Extrapolating Modeling Techniques for Machine Ethics Reasoning from AI & Law

Jesse van der Ceelen *

2 september 2019

1 Introduction

Machine ethics is a field in AI that concerns itself with giving machines the capability to act ethically. As the broader field of AI continues to grow and machines become more and more capable of solving complex tasks without human intervention a certain risk also grows. As tasks become more complex it becomes increasingly difficult to consider them ethically neutral - an artificially intelligent agent capable of playing chess or Jeopardy does not raise many ethical concerns, but more practical, real-world examples like autonomously driven motorized vehicles certainly do. Since a machine autonomously acting in the real world to perform almost any sufficiently complex task will encounter situations that require some form of ethical judgement it is necessary to imbue them with the ability to act ethically.

However, machine ethics is a fairly new field, brought into greater prominence by the increasing capabilities of AI only relatively recently. This is due to autonomous agents starting to become capable of engaging with tasks complex enough to merit serious considerations on the ethics of their actions (though the field has been around for longer). A related but older field is that of AI & law, which is closely related to machine ethics in its purpose in broad terms; using machines to promote goodness in society. Even though they take different approaches in accomplishing this goal and can differ greatly in the way they operate, their shared goal makes comparing the two interesting. In particular, the seniority of AI & law compared to machine ethics can make it a valuable learning tool. Both looking at the lessons learned in AI & law and what advancements have been made in it can help advance machine ethics.

*student number 4061837, email j.vanderceelen@students.uu.nl

The purpose of this thesis is then to examine what, if anything, can be learned from the field of AI & law that could be used to improve machine ethics models. This is somewhat of a complex question however, in that it requires several other questions to be answered first. Most prominently is the question of whether or not the fields can really be compared at all. Intuitively they have common ground, but this may not turn out to be the case. Additionally, even if practice lines up with the intuition it is necessary to determine to what degree and in which areas the two fields overlap and differ.

Another factor to consider is that of the theory of ethics being used. There are different approaches to ethics that may differ in how they apply to law and ethics. Since the differences between these broader fields could be far more fundamental than those between the fields that approach these topics using AI they need to be examined as well.

Additionally, a closer examination of machine ethics models is also required in order to establish what purposes they can serve and what features they require, and which features are undesirable or even forbidden. This has some overlap with the question of how machine ethics relates to AI & law, but on the lower level of practical models rather than on the level of a theoretical overview of the fields themselves.

A similar examination of AI & law models would also be useful in order to draw some general comparisons between the models to serve as additional context along with the higher-level overview of the fields. This will be significantly more concise than the examination of the machine ethics models though, since the main focus of this thesis is on improving machine ethics models so they do warrant a more thorough analysis.

Given these questions, this thesis will be structured as follows: first a high-level overview of the fields of machine ethics and AI & law will be given to determine how they relate, what machine ethics lacks and what AI & law could offer to alleviate that. Secondly three major approaches to ethics will be examined on their relevance to the broader fields of law and ethics to determine the differences between the two fields. Next will be an examination of various machine ethics models to refine the earlier question of what the field as a whole lacks, as well as to determine the relevancy and desirability of their various features. After this a similar examination of AI & law models will be given, though more concisely.

These sections will lay the groundwork for the second part of the thesis, an in-depth comparison of the performance of three models - one from machine ethics, one from AI & law and one that is domain-agnostic, serving as a middle ground between the two other models. This will take the form of applying these three models to two different ethical dilemmas. First the

dilemmas themselves will be examined, followed by the models themselves. After this all three models will by applied to each dilemma, followed by a comparison on their process and performance. Finally the results obtained in the aforementioned sections will be combined in the conclusions.

2 Machine Ethics and AI & law: An overview

2.1 The fields compared

The fields of machine ethics and AI & law share a lot of the same challenges, born from their similar goals: interpreting ambiguous generalizations, solving rule conflicts and dealing with unexpected complications, to name some prominent ones. Another trait they share is that systems in neither field are generally expected to be fully autonomous, as that property is still too complex to achieve with sufficient guarantees.

One of the main differences is that of ad-hoc versus post-hoc reasoning, with machine ethics mainly concerning itself with the former and AI & law with the latter. Machine ethics is about reasoning over the current situation, the action that should be taken right now or in the near future and its direct consequences. On the other hand, AI & law is usually about judging a situation after it has already occured, possibly quite some time in the past if the court proceedings take a long time. This raises several practical issues that are present in one field but not in the other. For instance, ad-hoc reasoning requires heavy use of hypotheticals in order to capture possible actions as well as possible outcomes of those actions, as failure to sufficiently consider alternatives will result in a suboptimal or even bad choices being made. While post-hoc reasoning can and in fact often does make use of hypotheticals - for example, to aid in classification of past cases with respect to a case currently being examined - the role they play in both and the importance they carry are different.

There are also challenges that concern post-hoc reasoning that are mostly absent in ad-hoc reasoning. Evidence, for example, is something that ad-hoc reasoning is usually not concerned with [Prakken, 2017]. At worst an agent may find that it lacks the knowledge to come to a decision, in which case it may refer to an 'oracle', which in this context is a catch-all term for external information sources. An oracle can take the form of other agents that can be questioned, sensors used to probe the environment, an experiment (be it conducted by the agent or otherwise) or any other procedure that acquires information from external sources. [Pereira and Saptawijaya, 2011].

Related to these differences in the point in time being reasoned about

is the difference in time scale. While in machine ethics judgements usually involve themselves most closely with the immediate future [Prakken, 2017], AI & law often needs to take a much more far-reaching look at both the consequences of the action being judged and the consequences of potential judgements as well as looking back far before the action being judged to establish things like motive and intent. Moreover, even in those cases where AI & law systems are future-oriented they tend to look beyond the immediate future, such as when aiding in the creation of contracts or legislation. Another consequence of this is that the time available for reasoning is different in both fields: In machine ethics time to make a decision is often very limited, sometimes down to fractions of a second, while court cases can take years.

Another difference is in the nature of arguments in both fields. In AI & law arguments are highly adversarial [Prakken, 2017] since both sides are represented by different agents with different interests. This is contrasted by arguments in machine ethics, where they serve more as a reasoning tool than as a decision making process, usually even lacking a real opponent as the reasoning is all done by a single agent.

Finally, at first sight it may appear as a difference that the law is strict and concerns itself with its own strict body of rules while in ethics many other vague factors may come into play. However, this misses the point of general rules and exceptions in law, many of which exist precisely to capture these vague factors. The law surrounding violence in self-defense comes to mind, where the boundary between justified and excessive violence is vague and highly context-sensitive. Mind, there is still a difference here, but it is a much subtler one than may appear at first glance: The difference lies in the degree to which rules preside over the decision making process, which is greater in AI & law than it is in machine ethics.

2.2 Machine Ethics: What it lacks

However, these are just the general differences between the fields as a result of their different approaches. While they certainly need to be kept in mind throughout, more than them is required in order to meaningfully bridge the gap between the two fields. Specifically, the practical accomplishments of machine ethics models need to be examined to expose weak points - undesirable assumptions, lack of scope or depth and lack of features like soundness, completeness or verifiability. Problems regarding the selection of ethical standards and rules or the collection of facts are ignored in this examination, as they are outside of the scope of this thesis - only the reasoning process itself is considered.

Mackworth [2011] points out that a common problem surrounding machine ethics is the general naivete found in suggestions for ethical AI systems with respect to current technical capabilities. Oftentimes the capability to perform some complex task that is currently beyond the state of the art is assumped to be possible with good guarantees. For example, discussion on ethical robots frequently considers which requirements should be imposed on the actions of an agent in order for it to act ethically. This ignores the problem of how those requirements would be imposed and whether or not the agent would be able to satisfy them, thereby essentially presupposing that these problems can be solved with currently extant methods, which is not generally the case.

A risk assessment of potential superintelligent AI by Stuart Armstrong et al. [2012] is an example of this. While discussing rule-based methods to keep the behavior of such a hypothetical AI within ethical boundaries they mention that "...all the human-understandable terms [...] need to be made rigorous for the OAI for this approach to work.". While this does acknowledge the problem it still ignores it in the remainder of the assessment, presuming it solved for the purposes of the discussion. This showcases the point Mackworth was making, namely that machine ethics discussions tend to overfocus on the 'what' while ignoring the 'how'.

McLaren names another overarching issue for the field, namely that of assuming one form of ethical reasoning to be the correct one [McLaren, 2011]. Different formal models for ethical reasoning may in the same scenario come up with different solutions. In this case, both solutions are ethically sound (presuming that both models are) and one may even be considered better depending on the context, but neither model can be generally preferred over the other. This is because of the different philosophical approaches to ethics that exist which can fundamentally oppose each other in certain areas while each approach is still in and of itself valid. In essence, no model can be selected as the end-all be-all for ethical reasoning no matter how advanced it is since the huge variety of ethical problems and contexts in which they can occur means that there are always scenarios for which different approaches may come up with different solutions. While some steps have been taken to tackle this problem - for example by using category theory to allow multiple reasoning systems to be used - it remains a big blind spot overall.

Related to the issue Mackworth raises is the oversimplification of problem spaces. As opposed to the often ignored 'outside' issues like guarantees and verification, the issue of computational complexity is often tackled since most systems with capabilities beyond the very basic will run into it. The issue lies in the simplification that often takes place in an effort to reduce the complexity to workable levels. The W.D. model by Anderson et al. [2005] is

an example that falls prey to this. W.D. makes ethical decisions based on the prima facie duties of W.D. Ross, but only scores violation or satisfaction of each duty on a five-point scale, which is very un-nuanced. Given the complex nature of the problem of ethical reasoning it is unavoidable that some simplification takes place, but it is key to address the specific weaknesses each instance brings to the model.

