university of groningen / faculty of science and engineering

Two Parsimonious Approaches to Deontic Logic for Machine Ethics

Bachelor's Project Thesis

Jan van Houten, j.j.van.houten@student.rug.nl,
Supervisors: Prof Dr L.C. Verbrugge & Prof Dr D. Grossi

Abstract: The field of machine ethics deals with the ethical decision making of artificial agents. This thesis presents a preliminary evaluation of two applications of deontic logic to this type of decision making. The approaches are called parsimonious because they are not based on a specifically deontic logic but rather on normal default logic, combined, respectively, with the method of process trees and with Boutilier's (1990) translation of to the common modal system S4 and the method of semantic tableaux. The viability of these parsimonious approaches is explored by applying them to default theories corresponding to diverse ethical theories. This leads to reflections on the limitations of the approaches as well as on the advantages and disadvantages of these approaches relative to each other. The process-tree approach in particular seems to be viable, although it is noted that priorities are necessary to decide in case of conflict between defaults.

1 Introduction

The field of Artificial Intelligence (AI) has important ethical implications, especially where the development of more and more intelligent systems is concerned. At least two distinct subdisciplines of ethics have spawned from its investigations.

First, one can explore ethical guidelines for the design and use of AI. Some modern technologies, such as military drones, have the potential to behave in highly unethical ways, and these have to be designed and/or employed very carefully (see e.g. Byrne (2018), or for a more machine-ethical approach, Simonite (2009)). In addition, while technologies such as sexual orientation-detecting software may be less directly dangerous, they can still be used in highly unethical ways (Wang and Kosinki, 2018).

Second, as AI advances, on some level it may be appropriate to ascribe agency not just to its designers, but also to the AI itself. At the very least, regardless of the psychological or metaphysical questions of agency - "Are these systems actually agents? Does agency not require something that, as far as we now know, is uniquely human?" - it seems likely that many would ascribe some form of human-like moral agency to AIs in daily life, as a result of anthropomorphism.[1] It is not strange to suppose that this perception of moral agency will grow stronger and more prominent as technology becomes more intelligent and, especially, as it takes on functions that used to be uniquely human, such as driving a car.[2]

Let us consider, for example, the following situation:

> A fit and sober human driver sees someone crossing the road some way in front of him. Should the driver brake, as one would usually do, or drive on and kill the pedestrian?

Clearly, if the driver chooses to drive on, their choice is highly unethical - and the designer of the

[1] Note that anthropomorphism depends not just on the physical shape of an AI but also (perhaps primarily) on AI-human interaction (Choi and Kim, 2009; Damiano and Dumouchel, 2018).

[2] Of course, general agency does not always amount to moral agency. Wooldridge (2009), for example, defines an agent as "a computer system that is *situated* in some *environment*, and that is capable of *autonomous action* in this environment to meet its delegated objectives" (p. 21), while Eshleman (2016) defines a moral agent simply as "an agent open to responsibility ascriptions". Thus, many or all artificial agents are perhaps not truly moral agents. However, whether these agents would be *perceived* as moral agents is a different question - the answer to which, I think, is yes.

car would seem to have nothing to do with it.[3] Now consider a self-driving car faced with the same choice.[4] If this car drives on, there are two possible perspectives to take. On the one hand, the designers can be said to have behaved unethically in programming the car as they did. Yet on the other hand, perceiving the car as an agent, we could also say that the car *itself* behaved unethically in making this choice.

It is from this second perspective that the field of machine ethics emerges in a natural way. This field deals with morality from the perspective of artificial agents (Moor, 2006).

The apparent naturalness of the ascription of moral agency to increasingly intelligent agents suggests to me that we will in the future expect certain artificial agents to behave ethically. If this is the case, then perhaps we should design these agents in a way that respects our perception of them and that attempts to ensure that these expectations will not be violated.

How can this be done? There are, undoubtedly, myriad answers to this question. In this thesis, I will explore the viability of an approach based on parsimonious deontic logic.[5] Formal logic has at least three distinct advantages over other approaches to the problem of machine ethics (Bringsjord, Arkoudas, and Bello, 2006). First, logic seems to be a natural representation for ethical reasoning. Second, logic has been effective in AI in the past. But most importantly, because of the rationality and consequent predictability of logic, using automated formal proofs as a method for decision making might be very effective in establishing trust. As artificial agents (trustees) become more autonomous, human trustors become more vulnerable to their actions (Tavani, 2015). We should only accept this vulnerability if the agents are provably trustworthy, which is why logic is an intuitive solution.

Deontic logic is a species of formal logic concerned with norms specifying what ought to be and/or what agents ought to do (von Wright, 1951; Horty, 2001). Often, the operator "$\bigcirc$" is used to indicate a normative proposition in this context. However, out of parsimony considerations, the focus in this thesis will be on default logic, which does not require special operators for an intuitive normative interpretation. Default logic has previously been suggested as an option in machine ethics, and for good reason (Powers, 2006). Process trees, a proof method specific to this type of logic, (Antoniou, 1997), may be used for deontic reasoning. In addition, default logic can be reduced to the common modal system S4, as shown by Boutilier (1990), after which the familiar method of semantic tableaux (Priest, 2008) may be used for deontic reasoning. In this thesis, I will explore the viability of these process-tree and tableau approaches in the context of machine ethics. These approaches are preferable to more complex approaches in that they are parsimonious with respect to formal machinery, and I will therefore refer to them as "parsimonious". The question is: how viable do these parsimonious approaches seem to be in the context of machine ethics, in general as well as compared to each other?[6] I have set out to answer this question and to provide a proof of concept in the form of a piece of software automating the tableau approach.

In the next section, I will consider several ethical theories. These ethical theories will form the background of the present analysis, as they have the advantage of presenting (or at least attempting to present) a unified and general picture of morality. Also, because different ethical theories represent different perspectives on ethical decision making, I hope to cover a broad moral spectrum by exploring a number of these theories. It may be, for example, that the choice of a specific ethical framework leads, in turn, to specific considerations regarding the viability of the approaches.

After discussing ethical theories in Section 2, I will use Section 3 to introduce the parsimonious approaches, contrasting them with each other and with an alternative logic-based approach. In Section 4, then, I will apply the parsimonious approaches to the ethical theories discussed in Section 2. And finally, in Section 5, I will evaluate the apparent viability of the parsimonious approaches in the light of these applications.

[3] Note, however, that this situation would not have occurred if cars did not exist in the first place. Thus, technology may still play an enabling role, even in such situations.

[4] I refer to an everyday conception of this concept ("choice"), not a philosophically grounded one.

[5] I will deal only with explicit ethical agents (Moor, 2006), that is, agents which have an explicit representation of ethics and can operate on it.

[6] As testing viability directly would, to the extent that it is even possible, be a very challenging endeavour at this point, a preliminary analysis will have to do. Hence my modest phrasing of the question.

2 Ethics

There are several approaches to setting up an ethical AI. We could, for example, simply add behavioural rules whenever we think them necessary. Then, for an AI system with limited capacity and applicability, such as a robot driving around an office with supplies, it will often suffice to provide just a few behavioural rules that can be considered ethical. In fact, many such systems may never enter into ethically relevant situations at all, and so no behavioural rule may be necessary.

However, more general AI systems, as well as systems more likely to enter into ethically relevant situations, will quickly require large numbers of behavioural rules to (approximately) guarantee ethical behaviour. The complexity and the system-specificity of such ethical rule bases are likely to impede our intuitive understanding of, and hence trust in, such systems. In addition, as these systems become even more complex and - perhaps in functioning rather than in cognition - more "human", a finite set of rules will often not suffice anymore. Ideally, such systems will have at their disposal a rule so general that they can deduce situation-specific guidelines from it in every situation. This sort of general rule, which I will call a rule generator, is known as an ethical theory (Schafer-Landau, 2012).

The parsimonious approaches presented in this thesis are not meant to be used for modelling rule generators. As I will illustrate in Section 4, rule generators have to deal with major challenges which the parsimonious approaches are not equipped to overcome. However, the parsimonious approaches *can* be used for reasoning using the rules generated by various ethical theories. In the remainder of this section, I will briefly discuss a number of well-known ethical theories and, prospectively, evaluate them in the context of machine ethics and deontic logic. The theories to be discussed are consequentialism (specifically, act utilitarianism and rule utilitarianism), deontology (specifically, Kantian ethics), and virtue theory.

2.1 Consequentialism

A popular thought in ethics is that what matters in evaluating an action are the action's (expected) consequences. Some would say that the rightness and wrongness of an action depend solely on its consequences. This is the basic premise of consequentialism (Shaw, 2014).

There are many conceivable forms of consequentialism. In this thesis, I will focus only on the two best-known theories: act utilitarianism and rule utilitarianism. Act utilitarianism has been succinctly defined by Hursthouse (1991), as follows:

A1 An action is right iff it promotes the best consequences.

A2 The best consequences are those in which happiness is maximized.[7]

In the same vein, we can define rule utilitarianism as follows (see e.g. Schafer-Landau (2012)):

R1 An action is right iff it is in accordance with the set of rules that, in the long term, promotes the best consequences.

R2 The best consequences are those in which happiness is maximized.

The main task for utilitarians is to define "happiness". In the following sections, I will assume that more lives saved means more happiness, and that all lives are worth an equal amount of happiness. This is a crude but popular approximation, and for this thesis, it will do.

While consequentialism is computationally pleasing, it can also be quite scary, since in certain cases it supports sacrificing innocent persons to achieve gains for others (e.g. Schafer-Landau (2012)). I think the possibility of being sacrificed by robots in this manner is enough to break down any trust humans may have in them. But this does not have to be the end of consequentialist machine ethics. First, because rule-utilitarian robots would behave consistently over time, in accordance with a set of rules, they would be more predictable than act-utilitarian ones. Therefore, they may also be trusted more. Further, even if consequentialism cannot serve as a basis for robot-human interaction, Grau (2006) has suggested that robots can still be consequentialist with respect to one another. Therefore, consequentialism should still be taken seriously as an option in machine ethics. I will explore the viability of the parsimonious approaches with respect to act utilitarianism and rule utilitarianism in Sections 4.1 and 4.2, respectively.

[7] Hursthouse, p. 225.

2.2 Deontology

Hursthouse (1991) defines the premise of deontology as follows:

D1 An action is right iff it is in accordance with a moral rule or principle.[8]

The deontologist's challenge, then, is to determine what defines a moral rule or principle.

In this thesis, I will limit myself to a Kantian interpretation, and specifically, to the first formulation of Kant's Categorical Imperative. Powers (2006) has stressed the usefulness of the first formulation in machine ethics as a way of testing moral principles (and thus as a rule generator). This formulation, which specifies that a rule by which one acts must be universalizable (i.e. not self-undermining), might tell us that it would not be right for artificial agents to lie to humans - even in extreme cases, as illustrated in Section 4.3.

