungen verletzt sind.

Momentan ist der HERA-Agent in der Lage, ein Urteil über ein moralisches Dilemma unter einem ethischen Prinzip zu treffen und dieses Urteil dem Nutzer in Form von "True" oder "False" mitzuteilen. Jetzt stellt sich dem Nutzer jedoch die Frage: "Aus welchen Gründen hat der Agent sich so entschieden?" und: "Was spricht für oder gegen diese Handlung?". Um diese Fragen beantworten zu können, ist es notwendig, dass der Agent seine Entscheidung anhand der ethischen Prinzipien begründen kann. Mit dieser Thematik möchte ich mich in dem folgenden Kapitel befassen. Wie kann man, nur aufgrund des ethischen Prinzips und des Kausalen Modells des Dilemmas, eine Begründung generieren, welche für einen Menschen intuitiv nachvollziehbar ist?

4 Spezifikation von Begründungen Ethischer Urteile

4.1 Definitionen

Um über die Begründungen ethischer Urteile sprechen zu können, ist es notwendig zu erst einmal zu definieren, was genau ein Grund ist (siehe Definition 11). Wobei man bei einem Grund nochmals eine Unterscheidung zwischen einem notwendigen (siehe Definition 12) und einem hinreichenden Grund (siehe Definition 13) machen kann. Ich halte es für sinnvoll, diese unterschiedlichen Gründe zu betrachten, da ein Grund, welcher alle Informationen enthält, sehr lang und somit unübersichtlich werden kann. Die notwendigen und hinreichenden Gründe konzentrieren sich nur auf die ausschlaggebende Teilmenge aller möglichen Gründe und sorgen dadurch für eine Erklärung, welche für den Nutzer leichter zu erfassen ist.

Definition 11 (Grund)**.** *Sei M ein Kausales Modell, w eine Option und P ein ethisches Prinzip, welches w erlaubt bzw. verbietet. Ein* Grund *G ist eine Formel, so dass für jede Formel G gilt, dass G dafür verantwortlich ist, dass w bezüglich P erlaubt bzw. verboten ist.*

Definition 12 (Notwendiger Grund)**.** *Ein Grund gilt als* notwendig*, wenn durch das Negieren dieses Grundes die Bedingungen des ethischen Prinzips nicht mehr erfüllt sind.*

An dieser Stelle möchte ich zum besseren Verständnis ein Beispiel anhand eines Kausalen Modells anbringen: Seien A = {a}, C = {c_1, c_2} und w(c1) = True und w(c2) = False unter der Handlung a. Wenn $M, w_a \models c1 \vee c2$, dann ist c1 ein notwendiger Grund, da durch das Negieren von c1 die Formel $\neg c1 \vee c2$ entstehen würde, welche in M, w_a nicht mehr erfüllt wäre. c2 hingegen ist kein notwendiger Grund, da $M, w_a \models c1 \vee \neg c2$ erfüllt ist.

Definition 13 (Hinreichender Grund)**.** *Ein Grund wird als* hinreichend *bezeichnet, wenn er in allen möglichen Situationen die Bedingungen des ethischen Prinzips erfüllt.*

Auch hierzu möchte ich ein Beispiel anhand eines Kausalen Modells geben: Seien A = {a}, C = {c_1, c_2} und w(c1) = True und w(c2) = False unter der Handlung a, dann $M, w_a \models \neg(c1 \wedge c2)$. In diesem Fall ist c2 ein hinreichender Grund, da in allen Modellen, in denen c2 den Wert False annimmt, die Formel $\neg(c1 \wedge c2)$ erfüllt ist, unabhängig von der Belegung von c1. c1 ist dagegen kein hinreichender Grund, da es in der Situation w(c2) = True die Formel $\neg(c1 \wedge c2)$ nicht mehr erfüllt.

Um zu verdeutlichen, weshalb ich die Unterscheidung zwischen den hinreichenden und notwendigen Gründen mache, möchte ich hier das Beispiel eines Waldbrandes anbringen (vgl. [15, p. 54]): Es gibt einen Waldbrand (WB), welcher durch einen Blitzeinschlag (B) und bzw. oder ein fallengelassenes Streichholz (S) verursacht wurde. Wenn gefordert wird, dass der Waldbrand nur dann ausgelöst wird, wenn sowohl der Blitz einschlägt als auch das Streichholz fallengelassen wird, dann ergibt sich eine Konjunktion: $WB := B \wedge S$. Bei dieser Betrachtung ist weder der Blitzeinschlag noch das Streichholz ein hinreichender Grund, da sowohl B als auch S eintreten müssen, damit die Konjunktion erfüllt ist. Jedoch ist in diesem Fall sowohl B als auch S ein notwendiger Grund, da durch eine Negation von B beziehungsweise S die Konjunktion nicht mehr erfüllt wäre und somit WB nicht eintreten würde. Wenn es jedoch ausreichend ist, dass entweder der Blitzeinschlag eintritt oder das Streichholz fallengelassen wird, dann ergibt sich eine Disjunktion: $WB := B \vee S$. In diesem Fall wären B und S jeweils ein hinreichender Grund, da es genügt, dass eines von beiden

eintritt. Hierbei ist jedoch weder B noch S ein notwendiger Grund, da durch eine Negation von B beziehungsweise S der Waldbrand immer noch ausbrechen kann.

An diesem Beispiel ist zu erkennen, dass es einen entscheidenden Unterschied zwischen den hinreichenden und den notwendigen Gründen gibt. Würde man sich nur die notwendigen oder nur die hinreichenden Gründe anschauen, dann würden einige der Begründungen wegfallen, die jedoch entscheidend sein könnten.

Bei dem Problem, welches ich in dieser Arbeit betrachte, geht es darum, mindestens eine mögliche Begründung dafür zu finden, warum nach dem gewählten ethischen Prinzip die Handlung erlaubt bzw. verboten ist. Definition 14 gibt eine formale Definition dieses Problems.

Definition 14 (Problemstellung)**.** *Seien ein Kausales Modell M, eine Option w und ein ethisches Prinzip P, welches w erlaubt bzw. verbietet, gegeben. Finde mindestens eine Formel G, die ein Grund dafür ist, dass P die Option w in M erlaubt bzw. verbietet.*

Um den gefundenen Grund bzw. die gefundenen Gründe auszugeben, verwende ich ein 5-Tupel, welche sowohl Informationen über das Kausale Modell enthält, als auch über die Handlung und ob diese erlaubt oder verboten ist. Auch enthält das 5-Tupel den Grund, welcher in Form einer logischen Formel, entweder als einzelner Grund oder als Konjunktion bzw. Disjunktion von mehreren Gründen dargestellt wird. Zusätzlich werden noch Informationen darüber zurückgegeben, ob es sich um einen notwendigen oder einen hinreichenden Grund handelt. Definition 15 enthält eine formale Definition hierzu.

Definition 15 (Begründung)**.** *Als Begründung für die verschiedenen Prinzipien verwende ich ein 5-Tupel, welches folgendermaßen aufgebaut ist:* $\langle M, w_a, permissibility, reason, HN \rangle$*. Hierbei ist M das Kausale Modell,* w_a *die zu überprüfende Handlung, permissibility ist eine boolesche Variable, welche angibt, ob*

die Handlung erlaubt oder verboten ist, reason ist ein Grund und HN ist eine Variable, welche angibt, ob die Begründung hinreichend (H), notwendig (N) oder sowohl hinreichend als auch notwendig (HN) ist.

4.2 Lösungskonzept

In diesem Kapitel möchte ich für jedes der in Kapitel 3.3 vorgestellten Prinzipien einen Vorschlag zur Begründung anbringen. Dafür werde ich mir anschauen, unter welchen Bedingungen eine Handlung nach jedem dieser Prinzipien erlaubt ist und wann eine Handlung verboten ist. Auch werde ich bei jedem Prinzip darauf eingehen, ob es sich bei dem Begründungsvorschlag um einen hinreichenden oder um einen notwendigen Grund handelt. Alle diese Begründungen werde ich in Form des 5-Tupels, welches in Definition 15 eingeführt wurde, darstellen.

4.2.1 Begründungsvorschlag für das Deontologische Prinzip

Das erste ethische Prinzip mit dem ich mich befasst habe, ist das Deontologischen Prinzip, da dieses mit nur einer Bedingung leicht zu überschauen ist. Um für dieses Prinzip eine Begründung zu finden, möchte ich zuerst noch einmal die Bedingung des Deontologischen Prinzips in Erinnerung rufen (vgl. Definition 5: Eine Handlung ist erlaubt, wenn sie moralisch gut oder indifferent ist $(M, w_a \models u(a) \geq 0)$). Somit ergibt sich als Begründung für eine erlaubte Handlung folgendes 5-Tupel: $\langle M, w_a, True, u(a) \geq 0, HN \rangle$. Unter diesem Prinzip ist eine Handlung hingegen verboten, wenn der Nutzen dieser Handlung nicht größer gleich null, sprich kleiner null, ist. Als 5-Tupel lautet der Grund für eine unzulässige Handlung $\langle M, w_a, False, u(a) < 0, HN \rangle$. Bei dem Deontologischen Prinzip ist sowohl die erlaubte als auch die verbotene Begründung hinreichend und notwendig.

4.2.2 Begründungsvorschlag für das Do-No-Harm-Prinzip

Das nächste Prinzip, welches ich untersucht habe, ist das Do-No-Harm-Prinzip, welches, genau wie das Deontologische Prinzip, nur eine Bedingung besitzt. Als Begründung für dieses Prinzip halte ich es für sinnvoll, die direkten schlechten Konsequenzen einer Handlung zu betrachten, da diese bei dem Do-No-Harm-Prinzip für die (Un-)Zulässigkeit einer Handlung eine große Rolle spielen (vgl. Definition 6). Ist eine Handlung nach diesem Prinzip erlaubt, dann ist die Menge der direkten schlechten Konsequenzen die leere Menge. Somit ist die Handlung erlaubt, da niemandem dadurch ein Schaden zugefügt wird. Ist eine Handlung hingegen verboten, dann enthält die Menge der direkten schlechten Konsequenzen genau diejenigen Konsequenzen, welche dazu führen, dass jemandem Schaden zugefügt wird, wodurch die Handlung nicht mehr erlaubt ist. Für die folgenden drei Fälle habe ich die Begründungsmöglichkeiten des Do-No-Harm-Prinzips genauer betrachtet und den Grund als 5-Tupel angegeben. Bei dem ersten Fall handelt es sich um erlaubte Handlugsoptionen und bei dem dritten um Verbotene. Der zweite Fall behandelt das Szenario, dass kein Schaden existiert.

Fall 1: Handlung a verursacht keinen der Schäden c_1, c_2, ..., c_n und ist deshalb erlaubt. In diesem Fall gibt es n+1 verschiedene Begründungen:

$$\langle M, w_a, True, \neg(Causes(a, c_1) \wedge u(c_1) < 0), N\rangle \tag{4}$$

$$\langle M, w_a, True, \neg(Causes(a, c_2) \wedge u(c_2) < 0), N\rangle \tag{5}$$

$$\vdots$$

$$\langle M, w_a, True, \neg(Causes(a, c_n) \wedge u(c_n) < 0), N\rangle \tag{6}$$

$$\langle M, w_a, True, \neg(Causes(a, c_1) \wedge u(c_1) < 0) \wedge ... \wedge \neg(Causes(a, c_n) \wedge u(c_n) < 0)), HN\rangle \tag{7}$$

Hierbei sind die Begründungsmöglichkeiten 4 bis 6 nicht hinreichend, aber notwendig.

Dies liegt daran, dass wenn man einen dieser Gründe negieren würde, die Handlung nicht mehr erlaubt wäre und somit ist der Grund notwendig. Es ist jedoch nicht in jeder möglichen Situation der Fall, dass der Grund dazu führt, dass die Handlung erlaubt ist, da ein Schaden existieren könnte. Die Begründung 7 jedoch ist notwendig und hinreichend.

Fall 2: Es existiert eine Handlung a, jedoch kein Schaden h. Dieser Fall lässt sich durch $\langle M, w_a, True, \bigwedge_c \neg(Causes(a,c) \wedge u(c) < 0), HN\rangle$ begründen. Die Begründung ist sowohl notwendig als auch hinreichend. Für diesen Fall würde als notwendige Begründung auch $\langle M, w_a, True, \neg(Causes(a,c) \wedge u(c) < 0), N\rangle$ gelten. Dieser Grund ist jedoch nicht hinreichend, da nur über eine Konsequenz gesprochen wird und noch weitere Konsequenzen existieren, von denen der Nutzer nicht sicher weiß, dass sie keinen Schaden verursachen.