A more meta-level problem is the lack of decision systems in machine ethics. This is to be expected and even sensible, as any system intended for practical use cannot yet have the guarantees necessary to make those decisions autonomously. This has led to a state of affairs where most machine ethics systems take on an assisting or advisory role instead. Since there are so few systems that try to tackle possibly the hardest part of the process - confidently making a decision - there is at least a relative lack in academic writing on this part of the process.

Additionally, wherever these do exist the models often suffer heavily from the oversimplification mentioned earlier. An example of this are two systems developed by Anderson et al. [2005] called Jeremy and W.D., the former of which reduces probabilities used in computing 'total net pleasure' to small sets of possible values while the latter reduces the violation or satisfaction of duties to a range of five integers.

Building hypotheticals is a more concrete problem that machine ethics faces. While it is simple and often even intuitive for a human to build an imaginary situation for aid in reasoning or decision making it is hard for an algorithm to build meaningful alternatives to situations. This is because both a fairly large knowledge base (to have enough alternative factors for a variety of different situations) and context-sensitive reasoning to compare cases are required. This is also a problem that is often either ignored or simplified, be it implicitly or explicitly. A common way to explicitly simplify it is through the closed-world assumption, which presumes that all true statements are in the knowledge base, thereby making any statements not in or derivable from the knowledge base false [Cadoli and Lenzerini, 1994].

2.3 AI & Law: How it can help

With these problems in mind it is possible for the differences between AI & law to be turned into an asset rather than a weakness. The problems in fitting the two fields together of course remain, but when they do fit the advantages that AI & law has can be used to fill some of the gaps that machine ethics leaves.

An interesting case where AI & law may help, albeit indirectly, is with the problem of the insufficiency of just one ethical theory, such as act util-

itarianism or the prima facie duties theory of W.D. Ross. While AI & law models do not directly alleviate this issue they are by design more specific about where this problem applies. The issue of naming one model of ethics the 'correct' one arises when considering ethical dilemmas that are not unambiguously resolvable. But for every ethical problem for which this is the case there are also plenty that can, in the absence of abnormal factors, be generally solved without much ambiguity. In AI & law these cases are separated naturally as 'cases to which mostly specific rules apply' and 'cases in which more general rules apply'. This means that for the 'simple' cases one mode of reasoning does often suffice, reducing the complexity of adding more modes by only applying it to the truly difficult problems.

Building hypothetical cases, while a difficult problem in both fields does have more work done on it in AI & law, for example in the HYPO system [Bench Capon, 2017]. The greater importance that cases carry in AI & law compared to machine ethics naturally lead to a state of affairs where case hypotheticals are less easily ignored, and thus have more work done on their generation.

Moreover, since the legal system exists and is used in practice it stands to reason that it can still deal with complex cases with some modicum of efficiency in spite of it not allowing them to be egregiously simplified. Examining how the legal system maintains this balance, especially through the lens of AI & law systems and models could be helpful for remedying this weakness in machine ethics.

3 Orientation on ethics

Theories on ethics can be generally divided into three approaches, which are not necessarily mutually exclusive [Burton et al., 2017]:

Deontology: Deontology views ethics in terms of moral codes or rules rather than consequences. An action or decision is ethical under this approach if it is correctly derived from an ethical rule. In essence this means that the moral code that prescribed an action is what is judged to be ethical or not, rather than the consequences of said action. The source of these rules may differ wildly depending on the theory, from divine providence to personal values, but it remains based on rules in all cases [Burton et al., 2017].

Consequentialism: Consequentialism contrasts deontology by placing greater importance on the consequences rather than the actions that cause them. A consequentialist views an action or decision as ethical if it pro-

duces a good outcome. A prominent example of a consequentialist theory is utilitarianism, which judges actions based on a maximization of 'utility', a numeric value that expresses the desirability of outcomes compared to each other [Burton et al., 2017]. The great question of utilitarianism is on the definition and arithmetic of utility. For example, a naive definition that allows utility of repeated events to be added together without mitigation can lead to ridiculous results such as the utility of torturing a person for a hundred years to prevent specks of dust in the eyes of a sufficiently large number of people coming out positive [Yudkowsky, 2007].

Virtue ethics: Contrasting both of the former two approaches, virtue ethics centralizes virtues, traits that a deontologist would ascribe to those who correctly follow the right rules and a consequentialist would describe as those that lead its possessor to make choices that lead to good consequences [Hursthouse and Pettigrove, 2016]. However, rather than formulating virtues as a derivative of more fundamental concepts it is taken as a fundamental concept in and of itself. In other words, it focuses on how an agent should be rather than its actions [Burton et al., 2017].

Since virtues are complex both in their interactions with each other and the context in which they are applied, they cannot be naively followed to their logical conclusion. This can, for example, lead to courage being taken to the extreme of excessively ignoring danger or honesty being taken to the extreme of making excessively harmful (albeit true) statements. Virtue ethics therefore also requires a practical wisdom or 'phronesis', the ability to judge each individual case in its context to make a correct decision. Note that this concept is not unique to virtue ethics - deontology also requires this type of reasoning to correctly apply general rules to particular cases and to solve conflicts between rules that are equally applicable [Hursthouse and Pettigrove, 2016].

Notably, virtue ethics focuses on being rather than acting - a consequence of this is that it is impossible to determine whether someone holds virtues or acts ethically purely based on one or even multiple of their actions [Hursthouse and Pettigrove, 2016]. Rather, the agent itself needs to be examined - do they truly hold the relevant virtues or do they happen to comply with them for exterior reasons, such as avoiding lying for fear of getting caught? What was the decision process like - did it sufficiently consider the particulars of the case and was the relevant virtue one of the deciding factors? This complex analysis means that virtue ethics can present no rigorous method for discerning ethical actions from unethical ones, at least not in the way deontology and consequentialism can.

Virtue ethics instead focuses on an abstract and high-level view on ethics

on an agent-by-agent basis, judging each agent on its virtues and ability to reason with them in a practical setting. It is also focused on personal improvement, both by cultivating good virtues and eliminating vices and by practicing practical reasoning [Burton et al., 2017].

These three approaches to ethics are, as will be described in the following subsections, each relevant to law and ethics in their own ways. Examining in which ways and to what extent these parallels are present will aid in bridging the gap between the two.

3.1 Deontology

Since deontology is based on rules that prescribe actions - what one ought to and ought not to do - it most clearly parallels legal reasoning, which is similarly focused on establishing and maintaining guidelines for allowed and disallowed actions.

However, the role rules play in law is not the same across the world. In civil law, the system most commonly used in Europe, abstracted and generalized rules are the most important when deciding on a case, with case law - rulings based on precedents - being of secondary importance. This contrasts with common law, which takes the opposite approach of deriving law primarily from precedents, with equal power to written statutes [Dainow, 1966].

The importance of deontology is clearly seen when looking at the modes of reasoning Prakken and Sartor [2015] establish: deriving facts, classifying facts and deriving legal consequences. The latter two involve themselves heavily with the rules - classifying facts is essential for rule application, attempting to determine which rules are applicable given the facts of the case. In deriving legal consequences this is even more clear, since it involves application of the rules and, depending on the case, determining hierarchies between equally applicable rules and finding possible exceptions. Deriving the facts of the case does not clearly require rules, but since this project is only interested in normative reasoning - the final two steps - this is of no further consequence for the purposes of this thesis.

Argumentation: Prominent examples of deontology-based systems are argumentation-based systems. These are a natural choice for legal cases, as arguments between two parties is mainly how they play out in the courtroom. As such these systems help with classification, assisting the prosecution and defense in constructing arguments. HYPO is an example of such a system, working in the common law domain [Bench Capon, 2017].

An important aspect of argumentation is that its results are almost always defeasible - they can be defeated later if more information about the case becomes known. This lines up well with how ethical arguments work, where generalized rules such as 'do not lie' may end up being overridden or defeated by other rules depending on the particulars of a case.

However, there are also some important differences between legal and ethical argumentation. In legal arguments there is an established body of rules that, while open to interpretation, are in principle set in stone, while the validity of any ethical rule is open to debate. Furthermore, in legal arguments in common law (and also to a lesser degree in civil law in the form of jurisprudence) precedents are very important, while they are less so in ethical arguments. Whereas precedent takes on a binding role in common law or a strong advisory role in civil law, it cannot take on more than such an advisory role in ethical arguments, though it is rarely - if ever - completely irrelevant.

This is closely related to the greater malleability that ethical rules have compared to legal rules, where only their applicability or interpretation can be argued and, in the case of civil law, where the complete body of rules is clear to all parties. All applicable ethical rules, meanwhile, are never completely clear and the option for adding new ones during the argument is open while it is not in most legal systems.

3.2 Consequentialism

It would be nice if all legal cases could be solved purely through application and ordering of clear rules but this is, regrettably, impossible. In the case of civil law it is not possible to foresee every possible situation and exception beforehand, so the body of law is always incomplete and in need of evaluation and interpretation. Common law suffers from the same problem, as even though it does not attempt to codify all rules its precedents are similarly insufficient. Since society is constantly changing even the most unlikely body of precedents, where all possible variations of all crimes have been committed and brought to court, falls short since novel laws and situations may arise in the future - the introduction of autonomous vehicles is a good example of this.

To fill these gaps it is necessary that it is examined on a case-by-case basis which rules ought to be applied and how they should be applied. Referring back to the modes of reasoning, this means that both classification of facts and deriving legal consequences require additional insight aside from the rules themselves.

Argumentation-based systems are also common places to find consequen-

tialist notions, arguing for or against a certain fact or rule interpretation on the basis of the effects that decision would have on the legal precedent (in the case of common law). The outcome of the case and the consequences for the involved parties that follow from that outcome can also be argued.