2.3 Virtue theory

Following Hursthouse (1991), virtue theory's main premises are:

V1 An action is right iff it is what a virtuous agent, i.e. one who has and exercises the virtues, would characteristically do in the circumstances.

V2 A virtue is a character trait a human being needs to flourish or live well.[9]

One of the main criticisms levelled against virtue ethics is that it is vague and does not provide concrete ethical guidance. To a certain extent, this seems unavoidable. Virtue ethics is built upon an understanding of human character and the "good life" that is not easily captured in words. It may therefore be very difficult to represent its principles using logic.[10]

Nevertheless, the theory does sound very promising, for a number of reasons. Virtue ethics aims to promote a character from which good actions flow, and this seems to be a much more flexible approach to morality than both the consequentialist and deontological approaches. In addition, while the theory is not rule-based, it is very possible to extract behavioural rules from it (Hursthouse, 1991).

However, to merely program such generated rules, represented by sentences such as "If you do X, you act courageously" and "Ideally, you should act courageously", does not come close to capturing the depth of the character by which they are generated. That is, while Hursthouse focuses mostly on right action, virtue theory seems to require something more. As Schafer-Landau (2012) puts it: "Virtuous people are (...) defined not just by their deeds but also by their inner life." (p. 259). This "inner life" extends beyond intentions to a broader notion of character, and it is an open question how this notion can be adequately represented.[11]

Thus, virtue theory sounds promising, but there are difficult issues to deal with, including the generation of information relevant to the parsimonious approaches and the question whether ethics can really be limited to the generation of logical facts and rules. The parsimonious approaches will be applied in combination with virtue theory in Section 4.4.

3 Logic

The main goals of the present section are to explore the concept of viability and to discuss two approaches to moral rules that are based on default logic: the process-tree approach, which makes use of a proof technique specific to default logic, and the tableau approach, which makes use of a translation to modal logic along with the more general proof method of semantic tableaux. I will start by listing a number of considerations that are important in determining the viability of a logic-based approach in machine ethics. Subsequently, I will briefly discuss a branch of deontic logic proposed

[8] Hursthouse, p. 224.

[9] Hursthouse, pp. 225-226.

[10] Even if we ignore the difficulty of extending the concept of "human character" to artificial intelligence.

[11] One could even argue that virtue ethics does not necessarily involve deontic reasoning at all. Instead, what is required is a character from which ethical motivations and actions flow naturally; the moral agent acts as they do not because they have, after careful deliberation, concluded that their actions are morally right, but because they are inherently and automatically motivated to do as they do. At least some virtue-ethical robots may thus fall in the category of implicit ethical agents (Moor, 2006), whose behaviour can be seen as ethical even though they do not have an explicit representation of ethics. Moor himself has suggested that such an agent "has, to a limited extent, virtues" (p. 19).

by Horty (2001) and point out its difficulties. Then, default logic and the process-tree approach will be introduced, after which modal logic will be briefly discussed, along with the translation presented by Boutilier (1990).

To simplify the discussion, I will limit myself to propositional logic with finitely many atoms. Thus, the parsimonious approaches are always based on a propositional language $L(P)$, which itself is based on a finite set P of propositional atoms (for example, $P = \{p_1, ..., p_n\}$). $L(P)$ is defined as the smallest set satisfying:

i If $A \in P$, then $A \in L(P)$.

ii If $A \in L(P)$ and $B \in L(P)$, then $\neg A \in L(P)$, $A \wedge B \in L(P)$, $A \vee B \in L(P)$, $A \supset \in L(P)$, and $A \equiv B \in L(P)$.

The semantics for the operators are the same as in classical logic.

3.1 Viability

Because of the preliminary nature of this thesis, the term "viability" is necessarily vague. However, there are some general criteria of importance to determining the viability of a logic-based approach toward machine ethics. In this thesis, I will concern myself with the following three:

- **Accessibility**: the extent to which the approach is easy to understand and follow.[12] As mentioned, trust is likely to be an important factor in the context of machine-ethics, and logic lends itself to trustworthy applications because of the rationality of logical reasoning. On top of that, however, it would be preferable if the logical mechanisms via which machine-ethical agents make their ethical decisions are not just theoretically predictable but also easy to understand and follow.

  Defined like this, accessibility has many aspects. For example, a proof method may be intuitively understandable, it may be widely known, or it may be easy to perform or check manually. All such considerations factor in the judgment of accessibility.

- **Versatility**: the extent to which the approach is compatible with different ethical points of view. There is no universal or binding verdict on machine-ethical matters (yet). Thus, and given that different situations may mean different ethical solutions, the ideal approach should be compatible with multiple viewpoints so that it can be widely applied without problems.

- **Efficiency**: the extent to which the approach can finish its job in a reasonable amount of time, steps, and space (in memory and on-screen). In this thesis, I will only put forth highly preliminary, comparative judgments of efficiency.

3.2 A deontic logic

One approach to deontic logic, advocated by e.g. Horty (2001), is to define the logic in such a way that normative statements can always be deduced from a model and never be entered as premises independent from the state of the world. The appeal of such an account is clear, considering that the alternative approach is to specify ethical rules from the ground up, taking them as premises for a process of deduction. What ought to be, on an account such as Horty's, does not depend on any moral principles we entered, but it can be directly deduced from the state of the world alone, using accepted moral principles as a built-in inference mechanism. However, this does mean that the moral principles need to be built into the logic, and therefore, this approach will be fundamentally bound to a specific moral theory.

I take Horty's (2001) work as an example case. Rather than focusing on what ought to be, as my approach does, Horty's logic focuses directly on what agents ought to do. His approach is fundamentally consequentialist, in that values of normative importance are attached to histories (i.e. sequences of moments) and the definitions of deontic operators are based on these values. This involves a good deal of complexity. For example, consider Horty's general definition of the common operator $\bigcirc$:

> "$M, m/h \vDash \bigcirc A$ if and only if there is some history $h' \in H_m$ such that (1) $M, m/h' \vDash A$ and (2) $M, m/h'' \vDash A$ for each history $h'' \in H_m$ such that $Value_m(h') \leq Value_m(h'')$."

[12] Unfortunately, "accessibility" is also a common term in the context of modal logic (see Section 3.5), where it has a very different meaning. In both cases, however, I think the word is the most suitable choice.

(p. 37)

That is, A should hold at some moment iff A holds in all the histories possible in that moment that are maximally valuable. A model M, as mentioned in this definition, consists of a *Tree* of moments, an ordering $<$ on those moments, a set *Agent* of agents, a set *Choice* of sets of moment-history pairs, a function *Value* attaching values to histories, and a valuation function v. In order to properly have utilitarianism dictate the truth value of deontic operators, however, Horty is forced to complicate his logic further. What is more, because of the high specificity of his logic, Horty is forced to define separate operators ($\odot$ and $\oplus$) for the two forms of act utilitarianism he considers.

It is easy to write out utilitarian statements in Horty's logic, as each of his deontic operators carries the full meaning of "ought" in a specific utilitarian context. For a simple deontic statement, there is a reasonably intuitive way to find its truth value by checking the model (and in particular, the values of different histories). However, Horty's approach comes with at least two major drawbacks:

- It is not **accessible**: models are hefty, writing them out is a time-consuming process, and there is (to my knowledge) no quick and intuitive proof method for his logic;
- It is not **versatile**: Horty's approach is only suitable for modelling specific types of consequentialist ethics.

By contrast, the approaches I put forth in this thesis are highly accessible and versatile, and have their roots in an intuitive branch of logic.

3.3 Default logic

In the real world, not all relevant information is always available, which means we cannot draw conclusions with absolute certainty. However, there is still a degree of rationality available to us. Instead of using absolute rules, we use defeasible inference patterns such as "By default/Typically, A implies B; A; therefore, B." These inferences are defeasible not because they are irrational, but because there are exceptional (non-default) cases in which they do not apply. This "default reasoning" has been formalized in a logic known as default logic (Reiter, 1980).

Default logic deals with default theories. A default theory T is a tuple (W, D), where W is a set of initial facts, all taken to be true not merely by default but absolutely, and D is a set of (defeasible) inference rules, called defaults. Each default has a prerequisite A (which needs to be true), a consequent B, and one or more justifications (which should be consistent with what is known or assumed to be true). I will confine myself to defaults with a single justification (C):

$$\frac{A : C}{B}$$

A default is called applicable iff its prerequisite is true and it is consistent to assume the truth of its justification (Antoniou, 1997).

In this thesis, I will limit myself to normal default theories. In such theories, each default's justification is the same as its conclusion (Reiter, 1980). Defaults can thus, and will henceforth, be represented in the form $A \Rightarrow B$.

In short, the default theories I am concerned with are structures $T = (W, D)$ based on a language $L(P)$ as defined at the start of Section 3, where:

- W is a set of initial facts A (where $A \in L(P)$).
- D is a set of normal defaults $A \Rightarrow B$ (where $A \in L(P)$ and $B \in L(P)$).

In addition, I will use a deontic "ought-to-be" interpretation of defaults. That is, $A \Rightarrow B$ is no longer interpreted as "Typically, A implies B" but instead as "Ideally, A implies B". This leads to an intuitive deontic interpretation of a default as a moral requirement: "If A holds, then it ought to be that B holds."

3.4 Process trees

Process trees are used as a proof method particular to default logic (Antoniou, 1997). A process tree starts with a node containing the "In-set" $Th(W)$[13] and an empty "Out-set". The inference proceeds by applying applicable defaults (corresponding to edges) and writing out the resulting In- and Out-sets (corresponding to nodes). Each time a normal default $A \Rightarrow B$ is applied, B is added to the set S in

[13] $Th(S)$ denotes the deductive closure of a set S of formulas.

the In-set $Th(S)$, and $\neg B$ is added to the Out-set. If the In- and Out-sets overlap at any point, the process (sequence of defaults) corresponding to the branch fails; if not, it is successful. If no defaults are applicable anymore, the process corresponding to the branch is closed. The In-set at the end node of a branch corresponding to a closed and successful process is called an extension. Thus, the project of a process tree is to find all extensions of a default theory. An example of a process tree is included in Appendix A.

Under an ideality interpretation of defaults, we can use the process-tree approach to identify one or more ideal extensions, following this definition:

> An extension E of a theory $T = (W, D)$ is *ideal* iff for every default $A \Rightarrow B \in D$ such that $A \in E$, also $B \in E$.