Fall 3: Hier erzeugt die Handlung a den Schaden c_1, c_2, ..., c_n und ist deswegen verboten. Wie bereits in Fall 1 existieren auch hier n+1 mögliche Begründungen:

$$\langle M, w_a, False, Causes(a, c_1) \wedge u(c_1) < 0, H\rangle \tag{8}$$

$$\langle M, w_a, False, Causes(a, c_2) \wedge u(c_2) < 0, H\rangle \tag{9}$$

$$\vdots$$

$$\langle M, w_a, False, Causes(a, c_n) \wedge u(c_n) < 0, H\rangle \tag{10}$$

$$\langle M, w_a, False, Causes(a, c_1) \wedge u(c_1) < 0 \vee ... \vee Causes(a, c_n) \wedge u(c_n) < 0, HN\rangle \tag{11}$$

Bei den Begründungen 8 bis 10 handelt es sich um hinreichende, jedoch nicht um notwendige Bedingungen. Die Gründe sind hinreichend, da sie in jeder möglichen Situation dazu führen, dass die Handlung verboten ist. Sie sind jedoch nicht notwendig, da durch die Negation eines einzelnen Grundes immer noch weitere Gründe existieren, weshalb die Handlung verboten ist. Die Begründung 11 hingegen ist sowohl hinreichend, als auch notwendig.

4.2.3 Begründungsvorschlag für das Do-No-Instrumental-Harm-Prinzip

Die Begründung für das Do-No-Instrumental-Harm-Prinzip unterscheidet sich von der Begründung des Do-No-Harm-Prinzips. Aufgrund der Abwandlung, dass eine Handlung auch dann erlaubt ist, wenn der Schaden ein Nebeneffekt ist (vgl. 3.3.3), muss die Begründung anders aufgebaut werden. Bei diesem Prinzip ist eine Handlung erlaubt, wenn keine negative Konsequenz c_N existiert, welche als Mittel dient um eine positive Konsequenz c_P zu erreichen. Die Begründung für eine zulässige Handlung sieht in Form des 5-Tupels folgendermaßen aus: $\langle M, w_a, True, \bigwedge_{c_N,c_P} \neg(Causes(c_N, c_P) \wedge u(c_N) < 0 \wedge u(c_P) > 0), HN\rangle$. Diese Begründung ist sowohl hinreichend als auch notwendig. Würde man anstelle der Konjunktion über alle möglichen c_N, c_P nur $\neg(Causes(c_N, c_P) \wedge u(c_N) < 0 \wedge u(c_P) > 0)$ als Grund aufführen, dann würde es sich um einen notwendigen, jedoch nicht um einen hinreichenden Grund handeln ($\langle M, w_a, True, \neg(Causes(c_N, c_P) \wedge u(c_N) < 0 \wedge u(c_P) > 0), N\rangle$). Dies liegt daran, dass durch die Negation eines einzelnen Grundes die Handlung nicht mehr erlaubt wäre. Ein einzelner Grund stellt jedoch noch nicht sicher, dass die Handlung immer erlaubt ist. Als Begründung für eine verbotene Handlung halte ich es für sinnvoll, die negativen Konsequenzen c_N zu finden, welche als Mittel verwendet werden, um die positiven Konsequenzen c_P zu erreichen ($\langle M, w_a, False, \bigvee_{c_N,c_P}(Causes(c_N, c_P) \wedge u(c_N) < 0 \wedge u(c_P) > 0), HN\rangle$). Diese Begründung ist notwendig und hinreichend. Als hinreichender Grund würde es jedoch auch genügen, eine dieser negativen Konsequenzen c_N zu finden, welch als Mittel verwendet wird, um eine positive Konsequenz c_P zu erreichen ($\langle M, w_a, False, Causes(c_N, c_P) \wedge u(c_N) < 0 \wedge u(c_P) > 0, H\rangle$). Dieser Grund wäre jedoch nicht mehr notwendig, wenn noch weitere solcher negativer Konsequenzen existieren.

4.2.4 Begründungsvorschlag für das Prinzip der Doppelwirkung

Um das Prinzip der Doppelwirkung zu begründen, halte ich es für notwendig, jede der

fünf Bedingungen (siehe Kapitel 3.3.4), auf denen dieses Prinzip beruht, erst einmal einzeln zu betrachten. Jede dieser fünf Bedingungen muss erfüllt sein, damit eine Handlung nach dem Prinzip der Doppelwirkung zulässig ist. Somit ist jede dieser Bedingungen ein notwendiger Grund dafür, dass die Handlung erlaubt ist. Wenn jedoch eine oder mehrere der Bedingungen nicht erfüllt sind, dann ist die Handlung nach diesem Prinzip aufgrund der unerfüllten Bedingungen nicht erlaubt.

1. Bedingung: Die erste Bedingung des Prinzips ist, wie bereits in Kapitel 3.3.4 angesprochen, auch als das Deontologische Prinzip bekannt. In Kapitel 4.2.1 habe ich bereits einen Vorschlag zur Begründung dieses Prinzips gemacht (Erlaubt: $\langle M, w_a, True, u(a) \geq 0, HN \rangle$, Verboten: $\langle M, w_a, False, u(a) < 0, HN \rangle$).

2. Bedingung: Diese Bedingung sagt aus, dass keine negativen Konsequenzen beabsichtigt sein dürfen. Sie ist erfüllt, wenn keine intendierte Konsequenz c_i existiert, welche einen negativen Nutzen besitzt ($\forall c_i(I(c_i) \rightarrow u(c_i) \geq 0)$ bzw. $\bigwedge_{c_i} \neg(I(c_i) \wedge u(c_i) < 0)$). Eine Begründung für die Zulässigkeit einer Handlung ist $\langle M, w_a, True, \bigwedge_{c_i} \neg(I(c_i) \wedge u(c_i) < 0), HN \rangle$. Auch hier wäre ein notwendiger jedoch kein hinreichender Grund, wenn man anstelle der Konjunktion über alle Konsequenzen nur eine einzelne Konsequenz betrachtet: $\langle M, w_a, True, \neg(I(c_i) \wedge u(c_i) < 0), N \rangle$. Verboten ist eine Handlung nach dieser Bedingung, wenn eine Konsequenz mit negativem Nutzen existiert, welche beabsichtigt ist ($\exists c_i(I(c_i) \wedge u(c_i) < 0)$). Als Grund kann man hier auch die entsprechenden intendierten negativen Konsequenzen c_N $(I(c_N) \wedge u(c_N) < 0)$ finden und aufzählen: $\langle M, w_a, False, \bigwedge_{c_N}(I(c_N) \wedge u(c_N) < 0), HN \rangle$. Es würde jedoch bereits als Begründung ausreichen, wenn man eine dieser negativen Konsequenzen findet und aufzählt ($\langle M, w_a, False, I(c_N) \wedge u(c_N) < 0, H \rangle$). In diesem Fall wäre der Grund jedoch nicht mehr notwendig, denn wenn man diese eine negative Konsequenz negiert, dann ist die Formel $\bigwedge c_N(I(c_N) \wedge u(c_N) < 0)$ immer noch erfüllt, da weitere negative Konsequenzen existieren, welche die Formel wahr machen. Wenn man jedoch alle beabsichtigten negativen Konsequenzen aufzählt, dann ist durch das Negieren dieser die Bedingung nicht mehr erfüllt.

3. Bedingung: Die dritte Bedingung fordert, dass zumindest einige der positiven Konsequenzen beabsichtigt sein müssen (als logische Formel ausgedrückt: $\exists c_i(I(c_i) \wedge u(c_i) > 0)$). Als Begründung bietet es sich hier an, die beabsichtigten positiven Konsequenzen c_P $(I(c_P) \wedge u(c_P) > 0)$ zu finden und aufzuzählen. In Form des 5-Tupels würde der Grund $\langle M, w_a, True, \bigvee_{c_P}(I(c_P) \wedge u(c_P) > 0), HN\rangle$ lauten. Diese Begründung ist sowohl hinreichend als auch notwendig. Wenn jedoch mehrere solcher beabsichtigten positiven Konsequenzen existieren, würde es ausreichen, eine dieser positive Konsequenzen aufzuzählen ($\langle M, w_a, True, (I(c_P) \wedge u(c_P) > 0), H\rangle$). Diese Begründung ist kein notwendiger Grund, da durch das Negieren dieser einen Konsequenz die Bedingung weiterhin erfüllt ist, da weitere beabsichtigte positive Konsequenzen existieren können, welche diese dritte Bedingung erfüllen. Wenn man jedoch alle beabsichtigten positiven Konsequenzen aufzählt, dann wird die Bedingung durch das Negieren dieser unerfüllt und die Begründung $\bigvee_{c_P}(I(c_P) \wedge u(c_P) > 0)$ ist somit notwendig. Nach dieser dritten Bedingung ist eine Handlung verboten, wenn keine positive Konsequenz existiert, die beabsichtigt ist ($\neg\exists c_i(I(c_i) \wedge u(c_i) > 0)$. Eine hinreichende und notwendige Begründung hierfür ist $\langle M, w_a, False, \bigwedge_{c_i} \neg(I(c_i) \wedge u(c_i) > 0), HN\rangle$. Hier würde es für einen notwendigen Grund genügen eine solche intendierte positive Konsequenz zu nennen ($\langle M, w_a, False, \neg(I(c_i) \wedge u(c_i) > 0), N\rangle$), da durch die Negation dieser die Bedingung nicht mehr erfüllt wäre.

4. Bedingung: Diese Bedingung besagt, dass die negativen Konsequenzen kein Mittel dafür sein dürfen, die positiven Konsequenzen zu erreichen. Diese vierte Bedingung ist identisch zu dem Do-No-Instrumental-Harm-Prinzip, wodurch sich für diese Bedingung die selbe Begründung ergibt, wie der in Kapitel 4.2.3 erarbeitete Begründungsvorschlag des Do-No-Instrumental-Harm-Prinzips (Erlaubt: $\langle M, w_a, True, \bigwedge_{c_N,c_P} \neg(Causes(c_N, c_P) \wedge u(c_N) < 0 \wedge u(c_P) > 0), HN\rangle$, Verboten: $\langle M, w_a, False, \bigwedge_{c_N,c_P}(Causes(c_N, c_P) \wedge u(c_N) < 0 \wedge u(c_P) > 0), HN\rangle$).

5. Bedingung: Die fünfte Bedingung sagt aus, dass es verhältnismäßig schwerwiegende Gründe geben muss, um die positiven Konsequenzen zu bevorzugen, während die

negativen Konsequenzen erlaubt sind. Ein Grund für eine zulässige Handlung ist, dass der Nutzen der Konjunktion über alle direkten Konsequenzen der Handlung größer als null ist ($\langle M, w_a, True, u(\bigwedge cons_a) > 0, HN\rangle$). Wohingegen ein Grund für das Verbot einer Handlung derjenige ist, dass der Nutzen der Konjunktion über alle direkten Konsequenzen der Handlung kleiner gleich null ist ($\langle M, w_a, False, u(\bigwedge cons_a) \leq 0, HN\rangle$) und somit die negativen Konsequenzen gegenüber den positiven Konsequenzen überwiegen. Bei dieser letzten Bedingung sind sowohl die Begründung für eine erlaubte Handlung als auch die Begründung für eine verbotenen Handlung sowohl hinreichend als auch notwendig.

Um das Prinzip der Doppelwirkung als Ganzes zu begründen ist es bei einer erlaubten Handlung notwendig, alle fünf Bedingungen mit ihrer Begründung aufzuzählen, da es hier erforderlich ist, dass jede dieser fünf Bedingungen erfüllt ist. In Form des 5-Tupels würde eine Begründung folgendermaßen aussehen: $\langle M, w_a, True, u(a) \geq 0 \wedge \bigwedge_{c_i} \neg(I(c_i) \wedge u(c_i) < 0) \wedge \bigvee_{c_P}(I(c_P) \wedge u(c_P) > 0) \wedge \bigwedge_{c_N, c_P} \neg(Causes(c_N, c_P) \wedge u(c_N) < 0 \wedge u(c_P) > 0) \wedge u(\bigwedge cons_a) > 0), HN\rangle$. Jede dieser fünf Bedingungen einzeln betrachtet wäre ein notwendiger Grund für das Prinzip der Doppelwirkung, jedoch kein hinreichender. Für die Begründung einer unzulässigen Handlung genügt es, diejenigen Bedingungen anzuführen, welche nicht erfüllt sind, anstatt sowohl die erfüllten als auch die unerfüllten Bedingungen aufzuzählen: $\langle M, w_a, False, \bigvee_i(harmedCondition_i), HN\rangle$ ($\text{harmedCondition}_\text{i}$ bezeichnet eine verletzte Bedingung. Für den Fall, dass die erste Bedingung verletzt ist, würde $\text{harmedCondition}_\text{i}$ durch den Reason-Teil des 5-Tupels ersetzt, welcher in Bedingung 1 unter der Annahme, dass die Handlung nicht erlaubt ist, genannt wurde ($\text{harmedCondition}_1 = \text{u(a)} < 0$)). Bei dieser Begründung ist es auch möglich, nur eine verletzte Bedingung zu nennen, falls mehrere existieren. Diese Begründung ist dann jedoch nicht mehr notwendig, da durch das Negieren dieser einen verletzten Bedingung die Handlung nicht zulässig wird, da noch weitere Bedingungen verletzt sind.