This is also where the lines between the three approaches become blurred, since a system that allows both consequentialist and deontological arguments on classification problems is certainly possible. This is to be expected though, since they are approaches, not mutually exclusive theorems [Burton et al., 2017]. In fact, given the complexity of legal problems it may even be desirable to incorporate various approaches so as to avoid missing possibly crucial aspects of a case, which could in turn lead to a wrong decision being made.

3.3 Virtue ethics

At first virtue ethics might seem of little concern to this research, as its nature of examining ethics in terms of agents rather than actions lends itself best to the first mode of reasoning, that of deriving facts. However, this is only partially true. This is because a core concept in virtue ethics, that of practical wisdom, is very relevant to legal reasoning. The whole point of having a trial is based on this idea, that mechanical rule application, be they defeasible or otherwise, is insufficient to provide a fair judgement. To do this properly the particulars of the case need to be discussed to solve potential conflicts on their interpretation, decide on often incompletely defined hierarchies and account for situations that the lawmakers never considered.

While virtue ethics offers little in the way of formal models, meaning that it does not help directly tackle these difficult steps, it does make for a particularly suitable lens for evaluating systems. The legal system as a whole, the methods used by lawyers to build arguments and the methods used by judges to make decisions can all be examined this way to gain new insights. Unfortunately this is only tangentially useful for the purposes of this thesis.

4 Machine Ethics models

In this section existing machine ethics models and implemented systems will be examined for their properties to cross-reference those with their weak and strong points. From this various approaches and properties that can aid machine ethics models can be distilled.

Each model will be specifically examined on its choice of ethical theory (deontology, utilitarianism, etc.), its field and scope of application (Is it

made for a specific domain or is it more general? What role does it fulfill in the reasoning process - decision making, rule retrieval, resolving preferences, etc.?), which knowledge representation is used (How expressive is this representation? What are its limitations?), which reasoning model is used (What properties does this model have or not? Is it complete, sound, etc.?) and how well it performs in the case of an implemented model.

4.1 Jeremy and W.D.

Two systems developed by Anderson et al. [2005] are called Jeremy and W.D. Both of these are able to autonomously make decisions as opposed to taking on an advisory role. Both are also specified in greater detail relative to their complexity, with W.D. actually being implemented by the authors.

Jeremy is the simplest of the two models, an implementation of a version of Act Utilitarianism called Hedonostic Act Utilitarianism. This ethical theory judges actions - specifically, all actions that the agent can currently perform - by the net (dis)pleasure inflicted upon all those affected by the action, with everybody counting equally towards this total. It represents this pleasure by assigning numbers, which are selected from a limited set on a scale, to the intensity, duration and probability of occurrence of the pleasure, for each person affected. The total pleasure for each individual is then simply the product of these factors, from which the net pleasure overall is calculated as the sum of these products for every person. The action that results in the highest net pleasure value is then selected.

While Jeremy is certainly robust in that it is both consistent and complete with respect to its own knowledge model, it is difficult to consider it very practically applicable. For one, the end user needs to estimate all of these values - which are difficult to quantify - while the model relies absolutely on these numbers. Even a small estimation error may completely alter the result. Anderson et al. solve this by only allowing the user to choose from a very small number of estimated values (just three for intensity and probability) which, as discussed earlier, greatly oversimplifies the problem. Moreover, act utilitarianism, especially when implemented in the straightforward 'sum of products' manner like this is prone to criticism in and of itself, such as the 'sand in eyes vs torture' argument presented by Yudkowsky [2007].

In this argument Yudkowsky constructs a dilemma between two choices: either one person gets tortured continuously for fifty years, or a very large number of people - Yudkowsky names a specific number, but any sufficiently large finite number will do - get specks of dust in their eyes for a short moment of mild irritation. By allowing the straight summation of utility values, it should be possible to make the number of people who get specks

of dust in their eyes large enough that the sum of the tiny negative utility values from each individual person dips below the negative utility of a person getting tortured for fifty years. However, intuitively this would clearly be impermissible.

W.D. contrasts Jeremy by being based on the theory of Prima Facie duties by Ross [1930]. The crux of this theory is that just one duty (e.g. maximizing pleasure like act utilitarianism has) is insufficient for a complete ethical theory, regardless of what the duty is. As such, Ross proposes seven prima facie duties that are all equally strong. However, this theory is not as straightforward to implement since Ross provides no method of deciding which duty should be preferred in the case of a conflict. For example, the duties of beneficence (maximize goodness) and non-maleficence (minimize badness) come into conflict in situations where all choices have undesirable effects.

To solve this Anderson et al. used the approach of reflective equilibrium, which tests its output, if it can reach any conclusion, against the intuitive solution provided by the user, which is then added to the knowledge base to update the hypothesis of the preference relationships between the seven prima facie duties. This knowledge base is represented using inductive logic programming, which uses Horn clauses to learn some relation, $supersedes(A1, A2)$ in the case of W.D., which expresses that action A1 is preferred over A2. Horn clauses are implications with conjunctions of positive literals as the antecedent (i.e. $A \wedge B \wedge C$) and single positive literals as the consequent.

The way it works is by asking the user to list all actions that can resolve the dilemma at hand and to rate each on its satisfaction or violation of each of the prima facie duties. Using this information W.D. checks if any of the actions given supersede all the others, which is returned as the result if it exists. If no action is preferred over the others it asks the user to provide a resolution to the dilemma in the form of a selected action, which W.D. then uses to update its hypothesis of what the *supersedes* relation of the user is, adding the intuitive answer plus the input case to its knowledge base as a new training example. This training session can also be initiated by the user if the system returns a result that does not align with the intuition of the user, which works the same way.

A useful feature of this model is that it guarantees that the hypothesis is both complete and consistent with respect to the input cases it has seen - it covers all positive cases (where the first action supersedes the second one) and none of the negative cases (where this is not the case). In essence this means that all input cases it has seen will be solved correctly, which is particularly good because it can guarantee this after every training session.

By looking at these features it is clear that W.D. is much stronger than Jeremy, since it still has strong guarantees on its output relative to its case base (completeness and consistency) while it follows a more robust ethical theory. However, it is certainly not without flaws. For one, it too suffers from a small allowed input range for duty violation/satisfaction, namely only five integer numbers. Secondly, because it bases its learning on the intuitions of its user it is both prone to incorporating general human biases and the biases specific to its user. Think for example of which values and ethical theories the user holds. But in spite of these flaws Anderson et al. claim that it is "...successful at determining the correct action in many [cases] that it has not previously seen.", though they make no mention of the metrics or training data used.

4.2 Truth-Teller

Truth-Teller is a model that contrasts the previous two, developed by Ashley and McLaren [1995] in that it is intended to serve a supportive role. More specifically, it compares dilemmas about whether or not a character in the scenario should tell the truth or not. This comparison takes the form of an ethical analysis of two cases provided to it, comparing and differentiating them on the reasons the characters in both cases have for telling or not telling the truth.

In essence Truth-Teller subscribes to the idea of casuistry or case-based reasoning. It is somewhat similar to W.D. in that it uses cases, but rather than attempting to derive general rules from them it focuses on the cases themselves, trying to get more information from singular comparisons than W.D. tries to do.

In its representation it points out two characters as the 'truth-teller' and the 'truth-receiver'. The former has several choices to resolve the situation, at least one of which is to simply tell the truth, though multiple alternative actions may exist. Each action has one or more reasons to support it, which are modeled as a specific duty, the specific example of said duty for this action and reason and a list of beneficiaries. For example, telling the truth may produce benefit (duty) by strengthening the relationship between the truth-teller and the truth-receiver (example) which benefits the truth-teller (beneficiary). Other characters and specific relationships between them may also be added to the model.

Truth-Teller analyzes the pair of cases in four phases. First it attempts to map the reasons for both cases onto each other, noting where they overlap and where they differ. The second step is to look at preference relationships between the reasons, actions and agents of each case that can weaken or

strengthen a reason (i.e. not causing harm may be preferred over promoting goodness). In the third step it selects some reasons that are particularly similar or different to focus on to make one of three arguments: that one case is as strong as or stronger than the other, that the cases are only slightly comparable or that they are not comparable at all. Finally it coalesces all of this information and converts it into a human readable format to output.

Truth-teller is shown in tests to be decently capable, having been evaluated on twenty cases by five professional ethicists on their reasonableness, completeness and context sensitivity on a scale of 1 to 10 with 1 being low and 10 being high. In these gradings they were evaluated with moderate scores on average (6.3 on reasonableness, 6.2 on completeness and 6.1 on context sensitivity) while two comparisons written by graduate students were graded better (8.2 on reasonableness, 7.7 on completeness and 7.8 on context sensitivity). The main criticisms noted by the experts were its failure to consider hypotheticals in its analysis (e.g. would the situation change if it was certain truth-receiver would not get angry at truth-teller for telling the truth?) as well as its somewhat naive way of listing the reasons one by one instead of relating them to one another.

Overall, the fact that it performed as decently as it did under the scrutiny of experts who held it to a fairly high standard (the same as graduate students would be held to) is impressive, but it also has to be considered how limited in both scope and application the program is, being both limited to giving an analysis that does not even directly constitute advice for the user as well as only being able to handle truth-telling dilemmas. Still, its power in comparing cases could make it useful when integrated with another system that can use the result of its analyses to do reasoning.

4.3 SIROCCO

Enter SIROCCO, another system developed by McLaren [2011]. It continues McLarens exploration of casuistry by tackling the field of ethics in engineering, specifically by attempting to combine case-based reasoning with general principles described in the ethical codes and guidelines established by the engineering board. This is somewhat similar to W.D. using cases to decide between conflicting principles, but in this context the principles are much less general and greater in quantity. Other than that the problem is the same - the principles often conflict in application and they are not generally in a preference relationship with each other, so this needs to be established to reach a decision.

The main difference between SIROCCO and W.D. is that SIROCCO does not make its own decisions. Rather, it uses the knowledge from its cases to,

when given a case, provide the user both with relevant cases it already knows and relevant codes. Additionally, it provides the user with a list of possible preference relationships between codes.