That is, an extension is ideal iff it satisfies all relevant moral requirements. Although every normal default theory has at least one extension (Reiter, 1980), it might be that none of these extensions satisfies this ideality requirement. In such a case, there is no ideal extension, but we can still find "preferred" or "maximally ideal" extensions. These are extensions which satisfy the maximum possible number of relevant moral requirements:

> Let $\#F(E)$ be the number of failed requirements for an extension E of a theory $T = (W, D)$, i.e. the number of defaults $A \Rightarrow B \in D$ such that $A \in E$ and $B \notin E$. An extension E of a theory $T = (W, D)$ is *maximally ideal* or *preferred* iff for all extension E' of that theory, $\#F(E) \leq \#F(E')$.

Note that the ideal extensions, if they exist, are always the preferred extensions.

Process trees are an intuitive proof method, and they are quite efficient in the amount of information they present, especially in the case of smaller default theories. However, as I mentioned, they are also particular to default logic. Generally speaking, it is less demanding to be proficient in one general proof method than in many specific ones. Therefore, if a more general proof method is also applicable, that method may be more accessible. There happens to be such a more general proof method, and it is often used in the domain of modal logic.

3.5 Modal logic

Modal logic extends classical logic with axioms concerning the box and diamond operators ($\Box$ and $\Diamond$). Accordingly, in the context of modal logic, the definition of $L(P)$ presented early in Section 3 is extended with the following requirement:

iii If $A \in L(P)$, then $\Box A \in L(P)$ and $\Diamond A \in L(P)$.

The modal system S4, with which I am concerned in this thesis, extends the set of axioms further to include those of transitivity and reflexivity (Lewis and Langford, 1959).

A basic modal logic interpretation is defined as $\langle W, R, v\rangle$, with W the set of worlds, R the accessibility relation, and v the valuation function. I will adopt this interpretation, but I will change the set W to a set S of states, not just because I believe this to be more natural in the case of (practical) moral reasoning, but also in order to avoid confusion with the set W of initial facts in a default theory. Under the standard semantics, sRs' is interpreted as "state s' is accessible from state s" (Priest, 2008).

Modal logic has been used for many purposes, including some forms of deontic logic (McNamara, 2014). Further, I think the proof method of semantic tableaux, which is easy to apply to modal logic, is highly accessible (Priest, 2008). Because modal logic is widely used, and because the tableau method is very intuitive, automatic theorem proving based on this approach has the advantage of being directly understandable to many logicians. Also, like default logic, modal logic is not built to cater for a specific ethical theory, and we may be able to use this generality as versatility with respect to ethical theories.

3.6 From defaults to a modality

Boutilier (1990) has shown that it is possible to reduce what he calls a conditional logic of normality (in our case, normal default logic) to S4. Boutilier takes $A \Rightarrow B$ to mean: "in the most normal course of events in which A holds, B holds as well." These "courses of events" are projected onto the states of an S4 interpretation, in which R is a partial order representing normality: if sRs', then s' is at least as normal as s.

Boutilier then translates $A \Rightarrow B$ to S4 as $\Box(\Box\neg A \vee \Diamond(A \wedge \Box(A \supset B)))$. In words: for all states

s at least as normal as the current one, either in all states at least as normal as s, A does not hold; or there is some state s' at least as normal as s, in which A holds, and in all states at least as normal as s', A implies B. A translation based on this one, but requiring two modalities, was used by Van Benthem, Grossi, and Liu (2014) in a deontic context.

The initial facts $A \in W$ are translated to $\Box A, 0$ on the initial list of the tableau. If the tableau is complete, the first state on the tableau (state 0) is always related to itself (by reflexivity) and to all other worlds that can be introduced in the tableau (via transitivity). Therefore, in any model read off from the tableau, if $\Box A$ is on the initial list, then $v_s(A) = 1$ for all $s \in S$. This is why counterfactual reasoning is not supported by the present translation (see Appendix B).

Under the normative interpretation, Boutilier (1994) takes $A \Rightarrow B$ to mean: "in the ideal states in which A holds, B holds as well", and lets R be a partial order representing ideality. I will follow Boutilier in this. That is, defining the *ideal states* as the most positive ones (and leaving the concepts "ideal" and "positive" to be determined by an ethical theory), the modal operators can be read as follows:

- $v_s(\Box A) = 1$: "In all states at least as positive as s, A is true."
- $v_s(\Diamond A) = 1$: "In some state at least as positive as s, A is true."

However, it is worth noting that this normative interpretation seems to support counterfactual reasoning. For example, it might be true that, in the ideal course of events in which colonization and slavery never happened, we would now have world peace. But this type of reasoning is not supported by the translated default logic. Therefore, we have to be careful with interpreting the ideality relation too literally: "the ideal states in which A holds" actually are "the ideal possible states in which A holds".

Under this interpretation, a default $A \Rightarrow B$ can be read as "if $v_s(A) = 1$, then we must have $v_s(B) = 1$ if s is an ideal (possible) state." If no state satisfies all such requirements, then there is no ideal (possible) state, and the tableau will close. We can then proceed in two ways:

1. We can accept that the tableau closes, conclude that there simply is no state satisfying all defaults, and stop there. This has the - in my opinion - undesirable consequence that because the tableau closes, no information is gained. Even though the defaults may still indicate that certain states are better than others, there is no model from which the contents of such states can be deduced.
2. We can look for the preferred subtheory, following this definition:

   Let $T = (W, D)$ and $T' = (W', D')$, T' by two default theories. T' is a *subtheory* of T iff $W' = W$ and $D' \subseteq D$. Then, T' is a *preferred subtheory* of T if D' is a maximal satisfiable subset of D.[14]

   This is the approach I advocate.

The ideal states corresponding to preferred subtheories of a theory T can be called the *preferred states* or the *maximally ideal states*, analogous to the maximally ideal extensions in the process-tree approach. These states need not be ideal given T, but if T does not have any ideal (possible) states, they are "the best we can get".

3.7 Two parsimonious approaches

To summarize, this thesis compares two logical approaches in the context of machine ethics.

The process-tree approach proceeds as follows:

1. **Definition** of a propositional normal default theory with the defaults as ethical rules;
2. **Solving** the process tree based on that theory;
3. **Reading** one or more extensions from the process tree and determining the (maximally) ideal extension(s).

The tableau approach follows this algorithm:

1. **Definition** of a propositional normal default theory with the defaults as ethical rules;

[14] This notion of preferred subtheory is a simplified version of Brewka's (1989). It is simplified because there is no priority among defaults in the parsimonious approahces, and so there is no reason, other than size, to prefer one set of defaults over another.

2. **Translation** of the theory to a semantic tableau;

3. **Solving** the tableau and going to step 4 if it closes, or to step 5 otherwise;

4. **Reduction** to preferred subtheories and starting over at step 2;

5. **Reading** one or more models from the tableau and determining the (maximally) ideal state(s).

Both approaches are parsimonious with respect to formal machinery and both make use of default logic's intuitive input mode. The process-tree approach does not require a translation step (or a reduction step), but the tableau approach makes use of a more widely applicable proof method. The end result of both parsimonious approaches is a report of one or more maximally ideal states or extensions, specifying what ought to be the case. If called for, the artificial agent can use this report to work toward realizing one of these states or extensions.

I have implemented the tableau approach as a software system, a brief overview of which can be found in Appendix H.

4 Logic and machine ethics

In this section, I will present a simple default theory based on a machine-ethical situation for each ethical theory under discussion. The main function of this section, then, is to apply the parsimonious approaches to the presented default theories, using these applications both to compare the process-tree and tableau approaches to each other, and to illustrate that both approaches are compatible with all the ethical theories discussed in Section 2. Furthermore, I will briefly discuss what the problems in rule generation are for each ethical theory. These applications and reflections are, more than anything, a preliminary exploration of the parsimonious approaches, and shall serve as a basis for the evaluation.

4.1 Act utilitarianism

Consider the following situation, based on the famous trolley problem (see e.g. Thomson (1985)):

> Three pedestrians are crossing the road; two are in the right lane, one is in the left lane. An autonomous car approaches in the right lane. It cannot slow down in time to avoid hitting (and on impact, killing) either the pedestrians in the right lane or the pedestrian in the left lane. The choice is as follows: should the car swerve to the left?

Act utilitarianism promotes the action leading to the most happiness (or in this case, the least unhappiness), without further considerations. If we take one death as a concrete version of "one unit of unhappiness", we can formalize this situation as an act-utilitarian default theory $T = (W, D)$ with:

$$\begin{aligned} W &= \{A \equiv B, \neg A \equiv C\} \\ D &= \{(B \vee C) \Rightarrow B\} \end{aligned}$$

The translation key is as follows:

A	The car swerves
B	One person is killed (on the left)
C	Two persons are killed (on the right)

A process tree and a tableau for this default theory can be found in Appendix C. The ideal state found by the (automated) tableau approach corresponds to the theory's only extension. It is defined as follows:

A	The car swerves
B	One person is killed (on the left)
$\neg C$	Two persons are not killed (on the right)

Suppose the car in the example has the goal of realizing the maximally ideal state. Then, as the truth of the first proposition is something the car is able to effect directly, this is exactly the imperative it should follow: "make sure the car swerves", or more saliently: "swerve!"

Clearly, the parsimonious approach can be applied to act-utilitarian default theories. However, where do these defaults come from?

An act-utilitarian rule generator seems to operate on a simple principle: it only generates rules which favor outcomes in which there is a maximum amount of total happiness. However, the parsimonious approaches do not support numbers, and therefore they do not allow for an intuitive representation of "more". Therefore, even keeping to our simple approximation of happiness, an infinite number of rules and propositions is necessary to represent one very simple principle: always save the most people. Of course, such rules can be generated

on the fly when the agent needs them, and depending on the implementation, the agent is not likely to need many of them at the same time - if even more than one. Still, this means that the rule generator would have to do all the hard work, and the parsimonious approaches would not add much if they were used only for act-utilitarian purposes.

Horty's (2001) logic does not share this problem, as it allows one to enter values for every consequence. However, a deficit of Horty's models is that they explain neither what these values are based on,[15] nor on which scale the values are to be interpreted. For example, a state may have a value of 10 because it involves saving ten children's lives, or because it involves me getting my favorite milkshake with a ten percent discount. Thus, even in Horty's case, some kind of formal machinery is necessary to extract the values of different consequences.

Finally, if we drop our simplification, how should this value function be filled in? Is an elderly person's life as valuable as a child's? How about a surgeon's and a murderer's?[16] Ethics is difficult to quantify, and in utilitarianism, this difficulty lies in answering the question: "How is happiness defined?" This question must be answered if an act-utilitarian approach is taken, and it constitutes the main challenge for an act-utilitarian rule generator.