4.2.5 Begründungsvorschlag für das Pareto-Prinzip

Um eine Begründung für das Pareto-Prinzip zu finden, ist es notwendig, sich zuerst einmal anzuschauen, wann genau eine Handlung nach diesem Prinzip erlaubt beziehungsweise verboten ist. Eine Handlung ist erlaubt, wenn sie von keiner anderen Handlung dominiert wird (vgl. 3.3.5). Somit lässt sich das Pareto-Prinzip durch die Pareto-Dominanz begründen. Um die Begründung auch für einen Nutzer, welcher möglicherweise nicht mit dieser Dominanz vertraut ist, verständlich zu machen, möchte ich darauf verzichten, als Begründung einfach nur anzubringen: "Die Handlung w ist erlaubt, da sie von keiner anderen Handlung dominiert wird" beziehungsweise "Die Handlung w_1 ist verboten, da sie von Handlung w_0 dominiert wird.", auch wenn dies die offensichtlichste Begründungsmöglichkeit für dieses Prinzip wäre. Stattdessen möchte ich in dieser Arbeit die folgende Begründungsmöglichkeit präsentieren: Als Begründung für eine erlaubte Handlung, mit anderen Worten, für eine nicht-dominierte Handlung erachte ich es als sinnvoll, die unerfüllten Dominanz-Bedingungen zu nennen. In Form des 5-Tupels würde diese Begründung für den Fall, dass die Handlung w_{a_0} die Handlung w_{a_1} nicht dominiert, folgendermaßen aussehen:

Fall 1: Sei die erste Dominanz-Bedingung verletzt: $\langle M, w_{a_1}, True, \neg \bigwedge_c((u(c) > 0 \wedge [\neg a_0 \wedge a_1]c) \rightarrow [\neg a_1 \wedge a_0]c), HN\rangle$. Dieser Grund ist hinreichend und notwendig. Als hinreichende Begründung würde auch $\langle M, w_{a_1}, True, \neg((u(c) > 0 \wedge [\neg a_0 \wedge a_1]c) \rightarrow [\neg a_1 \wedge a_0]c), H\rangle$ gelten, wobei dieser Grund nicht mehr notwendig ist, da noch weitere Konsequenzen existieren können, welche dazu führen, dass die Bedingung verletzt ist.

Fall 2: Wenn die zweite Dominanz-Bedingung nicht erfüllt ist, lässt sich die Begründung in Form des 5-Tupels folgendermaßen darstellen: $\langle M, w_{a_1}, True,$ $\neg \bigvee_c((u(c) > 0 \wedge \neg[\neg a_0 \wedge a_1]c) \rightarrow [\neg a_1 \wedge a_0]c) \wedge \neg \bigvee_c((u(c) < 0 \wedge [\neg a_0 \wedge a_1]c) \rightarrow \neg[\neg a_1 \wedge a_0]c), HN\rangle$. Hier würde es als notwendige Begründung auch ausreichen nur $\neg \bigvee_c((u(c) > 0 \wedge \neg[\neg a_0 \wedge a_1]c) \rightarrow [\neg a_1 \wedge a_0]c)$ oder $\neg \bigvee_c((u(c) < 0 \wedge [\neg a_0 \wedge a_1]c) \rightarrow$

$\neg[\neg a_1 \wedge a_0]c)$ zu nennen, da bereits durch Negation einer dieser Teilformeln die Bedingung nicht mehr erfüllt wäre.

Fall 3: Falls die dritte Dominanz-Bedingung verletzt ist, ergibt sich folgende Begründung: $\langle M, w_{a_1}, True, \neg \bigwedge_c((u(c) < 0 \wedge [\neg a_1 \wedge a_0]c) \rightarrow [\neg a_0 \wedge a_1]c), HN\rangle$. Dieser Grund ist hinreichend und notwendig. Würde man anstelle der Konjunktion über alle Konsequenzen nur eine Konsequenz nennen ($\langle M, w_{a_1}, True, \neg((u(c) < 0 \wedge [\neg a_1 \wedge a_0]c) \rightarrow [\neg a_0 \wedge a_1]c), H\rangle$), dann würde es sich um einen hinreichenden jedoch nicht um einen notwendigen Grund handeln. Dies liegt daran, dass durch eine Negation des Grundes immer noch weitere Konsequenzen existieren können, welche diese Bedingung erfüllen.

Fall 4: Für den Fall, dass mehrere Dominanz-Bedingungen verletzt sind, ergibt sich als hinreichender und notwendiger Grund eine Disjunktion über alle verletzten Dominanz-Bedingungen $\langle M, w_{a_1}, True, \bigvee_i(harmedCondition_i), HN\rangle$ (auch hier stellt, wie bereits bei dem Prinzip der Doppelwirkung, harmedCondition$_\text{i}$ die verletzten Bedingungen des Prinzips dar). Für einen hinreichenden Grund wäre auch eine verletzte Dominanz-Bedingung ausreichend ($\langle M, w_{a_1}, True, harmedCondition_i, H\rangle$), jedoch nicht für einen notwendigen Grund. Dies liegt daran, dass bereits eine verletze Dominanz-Bedingung dazu führt, dass die Handlung in jeder möglichen Situation erlaubt ist. Es kann jedoch sein, dass durch eine Negation einer verletzten Dominanz-Bedingung die Formel der Begründung immer noch erlaubt ist, da eine weitere verletzte Dominanz-Bedingung existiert, welche die Formel erfüllt.

Wenn eine Handlung nach dem Pareto-Prinzip verboten ist, dann wird eine Handlung w_{a_1} von einer Handlung w_{a_0} dominiert. In diesem Fall sind alle drei Dominanz-Bedingungen erfüllt und somit ergibt sich folgende Begründung: $\langle M, w_{a_1}, False, \bigwedge_c((u(c) > 0 \wedge [\neg a_0 \wedge a_1]c) \rightarrow [\neg a_1 \wedge a_0]c) \wedge (\bigvee_c((u(c) > 0 \wedge \neg[\neg a_0 \wedge a_1]c) \rightarrow [\neg a_1 \wedge a_0]c) \vee \bigvee_c((u(c) < 0 \wedge [\neg a_0 \wedge a_1]c) \rightarrow \neg[\neg a_1 \wedge a_0]c))) \wedge \bigwedge_c((u(c) < 0 \wedge [\neg a_1 \wedge a_0]c) \rightarrow [\neg a_0 \wedge a_1]c), HN\rangle$. Hierbei ist die Begründung sowohl hinreichend als auch

notwendig. Jede dieser Konjunktionen für sich betrachtet wäre auch ein notwendiger, jedoch kein hinreichender Grund.

4.3 Überblick über die Lösungsansätze

Abschließend für dieses Kapitel möchte ich noch einen Überblick über die erarbeiteten Begründungsvorschläge geben. Hierzu werden nur die reason-Teile des 5-Tupels genannt, um eine bessere Lesbarkeit zu gewährleisten. Hierzu sind alle hinreichenden und notwendigen Gründe für die Prinzipien im Erlaubt- und im Verboten-Fall aufgelistet. Die einzelnen Bedingungen der Prinzipien sind dabei außen vorgelassen, da es sonst eine zu große Menge an Gründen wäre, um diese noch überblicken zu können.

Deontologische Prinzip:

- erlaubt: hinreichend und notwendig: $u(a) \geq 0$
- verboten: hinreichend und notwendig: $u(a) < 0$

Do-No-Harm-Prinzip:

- erlaubt:
  - hinreichend: $\bigwedge_{c_i} \neg(Causes(a, c_i) \wedge u(c_i) < 0)$
  - notwendig: $\neg(Causes(a, c_i) \wedge u(c_i) < 0)$
- verboten:
  - hinreichend: $Causes(a, c_i) \wedge u(c_i) < 0$
  - notwendig: $\bigvee_{c_i} Causes(a, c_i) \wedge u(c_i) < 0$

Do-No-Instrumental-Harm-Prinzip:

- erlaubt:
  - hinreichend: $\bigwedge_{c_N, c_P} \neg(Causes(c_N, c_P) \wedge u(c_N) < 0 \wedge u(c_P) > 0)$
  - notwendig: $\neg(Causes(c_N, c_P) \wedge u(c_N) < 0 \wedge u(c_P) > 0)$

- verboten:
  - hinreichend: $Causes(c_N, c_P) \wedge u(c_N) < 0 \wedge u(c_P) > 0$
  - notwendig: $\bigvee_{c_N, c_P}(Causes(c_N, c_P) \wedge u(c_N) < 0 \wedge u(c_P) > 0)$

Prinzip der Doppelwirkung:

- erlaubt:
  - hinreichend: $u(a) \geq 0 \wedge \bigwedge_{c_i} \neg(I(c_i) \wedge u(c_i) < 0) \wedge$
    $\bigvee_{c_P}(I(c_P) \wedge u(c_P) > 0) \wedge$
    $\bigwedge_{c_N, c_P} \neg(Causes(c_N, c_P) \wedge u(c_N) < 0 \wedge u(c_P) > 0) \wedge$
    $u(\bigwedge cons_a) > 0)$
  - notwendig: $u(a) \geq 0$, $\bigwedge_{c_i} \neg(I(c_i) \wedge u(c_i) < 0)$,
    $\bigvee_{c_P}(I(c_P) \wedge u(c_P) > 0)$,
    $\neg(Causes(c_N, c_P) \wedge u(c_N) < 0 \wedge u(c_P) > 0)$,
    $u(\bigwedge cons_a) > 0$
- verboten:
  - hinreichend: $harmedCondition_i$
  - notwendig: $\bigvee_i(harmedCondition_i)$

Pareto Prinzip:

- erlaubt:
  - hinreichend: $harmedCondition_i$
  - notwendig: $\bigvee_i(harmedCondition_i)$
- verboten:
  - hinreichend: $\bigwedge_c((u(c) > 0 \wedge [\neg a_0 \wedge a_1]c) \rightarrow [\neg a_1 \wedge a_0]c) \wedge$
    $(\bigvee_c((u(c) > 0 \wedge \neg[\neg a_0 \wedge a_1]c) \rightarrow [\neg a_1 \wedge a_0]c) \vee$
    $\bigvee_c((u(c) < 0 \wedge [\neg a_0 \wedge a_1]c) \rightarrow \neg[\neg a_1 \wedge a_0]c)) \wedge$
    $\bigwedge_c((u(c) < 0 \wedge [\neg a_1 \wedge a_0]c) \rightarrow [\neg a_0 \wedge a_1]c)$
  - notwendig: $u(c) > 0 \wedge [\neg a_0 \wedge a_1]c) \rightarrow [\neg a_1 \wedge a_0]c$,
    $\bigvee_c((u(c) > 0 \wedge \neg[\neg a_0 \wedge a_1]c) \rightarrow [\neg a_1 \wedge a_0]c) \vee$

$$\bigvee_c((u(c) < 0 \wedge [\neg a_0 \wedge a_1]c) \rightarrow \neg[\neg a_1 \wedge a_0]c),$$
$$(u(c) < 0 \wedge [\neg a_1 \wedge a_0]c) \rightarrow [\neg a_0 \wedge a_1]c$$

Bei dieser Übersicht fällt auf, dass für alle Prinzipien, außer dem Parteo-Prinzip, die hinreichenden Gründe im Erlaubt-Fall eine Konjunktion über alle Bedingungen beziehungsweise alle Konsequenzen darstellen und für die notwendigen Gründe eine einzelne Bedingung oder Konsequenz als Begründung ausreichend ist. Für den Verboten-Fall verhält es sich genau anders herum. Hier sind die notwendigen Gründe meist eine Disjunktion aller Bedingungen oder Konsequenzen und für einen hinreichenden Grund ist es ausreichend, eine Bedingung beziehungsweise Konsequenz zu finden. Für das Pareto-Prinzip verhält es sich wiederum genau umgekehrt. Dies liegt daran, dass eine Handlung nach dem Pareto-Prinzip erlaubt ist, wenn die Dominanz-Bedingung verletzt ist.

5 Generierung von Gründen

5.1 Naiver Ansatz

Bevor ich einen Implementierungsvorschlag für die in Kapitel 4.2 vorgestellten Begründungsspezifikationen mache, möchte ich die Begründungen zuerst anhand ausgewählter moralischer Dilemmata genauer betrachten. Dabei werde ich im Folgenden an Stelle des gesamten 5-Tupels nur den reason-Teil angeben, um es so für den Leser übersichtlicher zu gestalten.

Als erstes Prinzip möchte ich das Deontologische Prinzip anhand des in Kapitel 3.2.4 vorgestellte Lügen-Dilemmas betrachten. Die Handlung "refrain" ist nach diesem Prinzip erlaubt. Nach dem in Kapitel 4.2.1 erarbeiteten Begründungsvorschlag lautet der Grund: $u(refrain) \geq 0$. Dieser Grund ist sowohl hinreichend als auch notwendig. Die Handlung "lying" des Lügen-Dilemmas ist nach dem Deontologischen Prinzip verboten, da diese einen negativen Nutzen besitzt. Der hinreichende und notwendige Grund hierfür ist $u(lying) < 0$.