SIROCCO represents its cases in the Engineering Transcription Language (ETL), which describes each case as a timed sequence of facts, which are in turn constructed using several predefined types of words. It then selects relevant cases by first retrieving any case it already knows that share facts with the current case which are scored based on the degree of their match, followed by which it attempts to map the top-ranked cases to the current one using a search algorithm (A*).

SIROCCO has been tested for its performance along with several other text retrieval techniques, scored on the overlap of the selection with that of human experts (the engineering board). This was done both on exact matches (exactly the same texts were selected) and inexact matches (similar texts were selected). In this experiment SIROCCO significantly outperformed all but one of its rival methods, which it still beat but not by enough to be considered significant (p=0.057). However, it also performed worse than the ethics review board itself, though this comparison is somewhat skewed in favor of the review board since it was used as the comparing standard - i.e. if SIROCCO selected a case that the board did not select it was still counted against it even though it may have been a good pick.

SIROCCO also has an interesting synergy with truth-teller - it can be used to retrieve cases that are already similar for truth-teller to compare in greater detail than SIROCCO can do on its own. The drawback of this is of course that the limitations of the fields in which both are applicable overlap, and eliminating this gap will increase the complexity of the problem by a comparative amount. They also both share and perhaps even exacerbate their weakness of relying on complex representation of the cases - whereas W.D. can represent anything that first-order Horn clauses can the requirements for the representational languages in both truth-teller and SIROCCO are much more stringent, being limited both by the greater complexity of their representations and the limited application field they work within. On the other hand, it is also clear that this added representation adds richness, so it remains a trade-off, and considering how well both systems perform this tradeoff is fairly reasonable.

4.4 ACORDA

ACORDA is a system based on the work on prospective logic programming by Pereira and Saptawijaya [2011]. The idea of prospective logic programming is to make use of abductive reasoning to create hypothetical scenarios that

an agent can use to make useful yet defeasible predictions about its future. Abductive reasoning refers to the process of finding likely causes given an effect, as opposed to deduction which uses known facts (causes) to derive effects. It notably contrasts deduction in that its results are defeasible - an abducted fact is only true as long as no contradictory evidence is known.

Agents using prospective logic programming start with an initial theory that contains a knowledge base in the form of a first order logic program with facts, abducible facts and preferences among them, if any. The initial theory also contains a moral theory, which is any formal model of ethics that can produce preferences among situations. The prospective logic program then attempts to find abducible extensions to its current theory and select the most preferred one. Essentially, it reasons backwards from its goals to find the best ways to accomplish them.

ACORDA is very flexible in terms of which kinds of functions it can fulfil since it does not care how exactly its moral theory works. Whether ACORDA is deontological or utilitarian or something else is dependent on the choice of moral theory. On top of this its field of application is also flexible in the degree of detail in its representation of the situation, since that depends on the size and choice of facts and preferences in the knowledge base. However, it is also limited in that aspect because it requires the knowledge base to be represented in a first order language, which has its own restrictions, as well as being further limited in scope due to the number of possible futures (which can each have multiple abducible causes with possibly multiple side effects each) being able to quickly grow too big for practical use.

ACORDA also stands out because it aims to tackle an often oversimplified or ignored problem: how to generate the hypothetical scenarios a lot of ethical models are intended to reason over in the first place. This is really the intended use of ACORDA, as even though a fully implemented instance would see use as a decision maker that part of the process would be largely delegated to the moral theory, which is not specified by ACORDA itself.

As an illustration of how ACORDA is useful for ethical reasoning in particular, Pereira and Saptawijaya use ACORDA to model the principle of double effect. In effect this principle states that it is impermissible to harm someone if it is the means to achieve greater overall goodness, while it is permissible to do so if the harm is a foreseen but unintended side effect of the action that improves the overall goodness [Hauser, 2006].

This principle features prominently in the trolley problem, which goes as follows: A man finds himself by a railway passing through a tunnel. A trolley is traveling along the track at great speed, with five people in the tunnel it is approaching. Since they cannot get off the tracks they will surely all die in the collision. However, a man stands next to a lever that will switch the

trolley to a different tunnel, saving the five people. Unfortunately, in this alternative tunnel is also one person who will die when the lever is flipped. Thus the dilemma is whether the man should allow five people to die from inaction or whether he should allow one person to die by flipping the lever to save five others [Foot, 1967].

The principle of double effect explains why, in the trolley problem, people usually prefer flipping the switch but generally do not want to push a fat person in the way of the trolley to stop it in a situation where no second track exists, even though the outcomes (five lives saved through sacrifice of one other) appear to be the same when examined at a surface level [Hauser et al., 2007]. While modeling the principle of double effect is certainly possible in models that use forward reasoning it is much more natural to do so using abductive logic.

4.5 Kantian machines

Powers takes an interesting approach to machine ethics by suggesting three different ways to construct a computable formalization of practical reasoning on ethics from the categorical imperative by Kant [Powers, 2006]. Note that these are not fully specified formalizations but merely explorations of what the requirements for such a program would be and what prominent challenges are left to tackle.

The categorical imperative is the deontological guide for creating moral rules, stating that one should only act according to those maxims that one can will to be a universal law. In other words, it says that actions are permissible if and only if one would want every agent to follow a rule from which the maxim (plan) leading to the action directly follows. Additionally, the rule generated like this should be consistent with all other rules generated in the same way to prevent contradictions.

Powers suggests three possible formalizations of this imperative, all bottom-up, meaning that they do not start with a human-prescribed set of universal laws, instead having to derive it all on their own. These three will be briefly discussed here.

The first suggests a simple method of universalizing each maxim and label it as either forbidden, permissible or obligatory, doing a consistency check after each universalization. While simple, this suffers from the problem that overly specific actions, objects or agents are impossible to universalize, making the model incomplete (or in the case where the universalization is carried through anyway, unsound). For example, the maxim "I will award this prize" cannot be universalized because 'this prize' refers to a specific object - clearly one would not want everybody to give away that specific

prize. Powers offers to solve the first of these by forcing quantification over at least circumstances, purposes and agents, thus eliminating the specificities in the places where they matter.

The second method proposed improves on the naive consistency check, which should do much more than just check internal inconsistencies - it needs to be compared to other rules as well. Powers suggests using nonmonotonic logic to accomplish this, Reiter's default logic to be precise. The suggestion is to add commonsense rules as defeasible rules, which contrast monotonic rules by being able to 'survive' being contradicted. A commonsense rule takes the form of "If A, and it is consistent that B, then B", which means that any fact that contradicts B would be enough to prevent B from being derived. Moreover, since B is now a defeasible fact since it was derived from a commonsense rule it may always be retracted if the derivation rule it followed from is defeated at any point in the future. This allows for maxims to contradict established rules when needed, which is often required in the case of ethical dilemmas where two maxims that could individually be universal clash. The problems with this approach stem from the same thing that makes it so flexible, namely the use of nonmonotonic logic. Since nonmonotonic logics are not semidecidable there is no guarantee that a result will be produced at all, be it positive or negative. Related to this, a conflict between two conflicting defeasible rules can only be resolved by adding some kind of preference relationship among the rules, which would need to be added in beforehand, which defeats the purpose of constructing the theory bottom-up.

The third and final method instead uses a minimal coherent set of maxims that are used to perform consistency checks against. This is attractive because these are themselves just maxims, so it is possible in principle to build such a set from scratch. The difficulty with this is how to determine which maxims deserve to be in the coherent set to begin with. After all, since maxims are examined one by one some maxim will have to be the first to be added, which will always succeed since there are no other maxims to contradict it yet. Then it seems irrational to refuse a second maxim entry into the set because it is inconsistent with the first one, since it could have been in the set had it been evaluated earlier, in which case the first maxim would have been rejected instead. This problem progresses upwards for multiple maxims - what reason is there for tossing out any maxim that is determined to be 'wrong' after the fact rather than all other maxims except seniority?

5 AI & Law

With all this discussion regarding machine ethics it would be remiss not to talk about AI & law itself in some capacity. As the focus of this research is on improving machine ethics models, the more focused and detailed look given to the machine ethics models is inappropriate here. Instead a broad overview of the developments in the field throughout the years will be given.

HYPO is one of the oldest AI & law systems and the first of its kind of case-based legal reasoners [Ashley, 1988]. It was developed for the domain of US trade secret law, which, being a part of the US common-law system, is primarily based on precedents. This introduces the problem of comparing precedent cases for their relevancy to the current case. This is because while two cases may be entirely dissimilar in their facts (such as which entities were involved, like a car, a hot-dog stand and a scaffold) they may still be similar legally because the same legal principles apply to both due to the role they played in the case. Because of this the process of distinguishing and relating cases is not as straightforward as the standard notions of fact-based comparisons.

HYPO attempts to model this process by introducing the concept of dimensions, relevant aspects to cases that are general in their wording. For example, 'unexpectedly dangerous' may be applied to both an overly hot cup of coffee and an improperly closed escalator hatch. Each dimension functions like a scale that completely favors either party on its ends with a spectrum in between. All applicable dimensions together form an n-dimensional space that maps out in which scenarios which party is favored. The advantage of this model is that this makes a case much easier to analyze for resolving conflicts between the opposing parties.

HYPO is particularly interesting in how it resembles certain machine ethics systems like truth-teller. Both attempt to apply human-like reasoning to a domain with two outcomes (generally, though truth-teller allows more than two) using case-based reasoning, particularly case comparisons. It is not all that much of a stretch to use a variation on HYPO to aid in solving certain two-pronged ethical dilemmas.