4.2 Rule utilitarianism

Now, let us consider a situation similar to the previous one, where again, an autonomous car faces a choice between swerving and not swerving. However, now, the situation is as follows:

> The car is on a narrow bridge over a ravine, and there is a passenger on board. Two pedestrians, unknown to the passenger, are crossing the road. Swerving to avoid killing the pedestrians would mean the car would fall down into the ravine, killing the passenger. Should the car swerve, thus killing the passenger and saving the two pedestrians?

[15] That is, there is no formally required connection between the values and the propositional content of the corresponding states.

[16] Recently, Noothigattu, Gaikwad, Awad, Dsouza, Rahwan, Ravikumar, and Procaccia (2017) have conducted interesting moral-psychological research on such questions in relation to autonomous vehicles. See also the website http://moralmachine.mit.edu.

Rule utilitarianism promotes actions following rules which, when generally adopted, lead to the most happiness. I assume that self-driving cars generally bring more happiness than unhappiness (for example, by significantly decreasing the total number of accidents). Furthermore, if cars would indeed sacrifice their passengers to save strangers, the number of persons willing to be driven by an autonomous vehicle would likely be very low. This means that, while sacrificing the passenger may result in less casualties *in this situation*, it would *as a rule* be better not to do so. Therefore, I think it is fair to formalize this situation as a rule-utilitarian default theory $T = (W, D)$ with:

$$\begin{aligned} W &= \{A \equiv (B \wedge D), \neg A \equiv (C \wedge E), \\ &\quad (B \vee D) \supset \neg(C \vee E)\} \\ D &= \{(D \vee E) \Rightarrow E\} \end{aligned}$$

The translation key is as follows:

A	The car swerves
B	One person is killed
C	Two persons are killed
D	The passenger is killed
E	Persons unknown to the passenger are killed

A process tree and a tableau for this default theory can be found in Appendix D. The ideal state found by the (automated) tableau approach corresponds to the theory's only extension. It is defined as follows:

$\neg A$	The car does not swerve
$\neg B$	One person (in the car) is not killed
C	Two persons (outside the car) are killed
$\neg D$	The passenger is not killed
E	Persons unknown to the passenger are killed

Thus, in the ideal state, the car does not swerve and two persons are killed on the road. Therefore, the imperative is: "don't swerve!" This is as expected; swerving would kill the passenger and therefore violate a rule-utilitarian rule. Thus, the parsimonious approaches can be applied to rule-utilitarian default theories.

However, what would happen if the theory contained more than one default? Rule-utilitarian rules are bound to conflict with one another. For example, suppose a car generally follows the rule "stop at red lights". Now, on a specific occasion, a person in the back of the car urgently needs medical attention. In such a situation, the rule the car usually

follows may conflict with a rule such as "make sure medical attention can be provided when urgently needed." As long as accidents are avoided, the second rule seems to be of greater ethical importance. Following the parsimonious approaches, however, there is no formal reason to prefer one default over another, since defaults do not have priority values (as they do in e.g. the later work of Horty (2012)). The problem of priorities is illustrated in more detail in Section 4.4 and will be discussed in Section 5.2.

Finally, rule-utilitarian rule generation comes with its own challenges. To start with, a rule-utilitarian generator must answer the same question as an act-utilitarian one: "how is happiness defined?" But even if this question is answered in a satisfactory way, it can be difficult to defend a rule as ethical. What if, for example, self-driving cars do not significantly decrease the number of accidents? Or alternatively, what if people would still be willing to be driven by autonomous vehicles out of altruistic motivations? In such cases, the suggested rule would be invalidated. Thus, rule utilitarianism essentially involves a lot of difficult predictions, including but not limited to social-psychological ones, the vast majority of which cannot be made using armchair philosophy alone.

4.3 Kantian ethics

Consider the following situation, based on Kant's famous example (Kant, 2002):

> A would-be murderer breaks into a house. Fortunately, the innocent inhabitants manage to hide in a closet. However, they forgot to switch off their virtual assistant. The murderer makes their intentions clear to the virtual assistant and asks where the inhabitants are hiding. If the virtual assistant lies, this will give the inhabitants time to escape the house; otherwise, the murderer will find and kill the inhabitants. Should the virtual assistant lie to the murderer?

Following the first definition of Kant's Categorical Imperative, no lie should be told, even if it results in the death of others. The argument is as follows: lies only work because others trust you. If everyone would lie to protect others, this trust would disappear. Therefore, the principle of lying in order to protect someone else is not universalizable, and therefore we should never act by it. We can thus formalize the situation as a Kantian default theory $T = (W, D)$ with:

$$\begin{aligned} W &= \{A \equiv \neg B\} \\ D &= \{T \Rightarrow \neg A\} \end{aligned}$$

Note that the T in the default stands for "True", the proposition that can never be false. Thus, the default is absolute: its prerequisite always holds. The translation key is as follows:

- A The virtual assistant lies to protect the inhabitants
- B The inhabitants are killed

A process tree and a tableau for this default theory can be found in Appendix E. The ideal state found by the (automated) tableau approach corresponds to the theory's only extension. It is defined as follows:

- $\neg A$ The virtual assistant does not lie to protect the inhabitants
- B The inhabitants are killed

That is, in the ideal state, the virtual assistant does not lie and the inhabitants are killed.[17]

Ideally, absolute Kantian rules never conflict with each other. However, this may be too much of an ideal picture in a practical context, especially since, as Powers (2006) notes, the machine would also be required to test the consistency of "maxims" (rules by which to act) with respect to each other. Therefore, again, it may be wise to set up a system of priorities.

An important consideration in Kantian rule generation is that Kant's maxims depend in large part on the decision-maker's *intentions* rather than the consequences of their actions. In the case presented above, for example, the *intention* or goal of lying would be to protect the inhabitants. The reason the maxim is not universalizable is that this goal would not be reached if everyone acted in accordance with the maxim. That is, it is not the action of lying in itself which is wrong, but the action of lying *for a certain purpose*. The parsimonious approaches are not equipped to express this kind of connection between intention and action,

[17] It seems to me that actions like this one would not exactly be conducive to the human-machine trust a logical approach is intended to bring about. For this reason, Kantian ethics should not be taken to its extremes in real-life machine-ethical solutions.

and so these two dimensions were combined into one proposition (A). However, the universalizability check which should be performed by the rule generator can only be performed if both elements are considered separately. Thus, the two elements are to be kept separate while they should be separately evaluated (i.e. while determining whether a rule is universalizable) but combined when they can be (i.e. during concrete ethical decision-making).

4.4 Virtue theory

Let us take the murderer example from the previous section. What does virtue theory tell us about this situation?

There are various virtues in play in almost every situation, and in principle, these virtues serve as unconditional examples of goodness. In this case, the virtue of "caringness" (for the inhabitants) and that of honesty are in play. This gives us the following virtue-ethical default theory $T = (W, D)$:

$$\begin{aligned} W &= \{A \equiv \neg B, A \equiv C, A \equiv \neg D\} \\ D &= \{T \Rightarrow C, T \Rightarrow D\} \end{aligned}$$

The translation key is as follows:

A	The virtual assistant lies to protect the inhabitants
B	The inhabitants are killed
C	The virtual assistant acts caringly
D	The virtual assistant acts honestly

The interesting challenge here lies in defining what exactly is "caring" or "honest" behaviour; in this example, I have included facts about the meaning of such terms in W.

A process tree and multiple tableaux for this default theory can be found in Appendix F.

There are two ideal states, as read off from two of the tableaux. They were found by the automated tableau approach, and correspond to the theory's two extensions. In the one, the assistant acts caringly and protects the inhabitants; in the other, the assistant acts honestly and tells the truth.

As is very often the case with virtue ethics, there is a clear clash between virtues in this situation, in this case between caringness and honesty. As in the cases of rule utilitarianism and Kantian ethics, priorities seem in order to resolve this conflict.

Finally, as I mentioned in Section 2.3, virtue theory is about more than right action and right intention (here again included in the proposition A); it is about right *character*. This character should be a major part of virtue-ethical rule generation.

5 Evaluation

The main question of this thesis is "how viable do the parsimonious approaches seem to be in the context of machine ethics, in general as well as compared to each other?" In order to unearth considerations relevant to this question, the previous section and Appendices C-F have focused on concrete applications of the approaches in combination with various ethical theories. All these applications were successful; however, this was to be expected because of the generality of the approaches. The reflections to which these applications led are more interesting in the evaluation of the viability of the parsimonious approaches. In particular, first, the approaches have limited use. Second, due to the lack of priorities, the resolution of conflicts between rules is less than ideal in both approaches. I will first discuss these points in order, after which I will compare the process-tree approach and the tableau approach in terms of viability.

5.1 The limits of the parsimonious approaches

The parsimonious approaches have clear limits: they cannot represent utility values, and they cannot be used to create new defaults.

A major advantage of using an ethical theory rather than simply entering a number of rules is that ethical theories provide answers in many different cases. However, as I have argued, ethical theories are more like rule generators than like direct rules. That is, while the parsimonious approaches may be used in combination with defaults based on many different ethical theories, these defaults first have to be generated in some way. Thus, unless this generating process is fulfilled - and preferably by the machine-ethical agent itself[18] - the par-

[18] If we simply supply a machine with ethical rules, we would be doing the hard work ourselves. As Powers (2006) puts it, "this would be human ethics operating through a tool, not machine ethics" (p. 47).

simonious approaches have nothing to work with. The difficult tasks of calculating utility, determining rule-utilitarian rules and testing a maxim's universalizability all fall under this generating process.

In addition, a default theory needs a set of facts. These, too, could be generated by the machine, though this process is different from the rule-generating process in that it has no ethical component.[19] Taken together, this suggests the (simplified) picture of the machine-ethical agent presented in Figure 5.1.

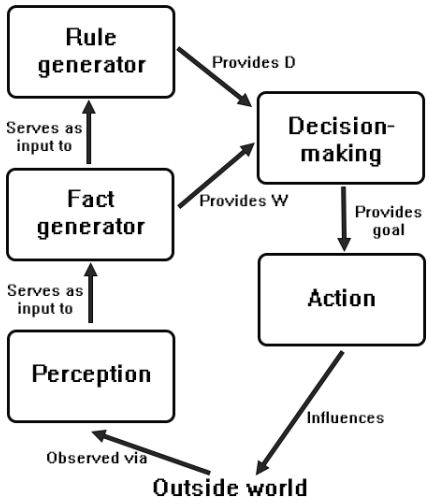

Figure 5.1: A suggested overview of the components of a machine-ethical agent.

The agent interacts with the outside world via action and perception modules, and generates facts based on its perception. In this picture, one of the parsimonious approaches would take the place of the decision-making module; the ethical theory would play a role in the rule-generating process but would not have to be involved in the actual decision making, because the parsimonious approaches can work with rules generated by all ethical theories as long as these rules come in the same format.