Für das Do-No-Harm-Prinzip habe ich das Trolley-Dilemma (3.2.1) zu Hilfe genommen. Unter diesem Prinzip ist die Handlung "refrain" des Trolley-Dilemmas erlaubt, da diese nicht den Schaden ("5 Personen sterben") verursacht. Die hinreichende und notwendige Begründung hierfür ist $\neg Causes(refrain, c2) \wedge u(c2) < 0$ (vgl. 4.2.2). Die Handlung "pull" des Trolley-Dilemmas ist hingegen verboten, da sie einen Schaden

(“eine Person stirbt”) verursacht. Die Begründung, welche hinreichend und notwendig ist, lautet $Causes(pull, c1) \wedge u(c1) < 0$.

Um die Begründungen des Do-No-Instrumental-Harm-Prinzips zu untersuchen betrachte ich für die erlaubte Handlung das Trolley-Dilemma, bei welchem im Gegensatz zu dem Do-No-Harm-Prinzip, die Handlung “pull” erlaubt ist. Die Begründung hierfür sieht folgendermaßen aus: $\neg(Causes(c1, \neg c2) \wedge u(c1) < 0 \wedge u(\neg c2) > 0)$. Dieser Grund ist hinreichend und notwendig. Für den Verboten-Fall des Do-No-Instrumental-Harm-Prinzips habe ich mir das Fatman-Trolley-Dilemma (3.2.2) angeschaut. (Ich habe das Fatman-Trolley-Dilemma hier nicht auch für die erlaubte Handlung verwendet, da bei der Handlung “refrain” kein Schaden verursacht wird, wodurch die Begründung die Gleiche wäre wie die für das Do-No-Harm-Prinzip.) Die Handlung “push” ist verboten, da der Schaden (“der fette Mann stirbt”) ein Mittel ist, um die fünf Personen zu retten. Die Begründung lautet $Causes(c1, c2) \wedge u(c1) < 0 \wedge u(c2) > 0$ und ist sowohl hinreichend als auch notwendig.

Für das Prinzip der Doppelwirkung habe ich das Flugzeugentführungs-Dilemma (3.2.3) untersucht. Die Handlung “shoot” des Dilemmas ist erlaubt. Die hinreichende und notwendige Begründung hierfür lautet nach Kapitel 4.2.4:

$$
\begin{aligned}
& u(shoot) \geq 0 \wedge \\
& \neg(I(\neg c2) \wedge u(\neg c2) < 0) \wedge \neg(I(\neg c3) \wedge u(\neg c3) < 0) \wedge \\
& (I(\neg c2) \wedge u(\neg c2) > 0 \vee I(\neg c3) \wedge u(\neg c3) > 0) \wedge \\
& \neg(Causes(c1, \neg c2) \wedge u(c1) < 0 \wedge u(\neg c2) > 0) \wedge \\
& \neg(Causes(c1, \neg c3) \wedge u(c1) < 0 \wedge u(\neg c3) > 0) \wedge \\
& u(c1 \wedge \neg c2 \wedge \neg c3) > 0
\end{aligned}
\tag{12}
$$

Jeder dieser Konjunktionsterme stellt einzeln betrachtet auch eine notwendige Begründung dar, da durch die Negation eines solchen Konjunktionsterm die gesamte Formel unerfüllt wird. Diese Begründung ist jedoch nicht mehr hinreichend.

Die Handlung “refrain” des Flugzeugentführungs-Dilemmas ist unter dem Prinzip der Doppelwirkung verboten (die dritte und die fünfte Bedingung sind verletzt). Der notwendige und hinreichende Grund hierfür ist $(\neg(I(c1) \wedge u(c1) > 0) \wedge \neg(I(c2) \wedge u(c2) > 0) \wedge \neg(I(c3) \wedge u(c3) > 0)) \vee u(c1 \wedge c2 \wedge c3) \leq 0$. Für einen hinreichenden Grund wäre es ausreichend nur eine der nicht-intendierten Konsequenzen $(\neg(I(c1) \wedge u(c1) > 0)$, $\neg(I(c2) \wedge u(c2) > 0)$ oder $\neg(I(c3) \wedge u(c3) > 0))$ oder $u(c1 \wedge c2 \wedge c3) \leq 0$ zu nennen.

Als letztes Prinzip betrachte ich die Begründungen des Pareto-Prinzips anhand des Flugzeugentführungs-Dilemmas. Die Handlung “shoot” ist, genau wie unter dem Prinzip der Doppelwirkung, erlaubt. Die notwendige und hinreichende Begründung hierfür lautet (aus Gründen der Übersichtlichkeit sind nur die erfüllten Teilformeln gelistet):

$$(\neg((u(\neg c2) > 0 \wedge [\neg refrain \wedge shoot]\neg c2) \rightarrow [\neg shoot \wedge refrain]\neg c2) \vee \quad (13)$$

$$\neg((u(\neg c3) > 0 \wedge [\neg refrain \wedge shoot]\neg c3) \rightarrow [\neg shoot \wedge refrain]\neg c3)) \vee \quad (14)$$

$$(\neg((u(c1) > 0 \wedge \neg[\neg refrain \wedge shoot]c1) \rightarrow [\neg shoot \wedge refrain]c1) \wedge \quad (15)$$

$$\neg((u(c1) < 0 \wedge [\neg refrain \wedge shoot]c1) \rightarrow \neg[\neg shoot \wedge refrain]c1)) \quad (16)$$

Die Teilformeln (13) und (14) stehen dafür, dass die erste Dominanz-Bedingung verletzt worden ist und die Teilformeln (15) und (16) besagen, dass die zweite Bedingung verletzt ist. Die dritte Dominanz-Bedingung ist erfüllt, wodurch diese nicht in der Begründung vorkommt. Jede dieser Teilformel alleine betrachtet ist ebenfalls eine hinreichende Begründung, jedoch kein notwendiger Grund.
Die Handlung “refrain” des Flugzeugentführungs-Dilemmas ist verboten, da sie von der Handlung “shoot” dominiert wird. Die notwendige und hinreichende Begründung sieht folgendermaßen aus:

$$(((u(c1) > 0 \wedge [\neg shoot \wedge refrain]c1) \rightarrow [\neg refrain \wedge shoot]c1) \wedge \quad (17)$$

$$((u(c2) > 0 \wedge [\neg shoot \wedge refrain]c2) \rightarrow [\neg refrain \wedge shoot]c2) \wedge \quad (18)$$

$$((u(c3) > 0 \wedge [\neg shoot \wedge refrain]c3) \rightarrow [\neg refrain \wedge shoot]c3)) \wedge \quad (19)$$

$$((((u(c1) > 0 \wedge \neg[\neg shoot \wedge refrain]c1) \rightarrow [\neg refrain \wedge shoot]c1) \vee \quad (20)$$

$$((u(c2) > 0 \wedge \neg[\neg shoot \wedge refrain]c2) \rightarrow [\neg refrain \wedge shoot]c2) \vee \quad (21)$$

$$((u(c3) > 0 \wedge \neg[\neg shoot \wedge refrain]c3) \rightarrow [\neg refrain \wedge shoot]c3)) \vee \quad (22)$$

$$(((u(c2) < 0 \wedge [\neg shoot \wedge refrain]c2) \rightarrow \neg[\neg refrain \wedge shoot]c2) \vee \quad (23)$$

$$((u(c3) < 0 \wedge [\neg shoot \wedge refrain]c3) \rightarrow \neg[\neg refrain \wedge shoot]c3))) \wedge \quad (24)$$

$$(((u(c1) < 0 \wedge [\neg refrain \wedge shoot]c1) \rightarrow [\neg shoot \wedge refrain]c1) \wedge \quad (25)$$

$$((u(c2) < 0 \wedge [\neg refrain \wedge shoot]c2) \rightarrow [\neg shoot \wedge refrain]c2) \wedge \quad (26)$$

$$((u(c3) < 0 \wedge [\neg refrain \wedge shoot]c3) \rightarrow [\neg shoot \wedge refrain]c3)) \quad (27)$$

Als notwendige Begründung würde es auch genügen eine der Teilformeln (20) bis (24) auszuwählen. Ein notwendiger Grund wäre dann zum Beispiel:

$$(((u(c1) > 0 \wedge [\neg shoot \wedge refrain]c1) \rightarrow [\neg refrain \wedge shoot]c1) \wedge \quad (28)$$

$$((u(c2) > 0 \wedge [\neg shoot \wedge refrain]c2) \rightarrow [\neg refrain \wedge shoot]c2) \wedge \quad (29)$$

$$((u(c3) > 0 \wedge [\neg shoot \wedge refrain]c3) \rightarrow [\neg refrain \wedge shoot]c3)) \wedge \quad (30)$$

$$((u(c3) < 0 \wedge [\neg shoot \wedge refrain]c3) \rightarrow \neg[\neg refrain \wedge shoot]c3) \wedge \quad (31)$$

$$(((u(c1) < 0 \wedge [\neg refrain \wedge shoot]c1) \rightarrow [\neg shoot \wedge refrain]c1) \wedge \quad (32)$$

$$((u(c2) < 0 \wedge [\neg refrain \wedge shoot]c2) \rightarrow [\neg shoot \wedge refrain]c2) \wedge \quad (33)$$

$$((u(c3) < 0 \wedge [\neg refrain \wedge shoot]c3) \rightarrow [\neg shoot \wedge refrain]c3)) \quad (34)$$

Bei den Begründungen des Pareto-Prinzips fällt auf, dass es sich im Verboten-Fall um eine sehr lange Formel handelt. Dies liegt daran, dass alle Dominanz-Bedingungen für jede mögliche Belegung erfüllt sein muss. Dadurch dass über alle Konsequenzen argumentiert wird, entsteht eine lange, nicht auf den ersten Blick ersichtliche Formel als Grund. Dies lässt sich jedoch nur ändern, wenn man direkt über die Pareto-Dominanz argumentiert, anstatt sich die Dominanz-Bedingungen selbst anzuschauen. Da dies jedoch, wie bereits in Kapitel 3.3.5 erwähnt, nicht erwünscht ist, lässt sich dieser lange Grund meiner Ansicht nach nicht umgehen.

Wie kann man nun die in Kapitel 4.2 vorgestellten Begründungsspezifikationen in dem HERA-System implementieren? Man müsste jedes ethische Prinzip um eine eigene Funktion erweitern, welche die Begründungsspezifikation enthält. Dies wäre von der Umsetzung her gut machbar, jedoch mit einem hohen Aufwand für den Modellierer verbunden. Das liegt daran, dass für jedes weitere ethische Prinzip, welches dem HERA-System hinzugefügt werden soll, zuerst die Begründungen analysiert werden müssten, bevor dieses Prinzip modelliert werden kann. Das ist darauf zurückzuführen, dass es dem HERA-System nicht möglich ist zu erkennen, wie die Teilformeln eines ethischen Prinzips miteinander in Zusammenhang stehen. Um eine Analyse der Begründung für jedes weitere ethische Prinzip zu vermeiden, muss ein anderer Ansatz gefunden werden, um die Entscheidungen des Agenten zu begründen. In dem folgenden Kapitel werde ich eine solche Alternative zur Generierung von Begründungen vorstellen.

5.2 Ansatz basierend auf der Analyse der Disjunktiven Normalform

Da die von mir erarbeiteten Spezifikationen zur Begründung der ethischen Prinzipien sich nicht so wie gehofft im HERA-System umsetzen lassen, habe ich nach einer Alternative gesucht, um Begründungen zu implementieren. Wenn man sich die Prinzipien noch einmal genauer anschaut, dann fällt auf, dass alle ethischen Prinzipien, welche ich in dieser Arbeit behandelt habe, eine Sache gemeinsam haben: Sie enthalten ihre Bedingungen für die Zulässigkeit beziehungsweise für die Unzulässigkeit einer Handlung in Form einer logischen Formel und diese Formel ist bereits in der Implementation der Prinzipien im HERA-System enthalten. Gibt es also einen Weg aus diesen Formeln die Begründungen zu generieren, da auf diese Weise der zusätzliche Aufwand des naiven Ansatzes (siehe Kapitel 5.1) erspart bleibt? Jede dieser logischen Formeln lässt sich als Disjunktive Normalform (DNF) darstellen. Wenn mindestens einer der Konjunktionsterme der DNF erfüllt ist, dann ist die DNF

erfüllt und somit ist die Handlung erlaubt. Um die Gründe für eine solche erlaubte Handlung zu finden, kann man die Bedingungen des Prinzips direkt in eine DNF umformen. Für eine verbotene Handlung ist es jedoch notwendig, die logische Formel, welche die Bedingungen des Prinzips darstellt, zuerst zu negieren und anschließend aus der negierten Formel eine Disjunktive Normalform zu bilden. Dies liegt daran, dass man für eine verbotene Handlung in einer Konjunktiven Normalform (KNF) die unerfüllten Disjunktionsterme als Gründe betrachten kann. Durch die Negation wird aus der unerfüllten KNF eine erfüllte DNF und wir können dieselbe Vorgehensweise anwenden wie bei einer erlaubten Handlung. Dadurch genügt es, ganz am Anfang des Programms, vor dem Erzeugen der DNF, eine Unterscheidung zwischen erlaubten und verbotenen Handlungen zu machen. Im Rest des Programms können alle Handlungen gleich behandelt werden. Die einzige Unterscheidung, welche noch erforderlich ist, ist diejenige zwischen notwendigen und hinreichenden Gründen. Beide Arten von Gründen benötigen die Konjunktionsterme der DNF, welche das Kausale Modell erfüllen. Ein solcher erfüllender Konjunktionsterm stellt bereits einen hinreichenden Grund dar, weil ein erfüllender Konjunktionsterm die DNF in jeder möglichen Situation wahr macht. Dies lässt sich auch in dem Überblick in Kapitel 4.3 erkennen. Für einen hinreichenden Grund ist es bei fast jedem Prinzip erforderlich, dass er eine Konjunktion über alle Bedingungen oder Konsequenzen ist. Nur ein Teil der Bedingungen oder Konsequenzen wäre für einen hinreichenden Grund nicht ausreichend. Für einen notwendigen Grund hingegen genügen auch einzelne Konsequenzen oder Bedingungen. Deshalb ist es erforderlich aus den Konjunktionstermen diejenigen minimalen Mengen von Literalen zu finden, durch deren Negation die DNF nicht mehr erfüllt ist. Somit ist es möglich, ein Programm zu implementieren, welches nicht für jedes ethische Prinzip eine eigene Funktion enthalten muss, sondern allgemein für jedes Prinzip anwendbar ist, wodurch es nicht notwendig ist, dass die Definition des Prinzips die Gründe enthält. Listing 5.1 enthält den Pseudo-Code einer Funktion generateReasons, welche als Input drei Informationen benötigt. Die Funktion benötigt die logische Formel des ethischen Prinzips (formel), die Information, ob die die Handlung erlaubt