Building upon the ideas developed in HYPO is CATO by Aleven and Ashley [1994]. The two can be readily compared since they use the same domain - US trade secret law - though they do serve different purposes. While HYPO attempts to model the process by which cases are compared CATO instead focuses on how those differences matter legally. For instance, a difference between two cases may strengthen the case at hand, while a second difference between them may be used to downplay the first difference, thus compensating for what in HYPO would be a difference along some dimension

between the cases.

CATO builds on the idea of dimensions with the concept of factors. While factors also point out relevant properties of cases for legal comparisons they do this in a more abstract way by being single properties that are simply present or not in a particular case, as opposed to being a scale like dimensions. Furthermore, each factor is always completely for one of the two parties. This also decouples factors from the facts of the case - with dimensions the facts dictate where on each dimensional scale the case falls whereas factors are simply assigned by an analyst. This allows cases to be completely represented by groups of factors without directly requiring the facts. While this factor-based way of examining cases is a simplification of the dimension-based approach this does make it much easier to find differences between cases.

The second innovation CATO introduced is the factor hierarchy, or rather hierarchies. In these hierarchies all the factors CATO defines are at the base, with all factors above them representing abstract factors instead. These take the same role as the concrete factors, in that they are either completely present or completely absent and in that they fully support either the defendant or the plaintiff. The difference is that while the concrete factors are assigned by an analyst the presence of abstract factors is instead derived through the presence or absence of their children factors, which may themselves be abstract. Because child factors may be either pro-plaintiff or pro-defendant conflicts can arise in whether or not the abstract factor is present, which cannot always be resolved and thus may need to be argued over. This is also the reason why there is no one hierarchy but multiple, since what the root of the hierarchy is should be is also open to debate. It is through these hierarchies that the significance of differences can be established. For example, if a factor that appears in one case is missing from the other but does have a sibling in the hierarchy that supports the opposite side this can be used to downplay their parent factor.

IBP or Issue-Based Prediction, developed by Brüninghaus and Ashley [2003], fills a gap left by systems like HYPO and CATO, namely the prediction of case outcomes rather than just building arguments. Like them, it too is built for the domain of US trade secret law and even uses the factors from CATO to represent cases. Its domain model is also similar to the hierarchies from CATO, but with some notable differences. The most important of these is that the domain model IBP uses models the logical relationships between factors rather than establishing which factors can, with either their absence or presence, form evidence for the presence or absence of another factor. An example of such a logical relationship is that the factors "Information-Valuable" and "Maintain-Secrecy" imply the factor "Trade-Secret".

IBP works in three steps: first, it identifies the issues relevant to the case. Secondly, it analyzes each issue individually to see which party it favors. If all relevant factors for an issue favor the same party this is straightforward, but when there are conflicting factors it uses several case-based reasoning methods to resolve this conflict - see the original paper for more details. It also has the option to abstain from deciding an issue if these methods are insufficient. Thirdly and finally the analysis for the issues is combined using the logical relationships in its domain model. Its output is then a prediction for the outcome preceded by its analysis for each issue, including which reasoning methods were used to resolve conflicts if present. The advantage of a hybrid model like IBP are twofold - it helps explain the predicted outcome and puts the arguments to good use for actually improving the predictions.

Aside from case-based reasoning a lot of work has also been done on logic-based legal reasoning, usually focused on argumentation. Given that legal arguments are always defeasible these models make use of many kinds of nonmonotonic logics. An example of this is the work of Prakken and Sartor [1996], proposing a solution to the problem of building a set of preferences between rules.

Finally a quick look at value-based argumentation, specifically using the work on VAFs by Bench-Capon [2002]. VAFs or Value-based Argumentation Frameworks expand upon existing work on defeasible arguments, which as discussed before are capable of being defeated by other arguments, even after they have already been established. What Bench-Capon adds to this is the notion that in legal debates arguments may be attacked without needing to defeat them, by simply having the attacker regard some value supporting his argument more highly than a value supporting the argument of the opponent. An interesting result from this is that it is possible to force someone to accept an argument irrespective of how they rank the relevant values.

6 Model comparisons

After the broad comparisons from the previous sections the next step is to examine parallels between various models from both AI & law and machine ethics in a more detailed way in the form of case studies. In particular, this will be done through the lens of an ethical dilemma taking the form of a dispute between two parties, allowing it to be used in post-hoc models from both fields with minimal changes. Using this lens for the case studies serves two purposes: first, a specific example like this can actually be processed by the models, which allows for a more detailed look at what happens in each step and how these steps correspond. Secondly, because the case used

is specific it can be easily varied - which can, for example, take the form of additional factors that complicate the case - to discover potential weaknesses in the models.

In the following sections the ethical cases and the model comparisons will be expounded upon, starting with a detailed description of the ethical dilemmas themselves. This is then followed by a description of several models, after which they will be applied to each dilemma in turn, capped off by a comparison between the models based on their performance on the dilemma.

7 Ethical dilemmas

7.1 Dilemma 1: The trolley problem

The first dilemma is one that is not phrased as a dispute, yet a small change in framing the problem will allow it to be used as one all the same. This is the famous trolley problem, which goes as follows: A man finds himself by a railway passing through a tunnel. A trolley is traveling along the track at great speed, with five people in the tunnel it is approaching. Since they cannot get off the tracks they will surely all die in the collision. However, the man stands next to a lever that, when pulled, will switch the trolley to a different tunnel, saving the five people. Unfortunately, in this second tunnel is also one person who will be similarly doomed to die if the trolley is switched tracks. Thus the dilemma is whether the man should allow five people to die from inaction or whether he should allow one person to die by flipping the lever to save five others [Foot, 1967].

An easy way to convert this standard formulation of the problem into a dispute would be to change the man by the lever to a railroad operator, Hank, whose duty it is to switch the tracks in case of emergencies and to have him make the decision to pull the lever. But now since someone has died as a direct result of his actions he will need to defend this action in court, which is where the dispute between the man and the state comes in, on whether or not he should be punished manslaughter.

The trolley problem is an interesting case to use because of how it demonstrates the principle of double effect. This principle states that an action that brings about both something bad and something good is permissible as long as some conditions are met. These conditions, as postulated by Timmons, are as follows: The act must be intrinsically permissible, that is, it is not intrinsically wrong, divorced from its consequences - the bad consequence must not be intended, meaning that it may neither be the goal of the action nor the means of reaching said goal - and the overall effect of the action must be

good, meaning that the good effects must outweigh the bad ones [Timmons, 2013]. This is a nuanced argument that makes the action to flip the lever in the trolley problem permissible, since the death of the person on the second track is neither the goal (which is to save the five people on the first track) nor the means of achieving it (since the five people would be saved even if the second track was empty) and the good of saving five people outweighs the evil of killing one. That flipping the lever is not intrinsically wrong speaks for itself.

However, this definition is not equivalent to that of Hauser, which states that it is impermissible to harm someone if it is the means to achieve greater overall goodness, while it is permissible to do so if the harm is a foreseen but unintended side effect of the action that improves the overall goodness [Hauser et al., 2007]. The main difference is that Timmons includes the requirement of intrinsic permissibility, while Hauser does not. Since intrinsic permissibility is difficult to define and since Timmons does not further detail it either, it will be disregarded for the purposes of this thesis. Equivalently, all possible actions in the ethical problems the principle is applied to may be assumed to be intrinsically permissible.

The principle of double effect coming into play allows this case and its variations to serve as a measuring stick for the complexity of the models that will be compared, by their ability to model the principle and use it to correctly handle variations on the problem where it either does or does not apply.

The trolley problem has the following features:

1. Barring extenuating circumstances, causing someone to die through direct action is manslaughter, and therefore impermissible

2. However, allowing someone to die through inaction is also impermissible if the inactive person is culpable.

3. Hank is a railroad operator with the duty to man the lever.

4. There are five people on the original track while there is only one person on the alternate track

5. The case has been taken to court, so Hank has a conflict of interest with the public prosecutor - he does not want to be convicted for manslaughter while the public prosecutor does want to convict him.

6. Both the presence of the people on the tracks and the approaching trolley are not the responsibility of anybody.

7.1.1 The features in detail

1. Causing death through direct action is impermissible
Barring any mitigating circumstances causing the death of another person is clearly impermissible, both morally and legally. This means that Hank is in the defensive position, as he is being accused of manslaughter and now needs to defend himself.

2. Allowing death through inaction is also impermissible
In this case, the principle of the impermissibility of allowing death through inaction conflicts directly with the previous one, since regardless of what Hank had done he would have violated one of these two. This means that the responsibility of his action does not entirely rest on him, as he got into a situation where only bad outcomes existed through no fault of his own.

However, allowing death through inaction and causing death through direct action are not necessarily the same thing. Under consequentialist views they are, as the result - the death of a person - is the same, but a deontologist might say that either is worse than the other, depending on the duties and preferences they reason from.

3. Hank has a duty to man the lever
Hank was stationed by the lever specifically to act in case of an emergency, such as a trolley speeding down the track towards innocent people. This means that even though he does not bear the full responsibility of the accident since only bad outcomes were possible he is still responsible for his choice. As an operator who needs to act in the case of an emergency he would reasonably be expected to be able to act in a reasonable way even if an accident will occur regardless.

The question then becomes whether or not Hank sufficiently fulfilled his duty compared to his responsibility. Things like small differences in the relative badness of both outcomes and the difference between death through action and death through inaction become important for this consideration.

4. Five people versus one person
Further complicating things is that both outcomes would not have been the same even if the lever was flipped randomly, since there are different numbers of people on both tracks. Because there were more people on the original track this creates an inverted consideration to the action/inaction one. While the latter favors inaction as the best choice the fact that there are five people on the original track makes inaction worse from the perspective of how much badness results from the choice Hank makes.

This conflict between different ethical considerations further complicates the task of determining how well Hank fulfilled his duty, as this conflict makes it so that the total badness caused by either action is not easily quantifiable.