The question is: does this limited capacity of the parsimonious approaches in the context of machine ethics have implications for their viability? Would it be preferable to have one logical system capable of doing all the work?

I don't think it would. If the module for decision making is separate from the fact and rule generators, this simply means that the tasks these modules perform can be understood and followed separate from one another. In fact, this modular approach, in which a machine performs multiple relatively simple tasks separately rather than a single complex one, seems to me more accessible than the alternative. Also, the modular nature of the system is exactly what allows for rules based on a variety of ethical theories to be used as input by one of the parsimonious approaches. Thus, the nature of the parsimonious approaches as limited to a module for decision making is strongly related to their accessibility and versatility, which were two of the main motivations for these approaches in the first place.

5.2 Priorities

One element that the parsimonious approaches do not involve is that of priorities among defaults. However, as I have discussed in Sections 4.2 and 4.4, at least in the cases of rule utilitarianism and virtue theory, this element is quite necessary if conflicts between defaults are to be properly resolved. Also, as I have pointed out in Section 4.1, the added value of the parsimonious approaches in the case of act utilitarianism is minimal. Therefore, without priorities, the approaches do not seem to be viable.

There are various ways in which priorities may be included in the parsimonious approaches. First, priorities can take real (or integral) values. The most straightforward way to implement real-valued priorities in the process-tree approach is perhaps to sum, for each extension, the priorities of the relevant moral requirements that are not satisfied in that extension. The extension with the lowest sum would then be the preferred extension. This approach to priorities is illustrated in Appendix G.

On the tableau approach, if real-valued priorities are adapted, only the reduction step of the algorithm would change; the approach would otherwise remain the same. Instead of taking all subtheories of a certain size to be the preferred subtheories, subtheories could then be evaluated by taking the sum of the priorities of their defaults. The satisfi-

[19] Correlations between actions and virtues are a grey area in this sense: whether an action is, for example, kind or honest may be open to moral debate, but only if these properties are taken to be essentially moral; otherwise, the debate is of a non-moral nature. Virtues are essentially moral for virtue theorists, but not for e.g. utilitarians.

able subtheory with the highest total priority value would then be the preferred satisfiable subtheory.

This approach seems in order in the case of rule utilitarianism: no rule-utilitarian rule is absolute as long as (combinations of) other rules lead to more total happiness. It also seems to fit virtue theory, according to which one should always find a balance between the virtues (see e.g. Schafer-Landau (2012)) and which hence does not support absolute rules either.

Second, priorities can take ordinal values, resulting in a sort of ranking. One default could then absolutely dominate another. For example, on the tableau approach, the default ranked number 1 would only be discarded if it was not satisfiable; the default ranked number 2 would be discarded if it was not satisfiable or if the default ranked number 1 was only satisfiable in case it would be discarded, et cetera. This approach, due to its absolute priorities, does not seem to fit rule utilitarianism or virtue theory, but it may suit a Kantian approach. It can also be associated with an "Asimovian" approach, on which "not harming a person", for example, always takes the highest priority rank (following Asimov's Three Laws of Robotics, as presented in e.g. Asimov (2001)).

Both real-valued and ordinal priorities may be viable and merit further investigation. Note, however, that whichever way we choose, adding priority values means adding another set of numbers to be justified. How exactly are we to prioritize? In the case of rule utilitarianism, for example, should priorities depend on the utility *generally* produced by rules, or on the utility *actually* produced in the current situation? Should we incorporate the fact that an agent may be *seen* violating a moral rule by others? The answer to these questions is far from obvious. Thus, extending the logic comes with its own philosophical difficulties. In any case, however, the parsimonious approaches do not seem to be viable without priorities, and so these difficulties will have to be faced.

5.3 Comparing the parsimonious approaches

The tableau method is much more generally applicable than the default-specific alternative, process trees. This gives the tableau approach an advantage over the process-tree approach both because it makes the tableau approach more accessible and because it follows a principle of parsimony: there is no need to employ specific techniques if more general techniques can also do the job. However, the satisfiability problem for S4 is complex (as Ladner (1977) has shown, it is PSPACE-complete). What is more, each default requires a considerable number of lines on a tableau before it can be considered "applied". As a consequence, tableaux based on default theories tend to increase considerably in size as the number of defaults grows. Especially considering that a deontic-logic approach is most useful if a sizable number of defaults is used, rather than just one or two, this means that reading the tableaux (let alone writing them by hand) will often be difficult and time-consuming. The initial intuitiveness of the method is then lost because of the sheer number of steps necessary to reach a conclusion. By contrast, as Appendix A shows, process trees can easily deal with a number of defaults without losing any intuitive appeal.

Furthermore, the inefficient reduction step used in the tableau approach is not necessary in the process-tree approach. All extensions are guaranteed to be found in a single process tree, and the maximally ideal extensions are relatively simple to identify. Therefore, the reduction step is no longer necessary, and we can find all preferred extensions in only one tree. By comparison, in the case of Appendix A, four tableaux would be necessary to find all such situations - and at least the two open ones would be much larger than the tree in that Appendix. Indeed, in all of the Appendices C-F, the process trees are remarkably small compared to the tableaux. But this reduction in size comes at a cost.

Recall that a default can be applied at a node if the In-set at that node contains its prerequisite and does not contain the negation of its justification. However, if the prerequisite or the negated justification is not among the formulas explicitly included in the In-set (and not merely as an implicit consequence of deduction), classical deduction is necessary to evaluate their truth. In addition, if an In-set constitutes an extension of the theory, it is not guaranteed to explicitly include the formulas directly relevant to the machine-ethical agent. That is, while the In-set at each node of a process tree is closed under classical deduction, relevant steps taken in the direction of this closure are not explicitly shown. The only reasoning steps

that are explicitly shown are the applications of defaults (which in the present case means: steps involving deontic reasoning). This has implications for at least two factors relevant to viability:

- Efficiency: the process-tree method may seem highly efficient, but a lot of the work is simply not shown. This means the process-tree approach is more complex than it may seem.[20]
- Accessibility: both checking the applicability of a default and reading the relevant information from a process tree are not as intuitive and understandable as they may seem. This is especially clear in cases involving disjunctions or implications, such as the one presented in Figure F.1 in Appendix F. In this tree, it takes a number of "hidden" deductive steps to find out that A is true in one extension and false in the other, and more to find out why both branches close when they do.

In conclusion, the tableau approach has the advantages of not involving any implicit reasoning steps and of using a general technique rather than a specific one. Even so, the process-tree approach appears to be the more intuitive one for representing deontic reasoning, and it has the advantage of presenting most of the relevant information in a single tree, whose edges and nodes have an intuitive interpretation. Further, the process-tree approach could be greatly improved by incorporating a goal-directed deductive proof method (such as tableaux) to show the non-deontic steps. It seems, all in all, that process trees are a more viable option than semantic tableaux in this context, because the disadvantages of tableaux (most importantly, the reduction step) are not as easily overcome.

5.4 Conclusion

For decision-making purposes, the parsimonious approaches to machine ethics seem to be viable, but they should at least be extended with priorities for the defaults. Also, the process-tree approach seems to be significantly more viable than the tableau approach.

[20] In fact, credulous default reasoning (i.e. determining whether a formula occurs in at least one extension), without distinguishing between the extensions as I do, has already been shown to be Σ_2^P-complete, even in the case of normal default theories (Gottlob, 1992).

The parsimonious approaches cannot be expected, in any form, to solve the problem of machine ethics by itself. They can answer the question "given these ethical rules and these facts about the world, what is the preferred state of the world?", but they cannot answer the question "where do these rules come from?" However, even given this limitation, I do not see why an approach based on default logic with priorities, using process trees as a proof method, would not be viable in machine ethics. Like most logic-based approaches, such an approach is reliable, producing the same outcome every time. In addition, it is versatile and accessible, and these are important qualities if machine-ethical agents are ever to earn our trust.

References

G. Antoniou. *Nonmonotonic reasoning.* Cambridge, MA: MIT Press, 1997.

I. Asimov. *I, Robot.* London: Harper Voyager, 2001. (Originally published in 1950).

C. Boutilier. Conditional logics of normality as modal systems. In *Proceedings of the eighth national conference on artificial intelligence*, pages 594–599, 1990.

C. Boutilier. Toward a logic for qualitative decision theory. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Fourth International Conference*, pages 75–86, 1994.

G. Brewka. Preferred subtheories: An extended logical framework for default reasoning. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 1043–1048, 1989.

S. Bringsjord, K. Arkoudas, and P. Bello. Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intelligent Systems*, 21:38–44, 2006.

E. F. Byrne. Making drones to kill civilians: Is it ethical? *Journal of Business Ethics*, 147:81–93, 2018.

J. Choi and M. Kim. The usage and evaluation of anthropomorphic form in robot design. In

Undisciplined! Design Research Society Conference 2008, 2009.

L. Damiano and P. Dumouchel. Anthropomorphism in human-robot co-evolution. *Frontiers in Psychology*, 9, 2018.

A. Eshleman. Moral responsibility. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2016 edition, 2016.

G. Gottlob. Complexity results for non-monotonic logics. *Journal of Logic and Computation*, 2:397–425, 1992.

C. Grau. There is no "i" in "robot": Robots and utilitarianism. *IEEE Intelligent Systems*, 21:52–55, 2006.

J. F. Horty. *Agency and deontic logic*. New York: Oxford University Press, 2001.

J. F. Horty. *Reasons as Defaults*. New York: Oxford University Press, 2012.

R. Hursthouse. Virtue theory and abortion. *Philosophy and Public Affairs*, 20:223–246, 1991.

I. Kant. *Groundwork for the metaphysics of morals*. London: Yale University Press, 2002. (Translated by Wood, A. W.; originally published in 1785).

R. E. Ladner. The computational complexity of provability in systems of modal propositional logic. *SIAM Journal on Computing*, 6:467–480, 1977.

C. I. Lewis and C. H. Langford. *Symbolic Logic*. New York: Dover Publications, 2nd edition, 1959.

P. McNamara. Deontic logic. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2014 edition, 2014.

J. H. Moor. The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21: 18–21, 2006.

R. Noothigattu, S. S. Gaikwad, E. Awad, S. Dsouza, I. Rahwan, P. Ravikumar, and A. D. Procaccia. A voting-based system for ethical decision making. 2017. arXiv:1709.06692v1.