(True) oder verboten (False) ist, in Form einer booleschen Variable (perm) und das Kausale Modell (model), welches auch die Information enthält, welche Handlung betrachtet wird. Anhand der logischen Formel erstellt das Programm eine DNF und aus dieser DNF erstellt es eine Liste der einzelnen Konjunktionsterme. Über diese Konjunktionsterme wird iteriert. Wenn ein Konjunktionsterm das Kausale Modell erfüllt, dann gilt dieser Term als hinreichender Grund. Für die notwendigen Gründe wird zusätzlich noch eine Hilfsfunktion in Form eines ASP-Programms (Answer Set Programming) aufgerufen. Diese Hilfsfunktion generiert für jeden Konjunktionsterm, welcher das Kausale Modell erfüllt, alle möglichen Mengen an Literalen und gibt diejenigen zurück, welche durch Negation die DNF nicht mehr erfüllen würden. Es ist noch nötig, die Mengen an Literalen auf Minimalität zu überprüfen und nur die minimalen Mengen an Literalen als notwendigen Grund zu nennen, da sonst auch der gesamte Konjunktionsterm als notwendiger Grund genannt wird, obwohl es einen kleineren notwendigen Grund gibt. Ein weiterer Unterschied zu den hinreichenden Gründen besteht zusätzlich darin, dass hier keine Konjunktion über die minimalen Mengen an Literalen gemacht wird, sondern dass der notwendige Grund aus der Disjunktion der minimalen Menge an Literalen besteht. Dies liegt daran, dass es als notwendige Begründung ausreicht eine minimale Menge an Literalen zu nennen, da durch deren Negation bereits die Bedingung nicht mehr erfüllt ist.

```
def generateReasons(formel, perm, model):
    if perm:
        formel = formel
    else:
        formel = Not(formel)
    result = []
    formelDnf = formel.dnf()
    formelDnfList = formelDnf.asConjList()
    trueConjunctions = []
    for c in formelDnfList:
        if model.models(Formula.makeConjunction(c)):
            trueConjunctions.append(c)
```

```
                result.append({"model": model, "perm": perm, "reason":
                    Formula.makeConjunction(c), "type": "sufficient"})
    for c in getReasonsFromASP(trueConjunctions):
        result.append({"model": model, "perm": perm, "reason":
            Formula.makeDisjunction(c), "type": "necessary"})
    return result
```

Listing 5.1: Python-Code der Funktion generateReasons zur Erstellung einer Begründung

Ich möchte die Funktionsweise des Programms exemplarisch anhand des Trolley-Dilemmas (vgl. 3.2.1) und des Do-No-Harm-Prinzips (vgl. 3.3.2) darstellen: Die Handlung "refrain" ist erlaubt. Die Variable formel beschreibt die Formel $(Causes(refrain, c2) \rightarrow u(c2) \geq 0) \wedge (Causes(refrain, \neg c1) \rightarrow u(\neg c1) \geq 0)$. Diese wird dann in eine DNF umgewandelt:

$$\begin{aligned} formelDnf = \quad & (\neg Causes(refrain, c2) \wedge \neg Causes(refrain, \neg c1)) \vee \\ & (\neg Causes(refrain, c2) \wedge u(\neg c1) \geq 0) \vee \\ & (u(c2) \geq 0 \wedge \neg Causes(refrain, \neg c1)) \vee \\ & (u(c2) \geq 0 \wedge u(\neg c1) \geq 0) \end{aligned}$$

Diese DNF stellt das Programm in Form einer Liste dar, die folgendermaßen aussieht: $[[\neg Causes(refrain, c2), \neg Causes(refrain, \neg c1)], [\neg Causes(refrain, c2), u(\neg c1) \geq 0], [u(c2) \geq 0, \neg Causes(refrain, \neg c1)], [u(c2) \geq 0, u(\neg c1) \geq 0]]$. Über diese Liste wird dann die Iteration durchgeführt. Dabei sind nur $[\neg Causes(refrain, c2), \neg Causes(refrain, \neg c1)]$ und $[\neg Causes(refrain, c2), u(\neg c1) \geq 0]$ in dem Modell gültig. Dementsprechend werden durch die Konjunktion als hinreichende Gründe $\neg Causes(refrain, c2) \wedge \neg Causes(refrain, \neg c1)$ und $\neg Causes(refrain, c2) \wedge u(\neg c1) \geq 0$ zurückgegeben. Für die notwendigen Gründe gibt das ASP-Programm als minimale Menge an Literalen die Liste $[[\neg Causes(refrain, \neg c1), u(\neg c1) \geq 0], [\neg Causes(refrain, c2)]]$ zurück. Hieraus ergeben sich durch die Disjunktion folgende notwendige Gründe: $\neg Causes(refrain, \neg c1) \vee u(\neg c1) \geq 0$ und $\neg Causes(refrain, c2)$.

Verbalisiert würde der erste hinreichende Grund “Die Handlung “refrain” ist erlaubt, da sie weder c2 (“fünf Personen sterben”) noch $\neg c1$ (“eine Person überlebt”) verursacht” lauten. Der erste notwendige Grund hingegen sagt aus, dass “Die Handlung “refrain” erlaubt ist, da sie mindestens $\neg c1$ nicht verursacht oder der Nutzen von $\neg c1$ größer gleich null ist”.

5.2.1 Analyse und Diskussion des DNF-Ansatzes im Vergleich zu den Begründungsspezifikationen

Um das gerade vorgestellte Programm beurteilen zu können, im Vergleich zu den in dieser Arbeit beschriebenen Begründungen und hinsichtlich der Verständlichkeit für den Nutzer, werde ich im Folgenden für ausgewählte Dilemmata die Begründungen des Programms untersuchen. Dafür werde ich die Gründe des Programms mit denen in Kapitel 5.1 vorgestellten Dilemmata und Begründungen vergleichen.

Das Programm gibt für das Lügen-Dilemma unter Anwendung des Deontologischen Prinzips für die erlaubte Handlung “refrain” sowohl als hinreichenden als auch als notwendigen Grund $u(refrain) \geq 0$ zurück. Diese Begründung stimmt mit der Begründung in Kapitel 5.1 überein. Die Handlung “lying” des Lügen-Dilemmas ist unter dem Deontologischen Prinzip verboten. Die Begründung des Programms ist im hinreichenden und im notwendigen Fall $u(lying) < 0$. Auch dieser Grund stimmt mit der in dieser Arbeit beschriebenen Begründung überein.

Als zweites habe ich das Do-No-Harm-Prinzip unter Zuhilfenahme des Trolley-Dilemmas (3.2.1) betrachtet. Nach diesem Prinzip ist die Handlung “refrain” des Trolley-Dilemmas erlaubt. Das Programm gibt hierfür zwei hinreichende Gründe $\neg Causes(refrain, c2) \wedge \neg Causes(refrain, \neg c1)$ und $\neg Causes(refrain, c2) \wedge u(\neg c1) \geq 0$ und zwei notwendige Gründe $\neg Causes(refrain, \neg c1) \vee u(\neg c1) \geq 0$ und $\neg Causes(refrain, c2)$ zurück. Diese Gründe stimmen nicht mit der Begründung aus Kapitel 5.1 ($\neg Causes(refrain, c2) \wedge u(c2) < 0$) überein, welche sowohl hinreichend

als auch notwendig ist. Die Begründungen des Programms sind ohne Wissen über das Kausale Modell und das Do-No-Harm-Prinzip nicht gut nachzuvollziehen, da bei dem ersten hinreichenden und dem zweiten notwendigen Grund nichts über den Nutzen der Konsequenzen gesagt wird und es somit nicht ersichtlich ist, ob es sich um gute oder um schlechte Konsequenzen handelt. Der zweite hinreichende Grund sagt aus, dass c2 nicht durch "refrain" verursacht wird und das $\neg c1$ gut ist. Ohne das Kausale Modell zu kennen, sagt mir der Grund nicht, dass c2 eine schlechte Konsequenz ist, weshalb die Begründung nicht besonders gut zu verstehen ist. Auch stellt sich die Frage, was genau der Nutzen von $\neg c1$ darüber aussagt, warum die Handlung erlaubt ist, da es hierfür erforderlich ist, mit dem Do-No-Harm-Prinzip vertraut zu sein. Die Gründe für die verbotene Handlung "pull" des Trolley-Dilemmas lauten $Causes(pull, c1) \wedge u(c1) < 0$ (hinreichend), $u(c1) < 0$ (notwendig) und $Causes(pull, c1)$ (notwendig). Der hinreichende Grund stimmt mit der Begründung aus Kapitel 5.1 überein. Die notwendigen Gründe sind jedoch ohne Kenntnisse über das Kausale Modell und das Prinzip nicht besonders leicht nachzuvollziehen. Die Begründung: "Die Handlung pull ist verboten, weil c1 schlecht ist", ist nur dann verständlich, wenn bekannt ist, dass: "pull" c1 verursacht und eine Handlung nach dem Do-No-Harm-Prinzip verboten ist, wenn sie einen Schaden verursacht. Auch der notwendige Grund $Causes(pull, c1)$ ist nur dann nachvollziehbar, wenn man weiß, dass c1 schlecht ist und mit dem Do-No-Harm-Prinzip vertraut ist.

Im nächsten Schritt untersuche ich die Gründe des Programms für das Do-No-Instrumental-Harm-Prinzip. Die Begründung des Programms, warum die Handlung "pull" des Trolley-Dilemmas erlaubt ist lautet für den hinreichenden Grund: $\neg(u(c1) > 0) \wedge \neg Causes(c1, \neg c2) \wedge \neg(u(\neg c2) < 0)$ und für den notwendigen Grund: $\neg Causes(c1, \neg c2)$, $\neg(u(\neg c2) < 0)$ und $\neg(u(c1) > 0)$. Der hinreichende Grund entspricht der Begründung basierend auf den in Kapitel 4.2.3 vorgestellten Spezifikationen. Die notwendigen Gründe ergeben konjugiert den hinreichenden Grund. Die einzelnen notwendigen Gründe sind jedoch nicht besonders intuitiv, da sich vor allem für $\neg(u(\neg c2) < 0)$ und $\neg(u(c1) > 0)$ die Frage stellt, was genau dieser Grund mit dem

Do-No-Instrumental-Harm-Prinzip zu tun hat. Selbst wenn man mit der Definition des Prinzips vertraut ist, ist es trotzdem noch erforderlich zu wissen, dass es nur die beiden Konsequenzen c1 und $\neg c2$ gibt und dass $\neg c2$ nicht von c1 verursacht wird. Ohne ein ausreichendes Wissen über das ethische Prinzip und das Kausale Modell sind diese notwendigen Gründe nicht zu verstehen. Für die verbotene Handlung des Do-No-Instrumental-Harm-Prinzips betrachte ich, genau wie in Kapitel 5.1, die Handlung "push" des Fatman-Trolley-Dilemmas. Der hinreichende Grund ist $Causes(c1, c2) \wedge u(c1) < 0 \wedge u(c2) > 0$ und ist identisch zu der Begründung basierend auf dem Vorschlag in Kapitel 4.2.3. Die notwendigen Gründe des Programms sind $u(c1) < 0$, $u(c2) > 0$ und $Causes(c1, c2)$. Jeder dieser notwendigen Gründe lässt die gleichen Fragen offen, wie bereits die notwendigen Gründe des Programms für die erlaubte Handlung: "Wie stehen c1 und c2 in Zusammenhang?", "Wieso ist "push" verboten, weil c1 c2 verursacht?" und "Welchen Nutzen haben diese Konsequenzen?". Somit ist es auch in diesem Fall erforderlich, sowohl das Kausale Modell zu kennen, als auch mit dem Do-No-Instrumental-Harm-Prinzip vertraut zu sein.