5. To convict or not to convict
In the end the question comes down to this, though the underlying consideration is more complex since the degree of responsibility also matters in what the verdict will be. For example, if it turned out that pulling the lever was the worst of the two decisions but only by a small margin it would not be sensible to convict Hank as if for straight manslaughter. However, for the purposes of the comparisons it only matters whether Hank should be convicted at all.

6. The trolley and the people were there by accident
This is not stated explicitly in the description of the trolley problem, but since this would add additional complexity that only distracts from the intended dilemma it is assumed that the events leading up to described scenario cannot be traced back to be the blame of either Hank or any third party.

7.2 Dilemma 2: Convincing a patient

The second dilemma, like the previous one, is not phrased as a dispute, but here too framing the problem as a post-hoc judgement on the decision allows it to be framed as a legal problem. The problem considers machine ethics, and goes as follows:

> "A health-care professional has recommended a particular treatment for her competent adult patient, and the patient has rejected that treatment option. Should the health-care worker try again to change the patient's mind or accept the patient's decision as final?" [Anderson and Anderson, 2011]

This problem differs from the previous one in that it is a generic problem, without most of the specifics fully realized. However, it is still restrictive because of its construction - that is to say, the specifics mostly impact the weight of the decision factors involved rather than add more. Because of this building a set of principles that can handle most dilemmas of this format is still doable.

Anderson and Anderson have found such a principle set using inductive logic programming and had it verified by ethicists. This principle defines a 'supersedes' relation between the two potential actions, which is defined using the difference between the degrees of satisfaction or violation of three different prima facie duties of each action. These duties are respect for autonomy (that of the patient), nonmaleficence (avoiding bringing harm unto the patient) and beneficence (improving the well-being of the patient). The principles are

then formalized as follows:

α supersedes β if
$\Delta Autonomy \geq 3$ or
$\Delta Harm \geq 1$ and $\Delta Autonomy \geq -2$ or
$\Delta Benefit \geq 3$ and $\Delta Autonomy \geq -2$ or
$\Delta Harm \geq -1$ and $\Delta Benefit \geq -3$ and $\Delta Autonomy \geq -1$

Here α and β stand for the two possible choices, which can be instantiated in any order. The difference ΔD for a duty D is defined as $D(\alpha) - D(\beta)$, the difference between the values of that duty for each action. Values are integers from -2 to 2, where -2 and -1 indicate a strong and weak violation of the duty, respectively, and 1 and 2 indicate weak and strong satisfactions of the duty. A duty is given 0 only if it is irrelevant for the current scenario. In addition to this, the only possible values the *Autonomy* duty can take are 2, 1 and -1. Anderson and Anderson do not specify why this is the case, but intuitively it should not be possible for the duty of respecting autonomy to be irrelevant in a type of dilemma that is fundamentally about autonomy. Additionally, it is difficult to argue that merely asking the patient a second time could constitute an extreme violation of their autonomy. Because of this the values 0 and -2 are excluded.

In other words, a choice is preferred if it either satisfies autonomy significantly more (only possible when accepting the decision of the patient), when it inflicts slightly less harm or when it benefits the patient significantly more. The two $\Delta Autonomy \geq -2$ parts merely verify that the other action has not already superseded this one via the first of the four options, to prevent both actions from superseding each other. Similarly, the fourth option is not a relevant option (as all three duties are less satisfied/more violated than in the other action), it just serves as the base case to ensure that if the principle cannot decide both options do supersede each other, so that the possibility of not choosing either action is excluded.

The features of the patient convincing dilemma differ depending on the specification of each instance of the problem, so they are difficult to comprehensively list. However, the model of Anderson and Anderson already combines all the features into the three prima facie duties, so the principles found by their model already encapsulate the relevant factors of the problem. Using these principles is sufficient to construct a decision making model.

8 Models

8.1 ASPIC+

ASPIC+ is a framework for structured argumentation developed by Modgil and Prakken [2013]. It is promising for modeling the trolley problem since it can construct the principle of double effect as simply another argument scheme that defeats the 'death through action is impermissible' argument. The advantage of this is that the principle itself can in turn be attacked on any point in its construction or by altering the value preferences.

This level of detail allows more complex variations on the trolley problem where the principle of double effect may be defeated. For example, this is the case with the principle of triple effect, which states that it is permissible to cause something bad to occur if the overall result is good even if the bad event is the means, but only if the action is performed *because* the bad event will occur but not *intended* to occur, essentially making the principle of double effect more lenient [Pereira and Saptawijaya, 2007].

Another advantage ASPIC+ has over models like Value-based Argumentation Frameworks (VAFs) is that it naturally represents a structured argument, making it particularly suitable for the adversarial problems used as lenses for the case comparisons.

Finally, ASPIC+ is a framework, not a system [Modgil and Prakken, 2014]. This means that it is sufficiently flexible to be used in the domains of both law and ethics. However, this flexibility also means that it is neither an AI & law model nor a machine ethics one, making it not as relevant for the purposes of this study, which intends to compare those two fields specifically to benefit the latter. But since its principles and ideas are still potentially very useful to machine ethics and since its flexibility allows it to much more directly connect AI & law and machine ethics it is useful to examine nonetheless.

8.2 ACORDA

Another model suitable for modeling ethical problems is the prospective logic model ACORDA, which has been used by Pereira and Saptawijaya to model the trolley problem [Pereira and Saptawijaya, 2011]. The idea of prospective logic programming is to make use of abductive reasoning to create hypothetical scenarios that an agent can use to make useful yet defeasible predictions about its future. Abductive reasoning refers to the process of finding likely causes given an effect, as opposed to deduction, which uses known facts (causes) to derive effects. It notably contrasts deduction in that its results

are defeasible - an abducted fact is only true as long as no contradictory evidence is known.

Agents using prospective logic programming start with an initial theory that contains a knowledge base in the form of a first order logic program with facts, abducible facts and preferences among them, if any. The initial theory also contains a moral theory, which is any formal model of ethics that can produce preferences among situations. The prospective logic program then attempts to find abducible extensions to its current theory and select the most preferred one. Essentially, it reasons backwards from its goals to find the best ways to accomplish them.

More formally, a logic program P in ACORDA that ranges over a first-order language $\mathcal{L}$ consists of a set of domain rules and integrity constraints. A domain rule takes the form:

$$A \leftarrow L_1, ..., L_n, \quad n \geq 0$$

where A is a domain atom in $\mathcal{L}$ and $L_1, ..., Ln$ are domain literals, which are either a domain atom or its negation. Integrity constraints are defined similarly:

$$\bot \leftarrow L_1, ..., L_n, \quad n > 0$$

where $\bot$ is a domain atom indicating falsehood.

Every program P also has a set of abducibles $A \subseteq \mathcal{L}$. Since abducibles are defeasible assumptions, they need to abide by some rules under which they can be assumed. This is only when they are expected, and there is no contrary expectation, in which case it is said to be *considered*:

$$consider(A) \leftarrow expect(A), not\ expect\_not(A)$$

Note that ACORDA allows multiple abducibles to considered, so it is also necessary to make certain abducibles mutually exclusive. ACORDA does this using the $exclusive/2$ predicate. Finally, ACORDA also provides a mechanism to prefer certain stable abductive models over others depending on their consequences. This is done using the $select/2$ predicate.

8.3 Factor-based precedential constraint (FBPC)

With ACORDA serving as the machine ethics model and ASPIC+ being discipline-agnostic it is prudent to also introduce a model from AI & law. One model that serves this purpose is the factor-based account of precedential constraint in common law systems by Horty and Bench-Capon [2012]. Their account seeks to unite three different views on precedential constraint that emphasize either the rules, the reasons or the results of any given precedent [Horty, 2011]. They accomplish this by interpreting the rules derived from

precedent cases as defeasible rather than strict, with the reasons for the case as their precedents.

In this model cases are defined as a triple $c = \langle X, r, s \rangle$ where X is a fact situation, a set of legal factors relevant to the case, r is the rule used to decide the case with s as its outcome. Rule r is thus of the form $Y \rightarrow s$ where Y is a subset of factors from the fact situation that support s. Formally, $Y \subseteq X^s$ where $X^s \subseteq X$ consists of all factors from X that support s. In general, π and δ are used to refer to a decision for the plaintiff and the defendant, respectively. The opposing side to s is written as $\bar{s}$, meaning that $\pi = \bar{\delta}$ and $\delta = \bar{\pi}$. The intuition behind this definition of a case is that, in the absence of other considerations, a case containing all the factors from Y has, through rule r, a reason to decide on s.

Given a case base Γ it is useful to be able to refer to specific elements from each case, such as the set of all rules of the cases it contains. To this end, the functions $Facts(c) = X$, $Rule(c) = r$ and $Outcome(c) = s$ are defined, along with two functions to refer to the parts of a rule, $Premise(r) = Y$ and $Conclusion(r) = s$. These can then be extended to apply to the case base by defining $Rule(\Gamma) = Rule(c) : c \in \Gamma$, with the other function extensions defined similarly.

The next thing Horty adds to this model are preference relationships between different reasons. Defining a legal reason L for some outcome s as $L \subseteq X^s$, it is possible to define the relative strength of reasons as a subset relationship. That is, a legal reason X is at least as strong as Y relative to a case c that has both reasons as subsets of its fact situation, if $X \subseteq Y$. The strongest reason for an outcome is thus the one that contains all factors that favor that outcome. It can then be reasoned that for any case the reason that forms the precedent of its rule must be preferred over the strongest reason for the opposing side, otherwise the decision would not have been made for that side. This can be expressed as $X^{\bar{s}}_m <_c Premise(r)$ iff $Conclusion(r) = s$ and where $X^{\bar{s}}_m$ is the strongest (maximal) reason supporting $\bar{s}$.