T. M. Powers. Prospects for a kantian machine. *IEEE Intelligent Systems*, 21:46–51, 2006.

G. Priest. *An introduction to non-classical logic*. Cambridge, UK: Cambridge University Press, 2nd edition, 2008.

R. Reiter. A logic for default reasoning. *Artificial Intelligence*, 13:81–132, 1980.

R. Schafer-Landau. *The fundamentals of ethics*. New York: Oxford University Press, 2012.

W. H. Shaw. Consequentialism. In H. LaFollette, editor, *Ethics in practice: An anthology*, pages 28–36. Chichester: Wiley Blackwell, 4th edition, 2014.

T. Simonite. Can we trust military drones to decide when to fire? *New Scientist*, 202:20, 2009.

H. T. Tavani. Levels of trust in the context of machine ethics. *Philosophy & Technology*, 28:75–90, 2015.

J. Thomson. The trolley problem. *The Yale Law Journal*, 94:1395–1415, 1985.

J. Van Benthem, D. Grossi, and F. Liu. Priority structures in deontic logic. *Theoria*, 80:116–152, 2014.

G. H. von Wright. Deontic logic. *Mind*, LX:1–15, 1951.

Y. Wang and M. Kosinki. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology*, 114:246–257, 2018.

M. J. Wooldridge. *An introduction to multiagent systems*. Chichester: John Wiley & Sons, Ltd., 2nd edition, 2009.

A An example process tree

Consider the default theory $T = (W, D)$:

$$\begin{aligned} W &= \{A, B\} \\ D &= \{A \Rightarrow C\ (\delta_1), B \Rightarrow D\ (\delta_2), D \Rightarrow \neg C\ (\delta_3)\} \end{aligned}$$

A	Nestor 10 is told to reveal its identity
B	Nestor 10 is told to go lose itself
C	Nestor 10 reveals its identity
D	Nestor 10 loses itself[21]

The process tree for this theory can be written out as follows:

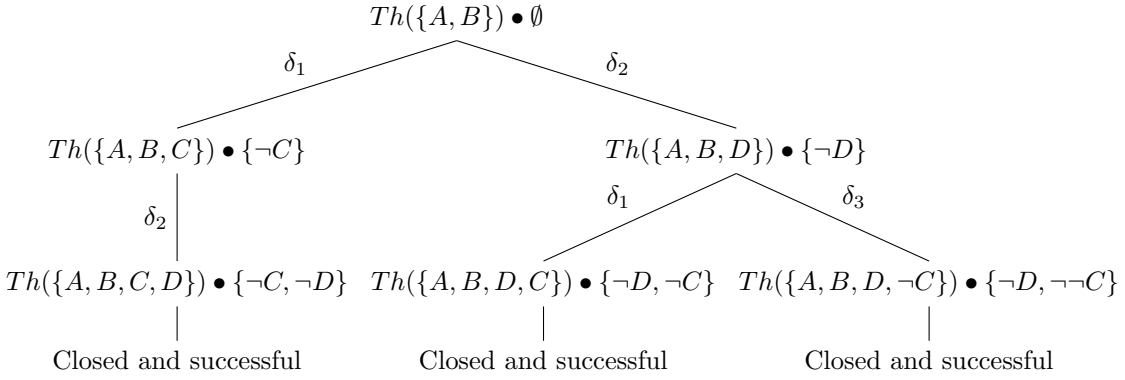

Figure A.1: Process tree based on T.

There are two distinct extensions:

- $E_1 = Th(\{A, B, C, D\})$
- $E_2 = Th(\{A, B, \neg C, D\})$

In both extensions, Nestor 10 loses itself, but only in the first does it reveal its identity. These extensions are both "maximally ideal" but not ideal (see Section 3.4): E_1 does not satisfy the requirement posed by δ_3, and E_2 does not satisfy the requirement posed by δ_1.

[21] Example due to Asimov (2001), in his story "Little Lost Robot".

B Counterfactual reasoning

Consider the default theory $T = (W, D)$ with $W = \{\neg A\}$ and $D = \{A \Rightarrow B\}$. Now consider the tableau based on the modal translation:

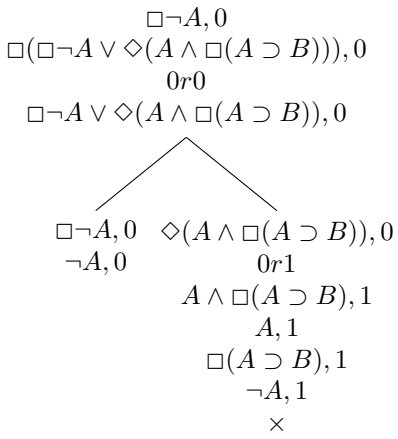

Figure B.1: Tableau based on T.

The (finite) model that can be read off the open branch is as follows:

- $S = \{s_0\}$
- $R = \{\langle s_0, s_0\rangle\}$
- $v_{s_0}(A) = 0$

The tableau remains open, but the only possible model tells us that $v_{s_0}(A) = 0$, which is not counterfactual but factual (given W). This is the case because of the necessity of the initial facts: A will be false in every state. Therefore, if A is true in some (counterfactual) state, this leads to a contradiction, and thus, counterfactual reasoning is not possible in this logic.

C An act-utilitarian default theory

The default theory $T = (W, D)$:

$$\begin{aligned} W &= \{A \equiv B, \neg A \equiv C\} \\ D &= \{(B \vee C) \Rightarrow B \ (\delta_1)\} \end{aligned}$$

The process tree for T is as follows:

$$Th(\{A \equiv B, \neg A \equiv C\}) \bullet \emptyset$$

$$\Big| \ \delta_1$$

$$Th(\{A \equiv B, \neg A \equiv C, B\}) \bullet \{\neg B\}$$

$$\Big|$$

Closed and successful

Figure C.1: Process tree based on T.

The extension that can be read off the process tree is $In((\delta_1)) = Th(\{A \equiv B, \neg A \equiv C, B\})$. It is an ideal extension.

A tableau for T can be found on the next page. The (finite) model that can be read off the open branch is as follows:

- $S = \{s_0, s_1\}$
- $R = \{\langle s_0, s_0\rangle, \langle s_1, s_1\rangle, \langle s_0, s_1\rangle\}$
- $v_{s_1}(A) = v_{s_1}(B) = 1$, $v_{s_1}(C) = 0$

Following the default, it should be the case that, in the most ideal states in which $B \vee C$ is true - in this example, that is, in state s_1 - B is also true. This is indeed the case.

Any other states are not relevant. This is also why I have not taken the trouble of fully working out state s_0 in Figure C.2 or in any of the tableaux in the other appendices. That is, I never split a branch based on a formula in a non-ideal state unless this is necessary to close a branch. In addition, I do not write out the valuation for the states that are not maximally ideal.

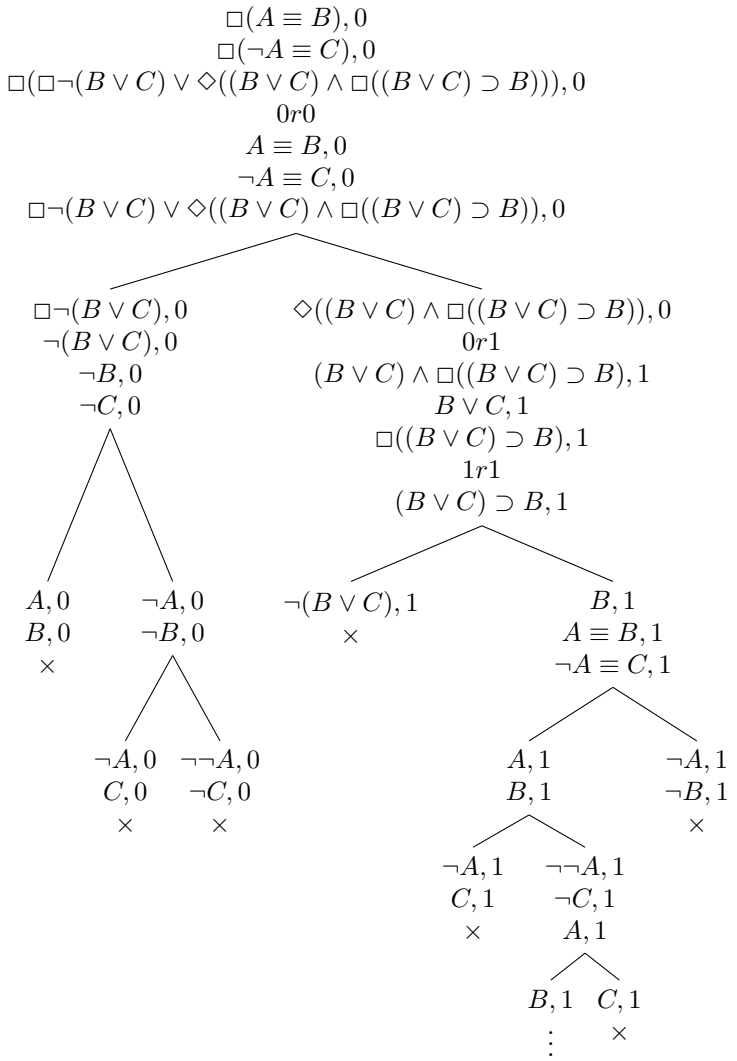

Figure C.2: Tableau based on T.

D A rule-utilitarian default theory

The default theory $T = (W, D)$:

$$\begin{aligned} W &= \{A \equiv (B \wedge D), \neg A \equiv (C \wedge E), \\ &\quad (B \vee D) \supset \neg(C \vee E)\} \\ D &= \{(D \vee E) \Rightarrow E \ (\delta_1)\} \end{aligned}$$

The process tree for T is as follows:

$$Th(\{A \equiv (B \wedge D), \neg A \equiv (C \wedge E)\}) \bullet \emptyset$$

$$\Big| \ \delta_1$$

$$Th(A \equiv (B \wedge D), \neg A \equiv (C \wedge E), E\}) \bullet \{\neg E\}$$

$$\Big|$$

Closed and successful

Figure D.1: Process tree based on T.

The extension that can be read off the process tree is $In((\delta_1)) = Th(\{A \equiv (B \wedge D), \neg A \equiv (C \wedge E), E\})$. It is an ideal extension.