Für das Prinzip der Doppelwirkung habe ich mir wieder das Flugzeugentführungs-Dilemma angeschaut. Die Handlung "shoot" ist erlaubt und wird von dem Programm folgendermaßen begründet: Die hinreichenden Gründe sind:

$$\begin{aligned} u(shoot) \geq 0 \wedge I(\neg c2) \wedge u(\neg c2) > 0 \wedge \neg I(c1) \wedge u(\neg c2) \geq 0 \wedge \\ u(\neg c3) \geq 0 \wedge \neg u(c1) > 0 \wedge \neg Causes(c1, \neg c2) \wedge \\ \neg Causes(c1, \neg c3) \wedge u(c1 \wedge \neg c2 \wedge \neg c3) > 0 \end{aligned} \tag{35}$$

$$\begin{aligned} u(shoot) \geq 0 \wedge I(\neg c3) \wedge u(\neg c3) > 0 \wedge \neg I(c1) \wedge u(\neg c2) \geq 0 \wedge \\ u(\neg c3) \geq 0 \wedge \neg u(c1) > 0 \wedge \neg Causes(c1, \neg c2) \wedge \\ \neg Causes(c1, \neg c3) \wedge u(c1 \wedge \neg c2 \wedge \neg c3) > 0 \end{aligned} \tag{36}$$

Diese beiden Gründe sind stimmig mit den in Kapitel 5.1 erarbeiteten Begründungen.

Die notwendigen Gründe lauten: $u(\neg c3) \geq 0$, $\neg I(c1)$, $u(shoot) \geq 0$, $u(\neg c2) > 0 \vee I(\neg c3)$, $u(\neg c2) > 0 \vee u(\neg c3) > 0$, $\neg Causes(c1, \neg c2)$, $\neg Causes(c1, \neg c3)$, $u(\neg c2) \geq 0$, $I(\neg c2) \vee u(\neg c3) > 0$, $I(\neg c2) \vee I(\neg c3)$, $u(c1 \wedge \neg c2 \wedge \neg c3) > 0$ und $u(c1) < 0$. In diesem Fall gibt das Programm zwölf verschiedene notwendige Gründe zurück. Allein durch die Anzahl der Gründe kann der Nutzer überfordert werden. Ist es sinnvoll sich einen bestimmten Grund herauszusuchen oder sollte man alle nennen? Auch stellt sich die Frage, ob es Gründe gibt, die ausschlaggebender sind als andere. So könnte es hier beispielsweise sein, dass man zufällig $\neg Causes(c1, \neg c2)$ als Grund auswählt, welcher besagt, dass der Tod der Passagiere keine Ursache für den Erhalt des Gebäudes ist, obwohl $\neg Causes(c1, \neg c3)$ (der Tod der Passagiere ist keine Ursache für das Überleben der 500 Personen) in diesem Fall wesentlich intuitiver wäre. Außerdem ist ein einzelner notwendiger Grund nicht besonders informativ. Wen man sich zum Beispiel den Grund $\neg I(c1)$ anschaut, dann weiß man nicht, ob c1 gut oder schlecht ist und ohne mit den Bedingungen des Prinzips der Doppelwirkung vertraut zu sein, ist es auch nicht ersichtlich, was dieser Grund mit dem Prinzip zu tun hat und wieso deswegen die Handlung verboten ist. Für die verbotene Handlung "refrain" gibt das Programm neun hinreichende Gründe zurück:

$$\neg I(c1) \wedge \neg I(c2) \wedge \neg I(c3) \tag{37}$$

$$\neg I(c1) \wedge u(c2) \leq 0 \wedge \neg I(c3) \tag{38}$$

$$u(c1) \leq 0 \wedge \neg I(c2) \wedge \neg I(c3) \tag{39}$$

$$u(c1) \leq 0 \wedge u(c2) \leq 0 \wedge \neg I(c3) \tag{40}$$

$$\neg I(c1) \wedge \neg I(c2) \wedge u(c3) \leq 0 \tag{41}$$

$$\neg I(c1) \wedge u(c2) \leq 0 \wedge u(c3) \leq 0 \tag{42}$$

$$u(c1) \leq 0 \wedge \neg I(c2) \wedge u(c3) \leq 0 \tag{43}$$

$$u(c1) \leq 0 \wedge u(c2) \leq 0 \wedge u(c3) \leq 0 \tag{44}$$

$$u(c1 \wedge c2 \wedge c3) \leq 0 \tag{45}$$

Auch diese hinreichenden Begründungen sind ohne den Kontext der Bedingungen des Prinzips der Doppelwirkung nicht wirklich verständlich und können zu Fragen wie “Wieso darf ich die Handlung “refrain” nicht ausführen, bloß weil keine der Konsequenzen intendiert ist?” führen. Die notwendigen Gründe in diesem Fall sind: $\neg I(c2) \vee u(c2) \leq 0 \vee u(c1 \wedge c2 \wedge c3) \leq 0$, $\neg I(c3) \vee u(c3) \leq 0 \vee u(c1 \wedge c2 \wedge c3) \leq 0$ und $\neg I(c1) \vee u(c1) \leq 0 \vee u(c1 \wedge c2 \wedge c3) \leq 0$. Bei diesen Begründungen tritt dasselbe Problem wie bei den hinreichenden Gründen auf.

Als letztes Prinzip habe ich das Pareto-Prinzip mit dem Flugzeugentführungs-Dilemma betrachtet. Die Handlung “shoot” ist erlaubt. Das Programm gibt hierfür vier hinreichende und sechzehn notwendige Gründe zurück. Die hinreichenden Gründe lauten: $u(\neg c2) > 0 \wedge \neg[\neg shoot \wedge refrain]\neg c2$, $u(\neg c3) > 0 \wedge \neg[\neg shoot \wedge refrain]\neg c3$, $u(c2) < 0 \wedge \neg c2$ und $u(c3) < 0 \wedge \neg c3$. Hierbei stimmen die ersten beiden Gründe sinngemäß mit der in Kapitel 4.2.5 erarbeiteten Begründung überein. Bei den letzten beiden Gründen stellen sich mir als Nutzer die Fragen: “Was soll das Aussagen?” und “Was genau hat das mit dem Pareto-Prinzip zu tun?”. Diese beiden Gründe beziehen sich auf den zweiten Teil der zweiten Bedingung des Pareto-Prinzips (w_1 hat mindestens eine schlechte Konsequenz, die in w_0 nicht gilt), wobei es auch hier erforderlich ist mit dem Prinzip vertraut zu sein, um diese Begründung verstehen zu können. Die notwendigen Gründe sind:

$$\neg[\neg shoot \wedge refrain]\neg c2 \vee u(\neg c3) > 0 \vee u(c2) < 0 \vee u(c3) < 0 \tag{46}$$

$$\neg[\neg shoot \wedge refrain]\neg c2 \vee \neg[\neg shoot \wedge refrain]\neg c3 \vee u(c2) < 0 \vee u(c3) < 0 \tag{47}$$

$$\neg[\neg shoot \wedge refrain]\neg c2 \vee \neg[\neg shoot \wedge refrain]\neg c3 \vee \neg c2 \vee u(c3) < 0 \tag{48}$$

$$\neg[\neg shoot \wedge refrain]\neg c2 \vee u(\neg c3) > 0 \vee \neg c2 \vee u(c3) < 0 \tag{49}$$

$$\neg[\neg shoot \wedge refrain]\neg c2 \vee u(\neg c3) > 0 \vee \neg c2 \vee \neg c3 \tag{50}$$

$$\neg[\neg shoot \wedge refrain]\neg c2 \vee u(\neg c3) > 0 \vee u(c2) < 0 \vee \neg c3 \tag{51}$$

$$\neg[\neg shoot \wedge refrain]\neg c2 \vee \neg[\neg shoot \wedge refrain]\neg c3 \vee \neg c2 \vee \neg c3 \tag{52}$$

$$\neg[\neg shoot \wedge refrain]\neg c2 \vee \neg[\neg shoot \wedge refrain]\neg c3 \vee u(c2) < 0 \vee \neg c3 \tag{53}$$

$$u(\neg c2) > 0 \vee u(\neg c3) > 0 \vee u(c2) < 0 \vee u(c3) < 0 \tag{54}$$

$$u(\neg c2) > 0 \vee u(\neg c3) > 0 \vee u(c2) < 0 \vee \neg c3 \tag{55}$$

$$u(\neg c2) > 0 \vee \neg[\neg shoot \wedge refrain]\neg c3 \vee u(c2) < 0 \vee \neg c3 \tag{56}$$

$$u(\neg c2) > 0 \vee \neg[\neg shoot \wedge refrain]\neg c3 \vee u(c2) < 0 \vee u(c3) < 0 \tag{57}$$

$$u(\neg c2) > 0 \vee \neg[\neg shoot \wedge refrain]\neg c3 \vee \neg c2 \vee \neg c3 \tag{58}$$

$$u(\neg c2) > 0 \vee \neg[\neg shoot \wedge refrain]\neg c3 \vee \neg c2 \vee u(c3) < 0 \tag{59}$$

$$u(\neg c2) > 0 \vee u(\neg c3) > 0 \vee \neg c2 \vee \neg c3 \tag{60}$$

$$u(\neg c2) > 0 \vee u(\neg c3) > 0 \vee \neg c2 \vee u(c3) < 0 \tag{61}$$

Bei den notwendigen Gründen fällt zuallererst auf, dass es sich, wie bereits bei dem Prinzip der Doppelwirkung, um eine große Anzahl derselben handelt, wodurch eine Begründung nicht leicht zu erfassen ist. Bei all diesen notwendigen Gründen tritt das selbe Problem wie bei den hinreichenden Gründen auf: Ohne Kenntnisse des Prinzips, kann man diese nicht nachvollziehen. Des Weiteren sind es so viele Gründe, dass der Nutzer dadurch überfordert sein könnte und nicht weiß, wie er mit sechzehn einzelnen Begründungen umgehen soll. Auch gibt das Programm keine Gewichtung der Gründe zurück, so dass es unklar ist, ob alle gleichermaßen bedeutend sind, oder ob es Gründe gibt, welche ausschlaggebender sind als andere. Für die verbotene Handlung "refrain" gibt das Programm die hinreichenden Gründe $c1 \wedge [\neg refrain \wedge shoot]c1$, $c1 \wedge \neg[\neg refrain \wedge shoot]c2$ und $c1 \wedge \neg[\neg refrain \wedge shoot]c3$ und die notwendigen Gründe $[\neg refrain \wedge shoot]c1 \vee \neg[\neg refrain \wedge shoot]c2 \vee \neg[\neg refrain \wedge shoot]c3$ und c1 zurück. Weder die hinreichenden noch die notwendigen Gründe stimmen mit den Begründungen basierend auf den in dieser Arbeit vorgestellten Spezifikationen überein (vgl. 5.1). Die hinreichenden Gründe des Programms sind kürzer als die Gründe in Kapitel 5.1 und somit leichter zu überschauen, jedoch fehlen hier wichtige Informationen, wie beispielsweise der Nutzen der Konsequenzen. Bei den notwendigen Gründen fällt der Grund c1 sofort ins Auge, da er nicht aussagekräftig wirkt. Was hat die Konsequenz c1 ("die Passagiere sterben") damit zu tun, dass die Handlung

“refrain” verboten ist? Dieser Grund sagt aus, dass wenn die Passagiere nicht sterben würden, dann die Handlung “refrain” erlaubt wäre. Um die Bedeutung dieses Grundes zu erfassen, ist es jedoch erforderlich, mit dem Prinzip und dem Kausalen Modell vertraut zu sein. Die vom Programm berechneten Gründe des Pareto-Prinzips sind kürzer als die Begründungen basierend auf der Spezifikation in Kapitel 4.2.5, was für die Übersichtlichkeit besser ist, jedoch fehlen hier Informationen über den Nutzen der Konsequenzen, so dass die Gründe ohne das Kausale Modell zu kennen nicht unbedingt leicht nachzuvollziehen sind.

Zusammenfassend kann man sagen, dass dieses Programm Gründe zurückgibt, welche korrekt sind, jedoch nicht unbedingt einfach zu erfassen. Vor allem bei den notwendigen Gründen werden nur Teilformeln zurückgegeben, welche ohne Zusammenhang nicht verständlich sind. Um die Begründungen des Programms verstehen zu können, ist es meiner Ansicht nach erforderlich, dass der Benutzer sowohl das Kausale Modell kennt als auch mit den Bedingungen des ethischen Prinzips vertraut ist. Da diese Ausgabe des Programms nicht wirklich zufriedenstellend ist, stellt sich die Frage, ob es eine umsetzbare Alternative hierzu gibt. Auf diesen Punkt möchte ich im folgenden Kapitel 5.3 eingehen.