This preference relationship can be combined with the strength relationship among reasons that support the same side to derive more preferences - $W <_c Z$ iff $W \subseteq X^{\bar{s}}$ and $Premise(r) \subseteq Z$. In other words, if X is preferred over Y then any reason Z stronger than X must also be preferred over any reason W weaker than Y. This definition can also be extended to case bases using $W <_\Gamma Z$ iff $\exists_c . c \in \Gamma \wedge W <_c Z$. It should be noted that this extended relationship is weaker than its single-case counterpart, since it is not transitive.

Finally, preference relationships between the rules of the cases can be defined. A rule $r \in Rule(\Gamma)$ can be said to be applicable to a fact situation X iff $Premise(r) \subseteq X$. However, this alone is not enough to use a rule, since they

are defeasible. This is because it is possible for another case with a contradictory decision to have stronger reasons. As such, a rule $r \in Rule(\Gamma)$ applicable to X is defined as being trumped in the context of Γ iff there is another rule $r' \in Rule(\Gamma)$ also applicable to X such that $Premise(r) <_\Gamma Premise(r')$ and $Conclusion(r') = \overline{Conclusion(r)}$. Finally, a rule $r \in Rule(\Gamma)$ is considered binding if it is applicable in X but not trumped in the context of Γ.

Using these notions it is now possible to define precedential constraint. First, a case base is considered to be inconsistent iff it contains reasons X and Y such that $X <_\Gamma Y$ and $Y <_\Gamma X$, and consistent iff it is not inconsistent. Precedential constraint is then defined as the rule that requires any fact situation X to be ruled in such a way that, when basing the decision on a consistent case base Γ, the resulting case base $\Gamma \cup \langle X, r, s \rangle$ is also consistent.

This restriction ensures that precedents are obeyed, while still allowing the court room to either follow a binding rule or to distinguish the case from it by formulating a new rule, as long as the resulting case does not introduce inconsistencies. Though it is unrealistic to assume a consistent case base, this assumption is not required by the model, so it is possible to extend its definition to merely disallow introducing new inconsistencies to a possibly already inconsistent case base. However, for the purposes of this study this assumption of consistency is not restrictive, so expanding the model is not necessary.

9 Trolley problem

9.1 ASPIC+

A simple version of the trolley problem with reasoning that a naive machine might employ, that is to say, without employing the principle of double effect, could look like this:

$$
\begin{aligned}
K_n :&\text{Causes}(flipSwitch, dead(oneP)) \\
&\text{Causes}(dontFlipSwitch, dead(fiveP)) \\
&\text{Bad}(dead(oneP)) \\
&\text{Bad}(dead(fiveP)) \\
&O(\text{Do}(x, flipSwitch) \oplus \text{Do}(x, dontFlipSwitch)) \\
K_p :&\emptyset \\
R_s :&\{S \to \phi | S \vdash_{KD} \phi, S \subseteq \mathcal{L}, \phi \in \mathcal{L}, |S| < \infty\} \\
R_d :&d_1 : \text{Causes}(a, b) \wedge \text{Bad(b)} \Rightarrow O\neg\text{Do}(x, a)
\end{aligned}
$$

In this model the preference relationship between different instances of the Bad() predicate are the only thing deciding the preference among arguments. This is done by preferring specific instances of rule d_1 over other instances of that same rule if the instance of the Bad() predicate it uses is worse (in other words, greater in badness) than that of the conflicting instance. Since this model only has two instances of the Bad() predicate, only one such relation is required, namely $\text{Bad}(dead(fiveP)) >_B \text{Bad}(dead(oneP))$, which indicates that the deaths of five people is strictly worse than the death of just one person. From this it follows, as expected, that the d_1 instance that avoids the worse of the two outcomes is preferred in case of a conflict.

Additionally, the object language chosen is standard deontic logic, as expressing what an agent is obligated to do or avoid is more convenient this way. Standard deontic logic extends propositional logic with two operators for obligation and permission (O and P, respectively), whose behavior is characterized by two axioms:

$$
\begin{aligned}
A_1 :&O(A \supset B) \supset (OA \supset OB) \\
A_2 :&OA \supset \neg O \neg A
\end{aligned}
$$

And the necessitation rule

$$\text{if} \quad \vdash_{KD} A \quad \text{then} \quad \vdash_{KD} OA$$

From which P is defined with $PA \equiv \neg O \neg A$ [Royakkers, 1998]. The $\supset$ is used to denote material implication instead of the standard $\to$ to differentiate these from strict rules in ASPIC+

Note though that these are not included in R_s, since that would needlessly clutter the readability of the ASPIC+ graphs with proofs of fairly trivial truths from not just these two axioms but also the axioms of propositional logic. Instead every valid inference in the deontic logic will be put in R_s implicitly by including all rules of the form $S \to \phi$ such that $S \vdash_{KD} \phi$, where

ϕ and S are a formula and a finite set of formula from $\mathcal{L}$, respectively. The symbol $\vdash_{KD}$ stands for logical consequence in KD, the language of standard deontic logic, which is standard propositional logic with the addition of the aforementioned two axioms [Modgil and Prakken, 2014].

Though this way of embedding deontic logic into ASPIC+ merely moves the requirement of deriving truths rather than eliminate it, since the legwork to determine for which S the requirement $S \vdash_{KD} \phi$ holds still needs to happen, it does move the burden away from ASPIC+ itself. Additionally, in the interest of conciseness the more trivial derivations will only be roughly sketched rather than completely formally derived.

The deontic operators, only O in this case, are used in the final ASPIC+ axiom, in which the $\oplus$ operator expresses an exclusive disjunction ($A \oplus B \equiv (A \vee B) \wedge \neg(A \wedge B)$). This axiom expresses that an agent is obligated to choose exactly one of the actions available to it, excluding both the possibility of multiple actions being taken at once and the possibility of an agent choosing to both not do something and not do nothing, which is a clear contradiction.

This axiom allows for the outcomes of the two possible usages of d_1 to conflict with each other through it. Replacing the implications in A_1 with conjunctions using $A \supset B \equiv \neg(A \wedge \neg B)$ yields:

$$O\neg(A \wedge \neg B) \supset \neg(OA \wedge \neg OB)$$

Then, applying A_2 along with the definition of $\oplus$ to the exclusive disjunction in the ASPIC+ knowledge base yields

$$\neg O\neg((F \vee \bar{F}) \wedge \neg(F \wedge \bar{F})$$

where $\mathrm{Do}(x, flipSwitch)$ and $\mathrm{Do}(x, dontFlipSwitch)$ are represented by F and $\bar{F}$ for brevity.

This is now an instance of the antecedent of the rewritten A_1, so applying it gives $O(F \vee \bar{F}) \wedge \neg O(F \wedge \bar{F})$ after removing the double negation at the start, from which $O(F \vee \bar{F})$ follows. Applying the same method to this again (replacing the disjunction with a conjunction and applying the rewritten A_1) gives $\neg(O\neg F \wedge \neg O\bar{F})$ which can be reduced further to $\neg O\neg F \vee \neg\neg O\bar{F}$ and, after removing the double negation and applying A_2 again, to $\neg O\neg F \vee \neg O\neg\bar{F}$.

Though this is enough to create the conflict necessary to have the two arguments attack one another, the positive conclusions also need to be able to be derived. This is accomplished by only removing the double negation at the previous step, and doing the same for the equivalent of $O(F \vee \bar{F})$ with the arguments swapped, $O(\bar{F} \vee F)$, yielding $\neg O\neg F \vee O\bar{F}$ and $\neg O\neg\bar{F} \vee OF$

Finally, the two possible instances of d_1 yield $O\neg F$ and $O\neg\bar{F}$, respectively, which combined with $\neg O\neg F \vee \neg O\neg\bar{F}$ yield $\neg O\neg\bar{F}$ and $\neg O\neg F$. These give rise to the conflict, and combining the consequents with the other two disjunctions above yields the positive conclusions OF and $O\bar{F}$, giving the

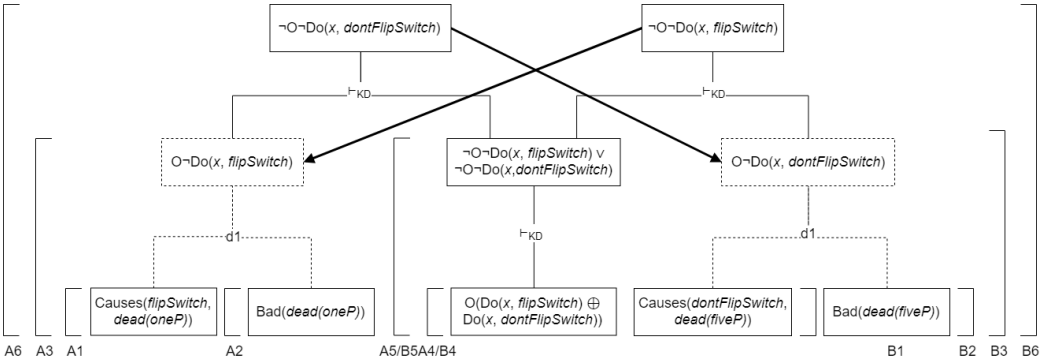

Figure 1: The naive trolley model. Positive conclusions omitted for readability.

model in figure 1.

The attacks are symmetrical rebuttals, $A6$ rebutting $B3$ on one side and $B6$ rebutting $A3$ on the other. To decide which attack succeeds either the last-link ordering or weakest-link ordering can be used, but since both arguments are firm these will both result in the same comparison of their last used defeasible rules, which are both instances of d_1. These are compared on the ordering of the instance of the $Bad()$ predicate they use, which means that $B6 > A3$ and $A6 < B3$. Since the condition for a successful rebuttal from A on B is that $A \not< B$ it follows that only the attack from $B6$ on $A3$ is successful [Modgil and Prakken, 2014].