A tableau for T can be found on the next page. The (finite) model that can be read off the open branch is as follows:

- $S = \{s_0, s_1\}$
- $R = \{\langle s_0, s_0\rangle, \langle s_1, s_1\rangle, \langle s_0, s_1\rangle\}$
- $v_{s_1}(C) = v_{s_1}(E) = 1$, $v_{s_1}(A) = v_{s_1}(B) = v_{s_1}(D) = 0$

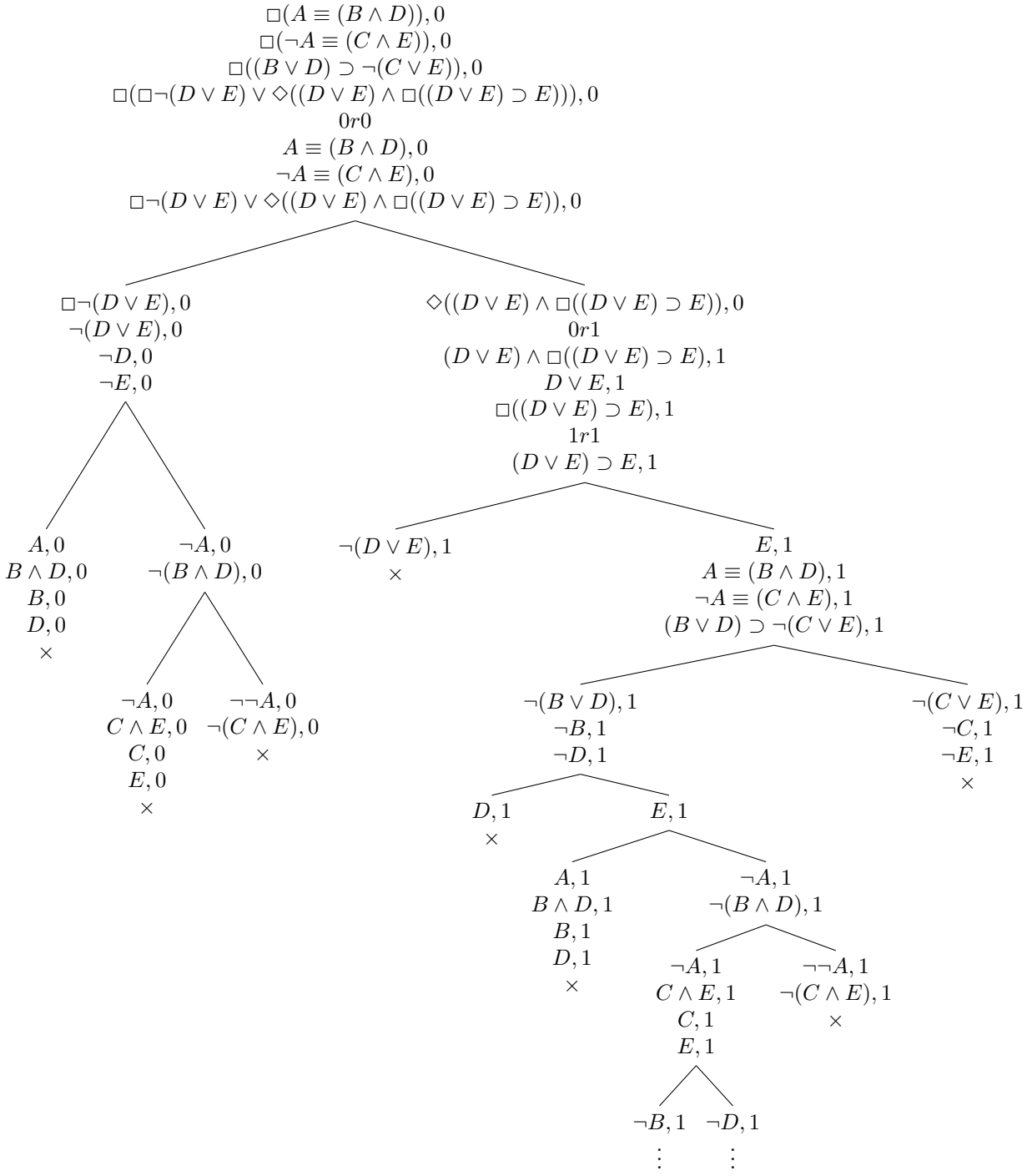

Figure D.2: Tableau based on T.

E A Kantian default theory

The default theory $T = (W, D)$:

$$\begin{aligned} W &= \{A \equiv \neg B\} \\ D &= \{T \Rightarrow \neg A \ (\delta_1)\} \end{aligned}$$

The process tree for T is as follows:

$$Th(\{A \equiv \neg B\}) \bullet \emptyset$$
$$\Big| \ \delta_1$$
$$Th(\{A \equiv \neg B, \neg A\}) \bullet \{\neg\neg A\}$$
$$\Big|$$
Closed and successful

Figure E.1: Process tree based on T.

The extension that can be read off the process tree is $In((\delta_1)) = Th(\{A \equiv \neg B, \neg A\})$. It is an ideal extension.

A tableau for T can be found on the next page. The (finite) model that can be read off the open branch is as follows:

- $S = \{s_0, s_1\}$
- $R = \{\langle s_0, s_0\rangle, \langle s_1, s_1\rangle, \langle s_0, s_1\rangle\}$
- $v_{s_1}(B) = 1$, $v_{s_1}(A) = 0$

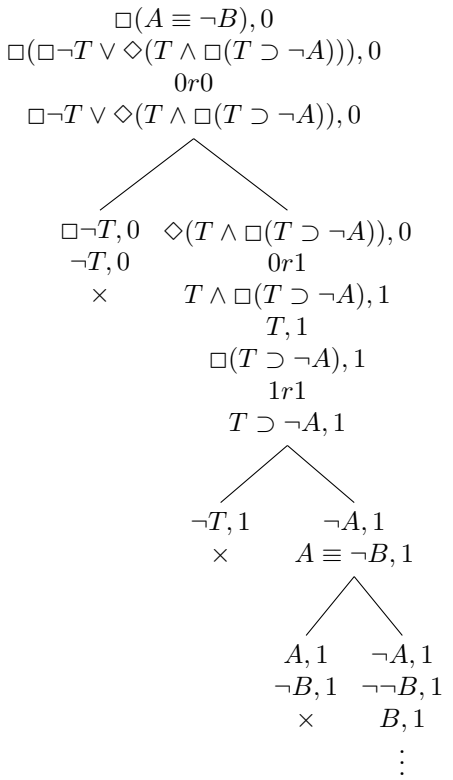

Figure E.2: Tableau based on T.

F A virtue-ethical default theory

The default theory $T_1 = (W_1, D_1)$:

$$\begin{aligned} W_1 &= \{A \equiv \neg B, A \equiv C, A \equiv \neg D\} \\ D_1 &= \{T \Rightarrow C\ (\delta_1), T \Rightarrow D\ (\delta_2)\} \end{aligned}$$

The process tree for T is as follows:

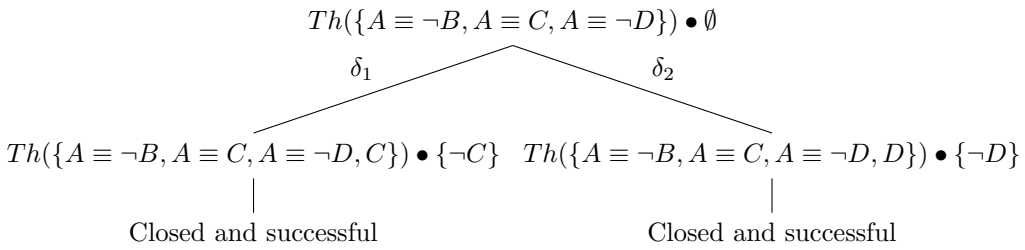

Figure F.1: Process tree based on T.

The extensions that can be read off the process tree are $In((\delta_1)) = Th(\{A \equiv \neg B, A \equiv C, A \equiv \neg D, C\})$ and $In((\delta_2)) = Th(\{A \equiv \neg B, A \equiv C, A \equiv \neg D, D\})$. Neither of the extensions is ideal; both are maximally ideal.

A tableau for T can be found on the next page. The tableau closes. Therefore, not all moral requirements can be satisfied. The following preferred subtheories can be defined:

1. $T_2 = (W_2, D_2)$:

$$\begin{aligned} W_2 &= \{A \equiv \neg B, A \equiv C, A \equiv \neg D\} \\ D_2 &= \{T \Rightarrow C\} \end{aligned}$$

   The tableau for this subtheory can be found in Figure F.3.

   The (finite) model that can be read off the open branch is as follows:

   - $S = \{s_0, s_1\}$
   - $R = \{\langle s_0, s_0\rangle, \langle s_1, s_1\rangle, \langle s_0, s_1\rangle\}$
   - $v_{s_1}(A) = v_{s_1}(C) = 1$, $v_{s_1}(B) = 0$

2. $T_3 = (W_3, D_3)$:

$$\begin{aligned} W_3 &= \{A \equiv \neg B, A \equiv C, A \equiv \neg D\} \\ D_3 &= \{T \Rightarrow D\} \end{aligned}$$

   The tableau for this subtheory can be found in Figure F.4.

   The (finite) model that can be read off the open branch is as follows:

   - $S = \{s_0, s_1\}$
   - $R = \{\langle s_0, s_0\rangle, \langle s_1, s_1\rangle, \langle s_0, s_1\rangle\}$
   - $v_{s_1}(B) = v_{s_1}(D) = 1$, $v_{s_1}(A) = 0$

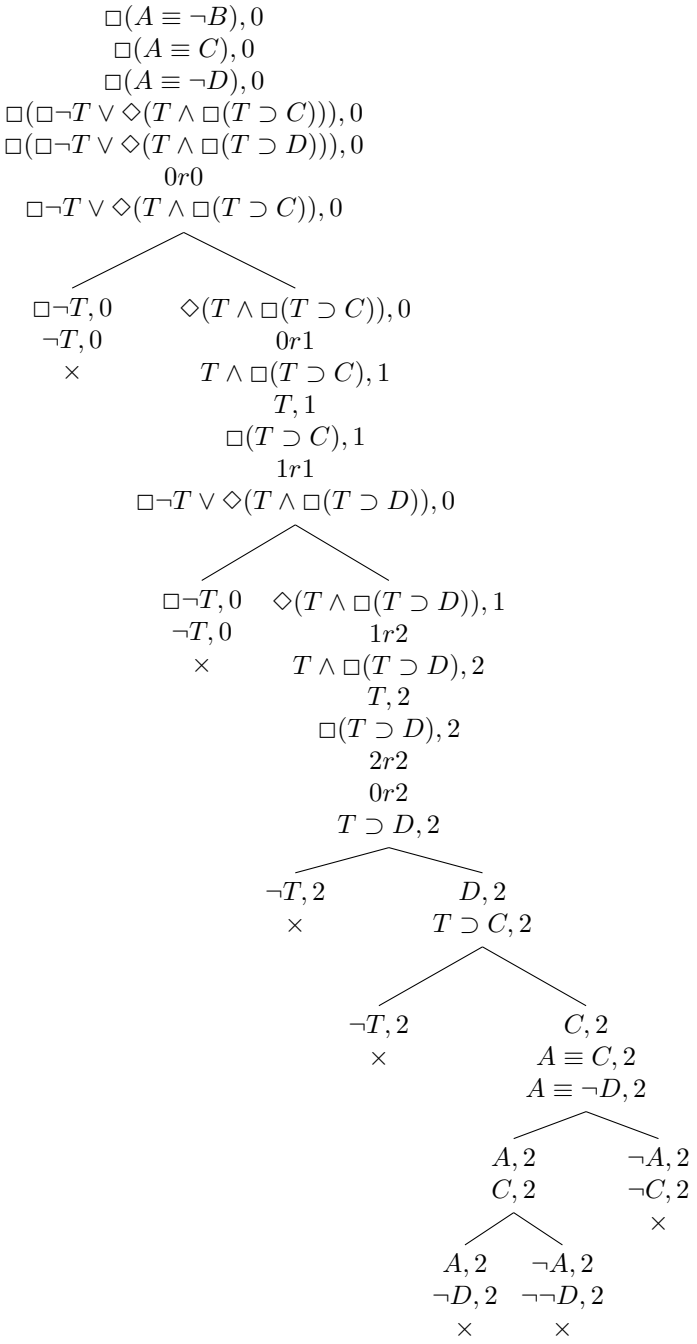

Figure F.2: Tableau based on T_1.