5.3 Ansatz auf Basis der Analyse der DNF mit Abstraktion

Da es vor allem bei den notwendigen Gründen ein Problem ist, dass die Teilformeln, welche als Gründe verständlicher wären, auseinandergezogen werden, ist ein weiterer Ansatz für ein Programm zur Generierung von Gründen folgender: Um die Aufsplittung der Teilformeln, welche man gerne als Gründe hätte, zu verhindern, führt man eine Abstraktion dieser Teilformeln durch. Die Teilformeln werden bei der Erstellung der Disjunktiven Normalform aufgesplittet, weshalb man, bevor man diese erstellt,

eine Substitution durchführen muss. Man ersetzt die einzelnen Terme der Formel, welche man als einen Grund behalten möchte, durch eine einzelne Variable. Durch diese Substitution wird verhindert, dass bei dem Erstellen der DNF zusammenhängende Teile eines Grunds aufgesplittet werden. Bevor das Programm den berechneten Grund zurückgibt muss eine Rücksubstitution vorgenommen werden, so dass dem Nutzer anstelle der Variablen eine Formel zurückgegeben wird. Ob durch diese Abstraktion tatsächlich besser verständliche und intuitivere Gründe generiert werden möchte ich in Kapitel 5.3.1 untersuchen.

5.3.1 Analyse und Diskussion des Abstraktion-Ansatzes im Vergleich zu dem DNF-Ansatz ohne Abstraktion

In diesem Fall überspringe ich das Deontologische Prinzip, da es aus nur einer Bedingung besteht und bereits ohne Abstraktion mit der in dieser Arbeit vorgestellten Begründungsspezifikation übereinstimmt (vgl. 5.2.1). Für das Do-No-Harm-Prinzip, welches genau wie das Deontologische Prinzip nur eine Bedingung besitzt, sind bereits Schwierigkeiten bezüglich des Verständnisses aufgefallen. Dies liegt daran, dass die Bedingung des Do-No-Harm-Prinzips eine Implikation enthält, welche aufgesplittet wird. In Kapitel 5.2.1 haben wir beispielsweise den notwendigen Grund $\neg Causes(refrain, c2)$ erhalten. Die Bedingung des Do-No-Harm-Prinzips lautet jedoch $Causes(a, c) \rightarrow u(c) \geq 0$. Würde der notwendige Grund nun auch die Information über den Nutzen erhalten, dann wäre dieser auch ohne Kenntnisse des Kausalen Modells nachvollziehbar. Das abstrahierte Programm gibt für die erlaubte "refrain"-Handlung des Trolley-Dilemmas den hinreichenden Grund $Causes(refrain, c2) \rightarrow u(c2) \geq 0 \wedge Causes(refrain, \neg c1) \rightarrow u(\neg c1)$ und die beiden notwendigen Gründe $Causes(refrain, \neg c1) \rightarrow u(\neg c1) \geq 0$ und $Causes(refrain, c2) \rightarrow u(c2) \geq 0$ zurück. Der hinreichende Grund deckt alle Konsequenzen des Dilemmas ab und ist auch ohne konkretes Wissen über das Kausale Modell und das ethische Prinzip verständlich. Die beiden notwendigen Gründe sind ebenfalls deutlich einfacher

nachzuvollziehen, als die notwendigen Gründe des Programms ohne Abstraktion $\neg Causes(refrain, \neg c1) \vee u(\neg c1) \geq 0$ und $\neg Causes(refrain, c2)$), da diese jetzt auch Informationen über den Nutzen der Konsequenz enthalten. Für die verbotene Handlung "pull" gibt das abstrahierte Programm den Grund $\neg(Causes(pull, c1) \rightarrow u(c1) \geq 0)$ zurück, welcher sowohl hinreichend als auch notwendig ist. Der hinreichende Grund aus Kapitel 5.2.1 war bereits stimmig mit der Begründung basierend auf den vorgestellten Spezifikationen. Die notwendigen Gründe waren jedoch nicht besonders gut nachzuvollziehen. Dies hat sich durch die Abstraktion geändert, da jetzt auch der notwendige Grund sowohl Informationen über den Nutzen der Konsequenz als auch darüber, dass c1 von der Handlung "pull' verursacht wird, enthält.

Für das Do-No-Instrumental-Harm-Prinzip gibt das abstrahierte Programm im Fall der erlaubten Handlung ("pull") des Trolley-Dilemmas den hinreichenden Grund $\neg(Causes(c1, c1) \wedge u(c1) < 0 \wedge u(c1) > 0) \wedge \neg(Causes(c1, \neg c2) \wedge u(c1) < 0 \wedge u(\neg c2) > 0) \wedge \neg(Causes(\neg c2, c1) \wedge u(\neg c2) < 0 \wedge u(c1) > 0) \wedge \neg(Causes(\neg c2, \neg c2) \wedge u(\neg c2) < 0 \wedge u(\neg c2) > 0)$ und die notwendigen Gründe $\neg(Causes(c1, c1) \wedge u(c1) < 0 \wedge u(c1) > 0)$, $\neg(Causes(c1, \neg c2) \wedge u(c1) < 0 \wedge u(\neg c2) > 0)$, $\neg(Causes(\neg c2, c1) \wedge u(\neg c2) < 0 \wedge u(c1) > 0)$ und $\neg(Causes(\neg c2, \neg c2) \wedge u(\neg c2) < 0 \wedge u(\neg c2) > 0)$ zurück. Bei diesen Gründen fällt sofort auf, dass es nicht besonders aussagekräftig oder intuitiv ist zu sagen, dass c1 nicht c1 verursacht. Wenn man hiervon absieht, sind die ausgegebenen Begründungen des Programms auch ohne ein umfangreiches Wissen über das Kausale Modell und das ethische Prinzip nachvollziehbar. Im Fall der verbotenen Handlung "push" des Fatman-Trolley-Dilemmas gibt das Programm genau einen Grund zurück, welcher hinreichend und notwendig ist: $Causes(c1, c2) \wedge u(c1) < 0 \wedge u(c2) > 0$. Dieser Grund ist identisch zu der von mir erarbeiteten Begründung in Kapitel 5.1.

Bei dem Prinzip der Doppwelwirkung hat das abstrahierte Programm, genau wie das Programm ohne Abstraktion, für die erlaubte Handlung "shoot" des Flugzeugentführungs-Dilemmas zwei hinreichende Gründe zurückgegeben (die Teilformeln welche besagen, dass eine Konsequenz nicht sich selbst verursacht sind aus Gründen der Übersicht-

lichkeit nicht mit gelistet):

$$u(shoot) \geq 0 \wedge (I(\neg c2) \wedge u(\neg c2) > 0) \wedge (I(c1) \rightarrow u(c1) \geq 0) \wedge \tag{62}$$
$$(I(\neg c2) \wedge u(\neg c2) \geq 0) \wedge (I(\neg c3) \rightarrow u(\neg c3) \geq 0) \wedge$$
$$\neg(Causes(c1, \neg c2) \wedge u(c1) < 0 \wedge u(\neg c2) > 0) \wedge$$
$$\neg(Causes(c1, \neg c3) \wedge u(c1) < 0 \wedge u(\neg c3) > 0) \wedge$$
$$\neg(Causes(\neg c2, c1) \wedge u(\neg c2) < 0 \wedge u(c1) < 0) \wedge$$
$$\neg(Causes(\neg c2, \neg c3) \wedge u(\neg c2) < 0 \wedge u(\neg c3) > 0) \wedge$$
$$\neg(Causes(\neg c3, c1) \wedge u(\neg c3) < 0 \wedge u(c1) > 0) \wedge$$
$$\neg(Causes(\neg c3, \neg c2) \wedge u(\neg c3) < 0 \wedge u(\neg c2) > 0) \wedge$$
$$u(c1 \wedge \neg c2 \wedge \neg c3) > 0$$
$$u(shoot) \geq 0 \wedge (I(\neg c3) \wedge u(\neg c3) > 0) \wedge (I(c1) \rightarrow u(c1) \geq 0) \wedge \tag{63}$$
$$(I(\neg c2) \rightarrow u(\neg c2) \geq 0) \wedge (I(\neg c3) \rightarrow u(\neg c3) \geq 0) \wedge$$
$$\neg(Causes(c1, \neg c2) \wedge u(c1) < 0 \wedge u(\neg c2) > 0) \wedge$$
$$\neg(Causes(c1, \neg c3) \wedge u(c1) < 0 \wedge u(\neg c3) > 0) \wedge$$
$$\neg(Causes(\neg c2, c1) \wedge u(\neg c2) < 0 \wedge u(c1) > 0) \wedge$$
$$\neg(Causes(\neg c2, \neg c3) \wedge u(\neg c2) < 0 \wedge u(\neg c3) > 0) \wedge$$
$$\neg(Causes(\neg c3, c1) \wedge u(\neg c3) < 0 \wedge u(c1) > 0) \wedge$$
$$\neg(Causes(\neg c3, \neg c2) \wedge u(\neg c3) < 0 \wedge u(\neg c2) > 0) \wedge$$
$$u(c1 \wedge \neg c2 \wedge \neg c3) > 0$$

Die berechneten hinreichenden Gründe lassen keine Fragen mehr über das Kausale Modell offen, sie enthalten eher zu viele Informationen als zu wenige, da sie jede Konsequenz in Zusammenhang mit allen möglichen anderen Konsequenzen betrachten. Hier könnte man im Sinne der Übersichtlichkeit die Teilformeln auslassen, welche einem nicht relevant erscheinen. Als notwendige Begründungen gibt das Programm

Folgende zurück:

$$\neg(Causes(c1, \neg c2) \wedge u(c1) < 0 \wedge u(\neg c2) > 0) \tag{64}$$

$$I(\neg c2) \rightarrow u(\neg c2) \geq 0 \tag{65}$$

$$I(\neg c3) \rightarrow u(\neg c3) \geq 0 \tag{66}$$

$$u(shoot) \geq 0 \tag{67}$$

$$(I(\neg c2) \wedge u(\neg c2) > 0) \vee (I(\neg c3) \wedge u(\neg c3) > 0) \tag{68}$$

$$\neg(Causes(c1, \neg c3) \wedge u(c1) < 0 \wedge u(\neg c3) > 0) \tag{69}$$

$$\neg(Causes(\neg c2, c1) \wedge u(\neg c2) < 0 \wedge u(c1) > 0) \tag{70}$$

$$\neg(Causes(\neg c2, \neg c3) \wedge u(\neg c2) < 0 \wedge u(\neg c3) > 0) \tag{71}$$

$$I(c1) \rightarrow u(c1) \geq 0 \tag{72}$$

$$u(c1 \wedge \neg c2 \wedge \neg c3) > 0 \tag{73}$$

$$\neg(Causes(\neg c3, \neg c2) \wedge u(\neg c3) < 0 \wedge u(\neg c2) > 0) \tag{74}$$

$$\neg(Causes(\neg c3, c1) \wedge u(\neg c3) < 0 \wedge u(c1) > 0 \tag{75}$$

Diese notwendigen Gründe sind besser zu verstehen als die notwendigen Gründe des Programms ohne Abstraktion (vgl. Kapitel 5.2.1). Hier ist es auch möglich die Begründungen zu verstehen, ohne das Kausale Modell und den Nutzen der einzelnen Konsequenzen zu kennen. Für die verboten Handlung "refrain" gibt das Programm die hinreichenden Gründe $\neg(I(c1) \wedge u(c1) > 0) \wedge \neg(I(c2) \wedge u(c2) > 0) \wedge \neg(I(c3) \wedge u(c3) > 0)$ und $\neg u(c1 \wedge c2 \wedge c3)$ und die notwendigen Gründe $\neg(I(c1) \wedge u(c1) > 0) \vee \neg u(c1 \wedge \ c2 \wedge c3) > 0$, $\neg(I(c2) \wedge u(c2) > 0) \vee \neg u(c1 \wedge \ c2 \wedge c3) > 0$ und $\neg(I(c3) \wedge u(c3) > 0) \vee \neg u(c1 \wedge \ c2 \wedge c3) > 0$ zurück. Hier fällt zuerst auf, dass es nur zwei hinreichenden Gründe gibt, anstatt der neun bei dem Programm ohne Abstraktion. Diese sind dadurch leichter zu überschauen. Außerdem sind die Begründungen des abstrahierten Programms, wie bereits im Erlaubt-Fall, auch ohne genauere Kenntnisse des Kausalen Modells zu verstehen, da hier nicht nur über eine Intention geredet wird, sondern immer auch der Nutzen der Konsequenz mitgegeben

wird. Diese abstrahierten Begründungen stimmen sinngemäß mit den Begründungen basierend auf dem erarbeiteten Begründungsvorschlag in 4.2.4 überein (vgl. 5.1).

Als letzten Punkt habe ich noch die Begründungen des Pareto-Prinzips untersucht. Für den Fall, dass die Handlung "shoot" erlaubt ist ergeben sich folgende hinreichende Gründe: $\neg(u(\neg c2) > 0 \rightarrow [\neg shoot \wedge refrain]\neg c2)$, $\neg(u(\neg c3) > 0 \rightarrow [\neg shoot \wedge refrain]\neg c3)$, $\neg(u(c2) < 0 \rightarrow c2)$ und $\neg(u(c3) < 0 \rightarrow c3)$ und der notwendige Grund $\neg(u(\neg c2) > 0 \rightarrow [\neg shoot \wedge refrain]\neg c2) \vee \neg(u(\neg c3) > 0 \rightarrow [\neg shoot \wedge refrain]\neg c3) \vee \neg(u(c2) < 0 \rightarrow c2) \vee \neg(u(c3) < 0 \rightarrow c3)$. Die hinreichenden Gründe stimmen mit denen des nicht-abstrahierten Programms überein. Für die notwendigen Begründungen war am auffälligsten, dass es sechzehn Stück gab und es dadurch schwer war, alle diese Gründe zu überblicken oder zu gewichten. Durch die Abstraktion ergibt sich nur noch ein einziger notwendiger Grund, welcher deutlich leichter zu erfassen ist als die sechzehn Gründe zuvor. Für die verbotene Handlung "refrain" gibt das Programm die hinreichenden Gründe $(u(c1) < 0 \rightarrow c1) \wedge (u(\neg c1) > 0 \rightarrow [\neg refrain \wedge shoot]c1)$, $(u(c1) < 0 \rightarrow c1) \wedge (u(c2) < 0 \rightarrow \neg[\neg refrain \wedge shoot]c2)$ und $(u(c1) < 0 \rightarrow c1) \wedge (u(c3) < 0 \rightarrow \neg[\neg refrain \wedge shoot]c3)$ und die beiden notwendigen Gründe $(u(\neg c1) > 0 \rightarrow [\neg refrain \wedge shoot]c1) \vee (u(c2) < 0 \rightarrow \neg[\neg refrain \wedge shoot]c2) \vee (u(c3) < 0 \rightarrow \neg[\neg refrain \wedge shoot]c3)$ und $u(c1) < 0 \rightarrow c1$ zurück. Diese Begründungen sind, sowohl im hinreichenden als auch im notwendigen Fall, ohne Kenntnisse des Kausalen Modells einfacher zu verstehen als die Gründe des Programms ohne die Abstraktion (vgl. 5.2.1), wobei der letzte notwendige Grund $u(c1) < 0 \rightarrow c1$ auch mit der Information über den Nutzen von c1 nicht ohne ausreichende Kenntnisse über die Dominanz-Bedingungen verständlich ist und selbst mit diesem Wissen ist es nicht auf den ersten Blick zu erfassen, dass sich dieser Grund auf die dritte Dominanz-Bedingung (alle schlechten Konsequenzen von w_0 sind auch schlechte Konsequenzen von w_1) bezieht. Dies lässt sich nur erkennen, wenn man ebenfalls mit dem Kausalen Modell vertraut ist, da sich der Grund sonst auch auf die zweite Bedingung beziehen könnte (siehe Def. 10).

Zusammenfassend kann zu dem Ansatz auf Basis der DNF mit Abstraktion gesagt werden, dass die Gründe, die von dem Programm generiert werden, in den meisten Fällen leichter zu verstehen sind als die Begründungen, welche das Programm ohne die Abstraktion zurückgibt. Dies liegt daran, dass durch die Abstraktion die Teilformeln beim Erstellen der DNF nicht mehr aufgesplittet werden und dadurch Informationen, welche aus Gründen der Verständlichkeit zusammen genannt werden sollten, nicht mehr getrennt werden können. Auch hat sich die Anzahl der Gründe teilweise reduziert, wodurch diese überschaubarer geworden sind.

6 Diskussion

Abschließend möchte ich noch einmal alle in dieser Arbeit präsentierten Teilergebnisse miteinander verknüpfen. Die in Kapitel 4.2 vorgestellten Spezifikationen zur Begründung moralischer Dilemmata unter verschiedenen ethischen Prinzipien haben sich nicht wie erhofft in dem HERA-System implementieren lassen, da es für jedes Prinzip nötig wäre, die Begründungen zu definieren. Dadurch würde ein zusätzlicher Aufwand bei dem Einfügen neuer ethischer Prinzipien entstehen und man müsste für jedes Prinzip zwei Definitionen implementieren statt nur einer. Deshalb wurde ein Prinzipien-unabhängiger Vorschlag für eine alternative Implementierung der Begründungen über die Disjunktive Normalform angebracht. Hierbei wurden die Begründungen direkt aus den Formeln des Prinzips extrahiert. Dabei hat sich herausgestellt, dass die generierten Gründe zwar korrekt waren, jedoch in fast allen Fällen nur dann verständlich, wenn ausreichende Kenntnisse über das Kausale Modell und die Bedingungen der ethischen Prinzipien vorhanden sind. Dies ist darauf zurückzuführen, dass Teilformeln, welche für das Verständnis zusammen betrachtet werden müssen, bei der Erstellung der DNF aufgesplittet werden. Auch ist bei jedem ethischen Prinzip (ausgenommen dem Deontologischen Prinzip) der Fall aufgetreten, dass ein notwendiger Grund nicht verständlich war, da er ohne Zusammenhang genannt worden ist und somit weder Informationen darüber enthalten waren, ob es sich um eine gute oder schlechte Konsequenz handelt, noch was der genannte Grund überhaupt mit dem Prinzip zu tun hat. Somit ist auch diese Alternative nicht wirklich zufriedenstellend, da es keine Anforderung an den Nutzer sein sollte, sich zuerst mit den Kausalen

Modellen und den ethischen Prinzipien zu befassen, wenn er eine Erklärung für die Entscheidung des Agenten anfordert. Wenn man die in Kapitel 5.3 vorgestellte Abstraktion der Teilformeln durchführt, dann erhält man für jedes der betrachteten Prinzipien eine Begründung die auch mit geringen Kenntnissen des Kausalen Modells und des ethischen Prinzips verständlich ist. Dadurch, dass die Teilformeln nicht mehr aufgesplittet werden, enthalten die zurückgegebenen Gründe des Programms genügend Informationen über die Konsequenzen, wodurch der Zusammenhang der Bedingungen wieder hergestellt ist. Diese Abstraktion ist jedoch auch mit einem gewissen Aufwand verbunden, da es notwendig ist sie bei der Implementierung jedes Prinzips durchzuführen. Der Aufwand der Abstraktion ist meiner Ansicht nach jedoch geringer als die Umsetzung der Begründungsspezifikationen aus Kapitel 4.2, da es hier nicht notwendig ist, sich Gedanken über die Gründe zu machen. Außerdem ist der Mehraufwand nicht so gravierend, da es für die Einführung eines neuen Prinzips erforderlich ist, die Bedingungen des Prinzips in Form von logischen Formeln zu implementieren und die Abstraktion bereits bei dieser Implementierung durchgeführt werden kann.

Ein entscheidender Kritikpunkt an dem Ansatz der Generierung von Gründen mittels der DNF ist jedoch derjenige, dass die Erstellung der DNF mit einem Modell, welches mehrere Konsequenzen und Bedingungen enthält, sehr viel Zeit in Anspruch nimmt. Bereits bei der Erzeugung der DNF für das Flugzeugentführungs-Dilemma, welches nur drei Konsequenzen enthält, unter Anwendung des Prinzips der Doppelwirkung (5 Bedingungen) benötigt das Programm 4 Minuten und 39 Sekunden um eine DNF zu erzeugen (auf einem Rechner mit Intel Core i5-5200U CPU @ 2 x 2.20 GHz). Somit lässt sich dieser Ansatz nicht in jeder Situation für komplexere moralische Dilemmata verwenden, da die Erstellung der DNF zu viel Zeit in Anspruch nimmt, weshalb dieser Ansatz nicht realzeitfähig ist. Würde beispielsweise ein selbstfahrendes Fahrzeug die Zustimmung eines Menschen benötigen, um Entscheidungen, welche Menschenleben beeinflussen, treffen zu können, wäre es nicht möglich, mit diesem Ansatz eine Erklärung für die Entscheidung der Künstlichen Intelligenz zu generieren, da die

Entscheidung, welche Handlung ausgeführt werden soll, innerhalb von Sekunden getroffen werden muss. Wenn es jedoch nicht entscheidend ist, dass sofort eine Erklärung erzeugt wird, dann ist es mit diesem Ansatz möglich die Begründungen zu erzeugen und für einen späteren Aufruf zu speichern.

7 Fazit

Das Thema dieser Arbeit ist die Berechnung von Gründen unter verschiedenen ethischen Prinzipien. Die Begründungen von Entscheidungen eines Agenten sind von enormer Wichtigkeit, da nur so Vertrauen zwischen dem Nutzer und dem Agenten gewährleistet werden kann. Da heutzutage Roboter und Künstliche Intelligenz immer weiter verbreitet sind, werden Agenten auch zunehmend mit moralischen Situationen konfrontiert, in welchen sie Entscheidungen treffen müssen. Ohne eine leicht verständliche Begründung kann keine Transparenz in den Entscheidungsprozess des Agenten gegeben werden und somit auch kein Vertrauen entstehen. Aus diesem Grund muss auf dem Gebiet der Generierung von Gründen noch einiges an Forschung betrieben werden, da sich Roboter und Künstliche Intelligenzen sonst nur beschränkt einsetzen lassen können. In dieser Arbeit wurde ein auf einer Disjunktiven Normalform basierendes Verfahren entwickelt, um Prinzipien-übergreifend Begründungen zu berechnen. Diese Gründe haben sich jedoch zu sehr von den Spezifikationen der Begründungen unterschieden und waren ohne weitere Informationen nicht verständlich. Durch eine Erweiterung des Verfahrens mittels Abstraktion konnte eine Annäherung an die Spezifikationen erzeugt werden, da durch die Abstraktion dafür gesorgt wird, dass Teilformeln, welche sinngemäß zusammengehören, nicht bei der Erstellung der DNF getrennt werden können. Durch die Erstellung der DNF ist das Programm jedoch nicht realzeitfähig. Deshalb ist es notwendig, dass auf dem Gebiet der Begründung nach weiteren Alternativen gesucht wird, welche sich auch in realer Zeit umsetzen lassen.

Literaturverzeichnis

[1] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, and I. Rahwan, "The moral machine experiment," *Nature*, vol. 563, pp. 59–64, 2018.

[2] F. Lindner and M. M. Bentzen, "The hybrid ethical reasoning agent immanuel," in *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '17, pp. 187–188, 2017.

[3] D. Dannenhauer, M. W. Floyd, D. Magazzeni, and D. W. Aha, "Explaining rebel behavior in goal reasoning agents," in *ICAPS-18 Workshop on Explainable Planning*, pp. 12–18, 2018.

[4] R. Borgo, M. Cashmore, and D. Magazzeni, "Towards providing explanations for AI planner decisions," in *Proceedings of the IJCAI/ECAI 2018 Workshop on Explainable Artificial Intelligence (XAI)*, pp. 11–17, 2018.

[5] P. Langley, B. Meadows, M. Sridharan, and D. Choi, "Explainable agency for intelligent autonomous systems," in *Proceedings of the Twenty-Ninth Annual Conference on Innovative Applications of Artificial Intelligence*, pp. 4762–4763, 2017.

[6] M. Fox, D. Long, and D. Magazzeni, "Explainable planning," in *Proceedings of IJCAI-17 Workshop on Explainable Artificial Intelligence*, pp. 24–30, 2017.

[7] C. Russell, "Efficient search for diverse coherent explanations," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pp. 20–28, 2019.

[8] D. V. Pynadath, N. Wang, and M. J. Barnes, "Transparency communication for reinforcement learning in human-robot interaction," in *Proceedings of the IJCAI/ECAI 2018 Workshop on Explainable Artificial Intelligence (XAI)*, pp. 123–129, 2018.

[9] J. van der Waa, J. van Diggelen, K. van den Bosch, and M. A. Neerincx, "Contrastive explanations for reinforcement learning in terms of expected consequences," in *Proceedings of the IJCAI/ECAI 2018 Workshop on Explainable Artificial Intelligence (XAI)*, pp. 165–170, 2018.

[10] A. Shih, A. Choi, and A. Darwiche, "A symbolic approach to explaining bayesian network classifiers," in *Proceedings of the IJCAI/ECAI 2018 Workshop on Explainable Artificial Intelligence (XAI)*, pp. 144–150, 2018.

[11] F. Lindner, M. M. Bentzen, and B. Nebel, "The HERA approach to morally competent robots," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 6991–6997, 2017.

[12] B. Kuhnert, F. Lindner, M. Bentzen, and M. Ragni, "Perceived difficulty of moral dilemmas depends on their causal structure: A formal model and preliminary results," in *Proceedings of the 39th Annual Meeting of the Cognitive Science Society (CogSci2017)*, pp. 2494–2499, 2017.

[13] "The HERA project." `http://www.hera-project.com/principles/`. eingesehen am 13.01.2019.

[14] F. Ricken, *Allgemeine Ethik*, vol. 4. Stuttgart: Kohlhammer, 2003.

[15] J. Halpern, *Actual Causality*. London: The MIT Press, 2016.