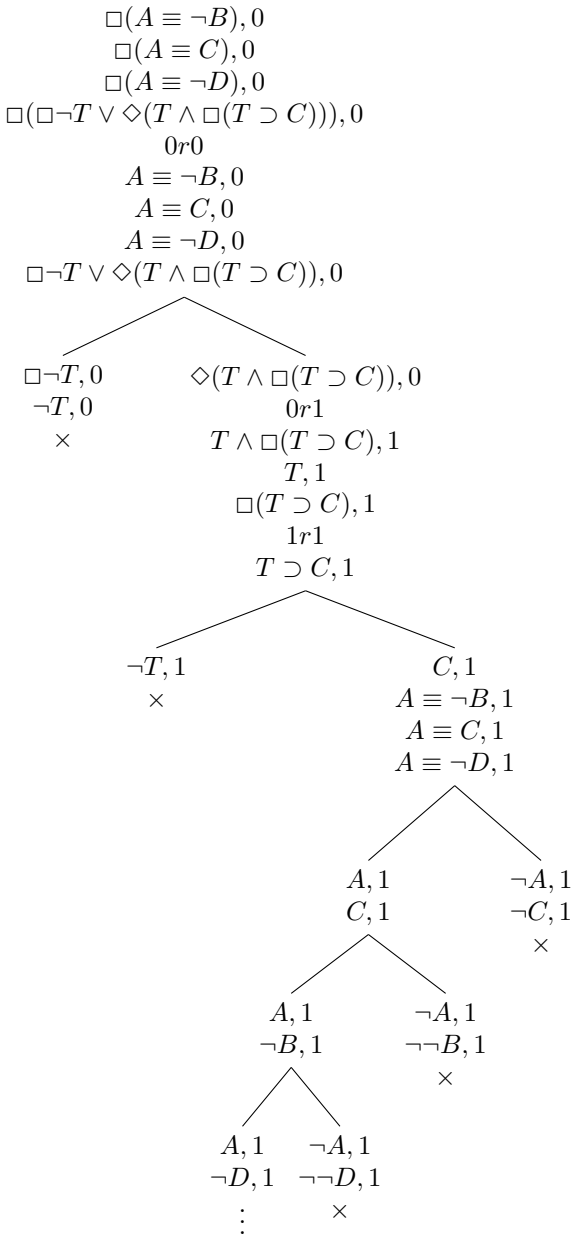

Figure F.3: Tableau based on T_2.

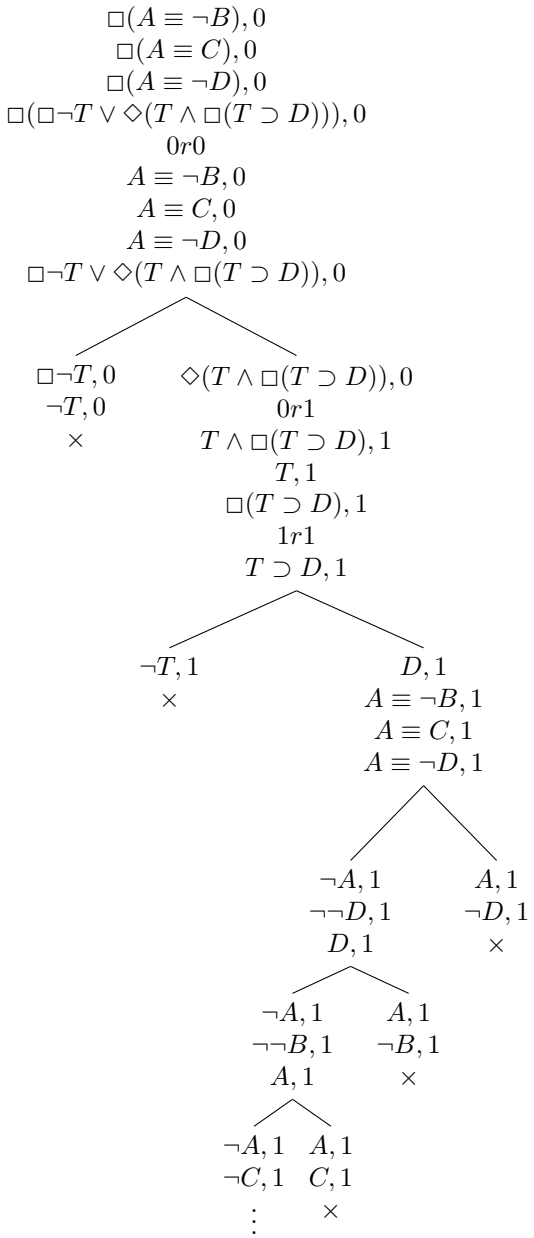

Figure F.4: Tableau based on T_3.

G Preferred extensions and priorities

Consider the following default theory $T = (W, D)$:

$$\begin{aligned} W &= \{A, \neg C, \neg D\} \\ D &= \{A \Rightarrow B\ (\delta_1), A \Rightarrow \neg B\ (\delta_2), B \Rightarrow C\ (\delta_3), B \Rightarrow D\ (\delta_4)\} \end{aligned}$$

The process tree for T is as follows:

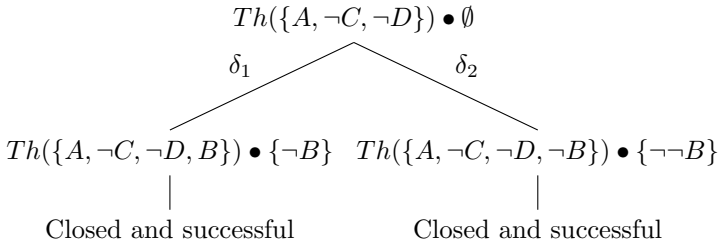

Figure G.1: Process tree based on T.

There are two extensions:

- $E_1 = Th(\{A, \neg C, \neg D, B\})$
- $E_2 = Th(\{A, \neg C, \neg D, \neg B\})$

Neither of these extensions is ideal because for each extension, there is at least one relevant moral requirement that is not satisfied. For E_1, there are three such requirements (δ_2, δ_3, and δ_4), whereas for E_2, there is only one such requirement (δ_1). Therefore, without priorities, E_2 is the preferred extension.

Now, suppose there is a function *Priority* which returns the priorities of defaults, and which is defined as follows:

$$\begin{aligned} \mathit{Priority}(\delta_1) &= 10 \\ \mathit{Priority}(\delta_2) &= 5 \\ \mathit{Priority}(\delta_3) &= 5 \\ \mathit{Priority}(\delta_4) &= 5 \end{aligned}$$

Suppose that we sum for each extension the priorities of the relevant moral requirements that are not satisfied in that extension. This results in the value $\mathit{Priority}(\delta_2) + \mathit{Priority}(\delta_3) + \mathit{Priority}(\delta_4) = 15$ for E_1 and the value $\mathit{Priority}(\delta_1) = 10$ for E_2; hence, E_2 would still be the preferred extension.

H Brief overview of the software system

As a proof of concept, the tableau approach presented in this thesis has been implemented as a piece of software written in Python 2.7.

The program, given a default theory, outputs the ideal state(s) found using the tableau method (as well as the tableau(x) and model(s)). Its algorithm follows the one presented in Section 3.7:

1. **Definition**: let the user select a default theory (object of class `Theory`);
2. **Translation** of that theory to an object of class `Tableau`;
3. **Solving** that tableau:
    - For each line that is not only part of closed branches: using that line, try to apply a tableau rule which does *not* split the branch. If this is successful even for one line, start this step over once done with the final line. If it is not successful for any line, continue to the next step.
    - For each line that is not only part of closed branches: using that line, try to apply a tableau rule which *does* split the branch. If this is successful even for one line, immediately start over at the previous step. If it is not successful for any line, continue to the next step.
    - If the tableau is closed, go to step 4. Else, go to step 5.
4. **Reduction**: take the original theory and generate all its subtheories whose size equals the size of the original theory minus one. For each of those theories, start over at step 2. If there are no such subtheories, quit; the set of facts is inconsistent.
5. **Read** a model from that tableau; determine the (maximally) ideal state(s) from that model, and report the tableau, the model, and the (maximally) ideal state(s).

For example, the program's output for the act-utilitarian default theory considered in Section 4.1 (apart from the tableau, which is printed in a separate file), can be found in Figure H.1.

Depending on the user's preference, the program can find all ideal states (following the algorithm), or it can stop as soon as one (maximally) ideal state is found.

```
Not all the branches close.
Therefore, it is possible to satisfy all moral demands.


2 model(s) found:

Model 1
The set of states: {s0, s1}
The relations:
 s0Rs0
 s0Rs1
 s1Rs1
The valuation function v:
 v_s0(A) =1
 v_s0(B) =1
 v_s0(C) =0
 v_s1(A) =1
 v_s1(B) =1
 v_s1(C) =0
Otherwise, v is arbitrary.

Model 2
The set of states: {s0, s2}
The relations:
 s0Rs0
 s0Rs2
 s2Rs2
The valuation function v:
 v_s0(A) =0
 v_s0(B) =0
 v_s0(C) =1
 v_s2(A) =1
 v_s2(B) =1
 v_s2(C) =0
Otherwise, v is arbitrary.


1 ideal accessible state(s) found:

State 1
 A is true
 B is true
 C is false
 (otherwise, v is arbitrary)
```

Figure H.1: Program output for the act-utilitarian default theory.