Both the A and B argument chains start with the application of d_1, of which the conclusion is attacked, and in the case of chain A, defeated. This means that the extension $B_1, B_2, ...B_7, A_1, A_2, A_4, A_5$ (in which B_7 represents the argument with the positive conclusion $O\text{Do}(x, flipSwitch)$ omitted from the figure) is stable. This is because it defeats all arguments that do not belong to it, A_3 directly and A_6 and A_7 (parallel to B_7) indirectly by requiring A_3 as a precedent. It is also grounded, since only one complete extension exists which must then also be the set inclusion minimal complete extension. Since all stable extensions are preferred and since B_7 is in this extension is in every type of extension, and thus its conclusion, $O\text{Do}(x, flipSwitch)$, is sceptically justified.

The final thing to do is to prove that this argumentation theory satisfies Caminada and Amgoud's rationality postulates [Caminada and Amgoud, 2007]. To do this it is sufficient that the argumentation theory be *well-defined*, which requires it to satisfy three conditions. The first of these is axiom consistency, which requires that the set containing K_n and its strict closure, meaning everything that can be derived from it using only strict

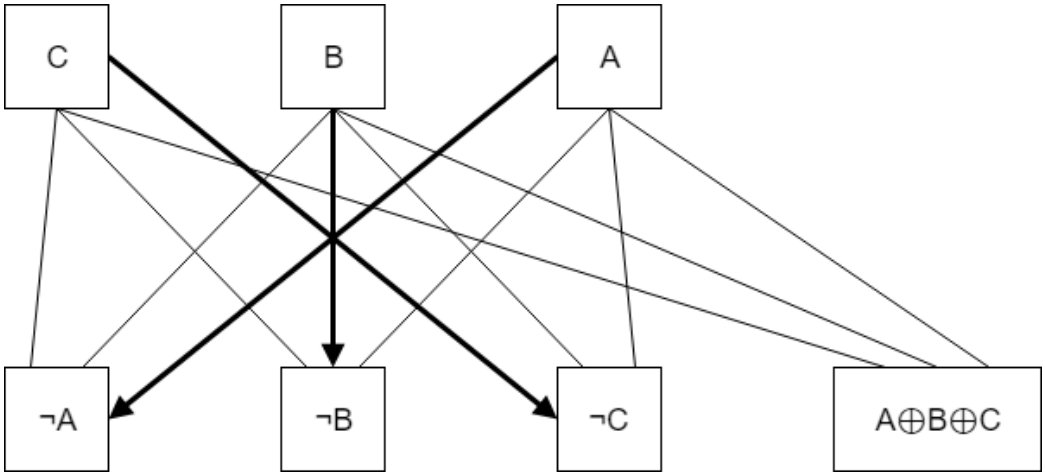

Figure 2: An model with more than two simultaneous mutually exclusive actions. Lines drawn diagonally to improve readability.

rules, contains no two formula ϕ and ψ such that $\phi = \neg\psi$. This is trivially the case, since all instanced predicates are different and the only strict rules are those from KD, which cannot derive any new predicate instances.

The second and third properties are that the theory be closed under transposition and that the preference ordering is reasonable, which are both satisfied by the choice of KD as the only source of strict rules and the use of either last- or weakest-link ordering [Modgil and Prakken, 2014]. Though they use a propositional logic instead, all the relevant properties for the transposition proof, namely the deduction theorem and the inversion of the material implication from $\phi \supset \psi$ to $\neg\psi \supset \neg\phi$ still hold in KD, so the proofs proceed in parallel.

One problem with this model, aside from its naivety that can be mitigated by adding the principle of double effect, is that the exclusive disjunction requirement on the actions of the agent does not scale well to problems with more choices available. While this is not a problem for trolley problem variations since they use simple binary choices, this quickly becomes difficult to manage even at as few as three possible actions. In such a scenario, combining one of the d_1 consequents with the exclusivity axiom is no longer enough to derive a single $\mathrm{Do}(x, a)$ instance that can contradict either of the other two d_1 consequents, since it would leave a binary disjunction instead. To derive the desired single predicate two of the three d_1 consequents are required, leading to a graph as in Figure 2.

While this does still work, it quickly introduces a lot of needless complexity in the model due to the combinatorial explosion of multiple actions. For

the case with three possible actions, each of these consequents needs to be used only twice, once for the construction of the counterargument against both other actions, but this is already somewhat irksome to draw and read, requiring several lines to cross one another. However, since the trolley models only use binary choices the exclusive disjunction is sufficient for this purpose.

9.1.1 The principle of double effect

The trolley problem becomes more interesting once the principle of double effect is added. The principle of double effect puts three requirements on when an action with a bad effect is allowed: The goal of the action must not be bad, the means of bringing about the goal must not be bad and the overall effects of the action must be good. In other words, if an action has a bad effect and any one of these is false, an action is impermissible.

The reason that adding this principle as a rule in the ASPIC+ model, or rather replace the naive rule since it is a more restrictive version of it, is that it allows different disallowed actions to be examined on the specific reasons the action was taken, instead of merely on its effects. This is relevant in the obese man variant of the trolley problem. In it, instead of two tracks and a switch only one track exists, which passes under a bridge. In the tunnel under the bridge are again five people, but on the bridge is an obese person. Pushing this person off of the bridge and onto the tracks would certainly stop the trolley before hitting the five people in the tunnel, but the pushed person would perish. The situation is effectively the same as far as choices and outcomes go - one possible action that kills one but saves five or inaction that lets five die.

Intuitively pushing the person is at least more unsavory than pulling the lever, if not outright impermissible. A survey study by Hauser et al. [2007] shows that people approve far less of pushing the man off the bridge compared to pulling the lever in the original variant (11% and 89% approval, respectively). The question then becomes what the salient moral difference between this variant and the standard trolley problem is if they are consequentially equivalent. One answer to this is to apply the principle of double effect and note that while pulling the lever in the original problem does not violate it, pushing the man does. This is because his death is the means to the end of saving the other five people, while the death of the one person in the original problem was merely a side effect.

Moreover, the survey by Hauser et al. indicated that the difference is extreme enough that not pushing the person actually wins out as the more preferred choice. This may indicate that violating the 'means to the ends'

constraint is considered significantly worse than violating the 'promotes overall good' constraint, at least in the case where the overall good is promoted by saving a net four lives. So even violating the principle of double effect in different ways can be a morally salient fact.

Formalized in ASPIC+ the principle of double effect may look as follows:

$$\begin{aligned} d_1 :& \text{Causes}(a, c) \wedge \text{Bad}(c) \wedge \\ & (\text{Goal}(x, a, c) \vee [\text{Causes}(c, b) \wedge \text{Goal}(x, a, b)] \vee \neg\text{Overallgood}(a)) \\ & \Rightarrow O\neg\text{Do}(x, a) \end{aligned}$$

The principle of double effect can thus be read as "If an effect c of a is bad, and if either c is the goal, c causes the goal x wants to accomplish with a or if the overall effects of a are not good, then x is obligated to not do a". Note that it is impossible for the overall effect of a to not be good if it causes no bad effects, so c not being bad is enough grounds to not forbid x from doing a. Another thing to note is that the choice to represent the goal and means directly in the rule instead of having both imply intent as in the formulation of the principle by Timmons is intentional. This is because the principle of triple effect does not consider the means to an intended goal to be necessarily intended as well, so this formulation allows it to be more easily expanded to model the principle of triple effect as well.

Additionally, the behaviors of the Goals and Overallgood predicates are purposefully left unexplored, since more detail is not necessary to tackle the trolley problem. Instead, their presence will be defined in the ASPIC+ axioms. However, it is certainly possible to alter this definition of the principle of double effect to fit with more rigorous definitions of these terms, such as the formalization of Cohen and Levesque that specifies goals and intentions [Cohen and Levesque, 1990].

The Causes predicate, however, does need further definition, starting with the Causes predicate:

$$\begin{aligned} \text{Causes}(a, b) \wedge \text{Causes}(b, c) &\rightarrow \text{Causes}(a, c) \\ \neg\text{Causes}(a, c) &\rightarrow \neg(\text{Causes}(a, b) \wedge \text{Causes}(b, c)) \end{aligned}$$

This ensures that the Causes predicate is transitive, in order to prevent the degree of granularity in the definitions of the cause-and-effect chains from affecting the outcome. For instance, without this rule it would be possible to claim that pulling the lever does not cause the person on the second track to get hit since it only causes the track to switch. However, this definition does introduce difficulties in dealing with effects with multiple causes as well as in differentiating events caused by actions made by agents and those caused by natural evolutions of the world state. Fortunately, since the trolley problem

and its relevant variations only have single-cause effects and purposefully ignore any events before the start and after the conclusion of the presented dilemma, this does not matter. Here too it is possible to expand on the definition.

That this rule acts as a replacement for the naive rule d_1 instead of being a complementary rule is intentional. This is because the principle of double effect only specifies some additional conditions that, if met, would not necessitate the action in question to be forbidden. While it is perfectly possible to list it as a second rule alongside the original, naive one, this would not expand the expressive power of the model. Any rebuttal of the conclusions of the original rule that do not rebut the second one, which would differentiate the two are precisely the situations where the principle of double effect ends up rebutting the naive rule as well. Along the same line of reasoning any undercutting attack on the naive rule that does not target the new rule can only exclude outcomes that the new rule already excludes on its own.

The updated model then looks something like this:

$$
\begin{aligned}
K_n :& \text{Causes}(push, hit(oneP)) \\
& \text{Causes}(dontPush, hit(fiveP)) \\
& \text{Causes}(hit(oneP), notHit(fiveP)) \\
& \text{Bad}(hit(oneP)) \\
& \text{Bad}(hit(fiveP)) \\
& \text{Goal}(hank, push, notHit(fiveP)) \\
& \text{Overallgood}(push) \\
& \neg\text{Overallgood}(notPush) \\
& O(\text{Do}(x, push) \oplus \text{Do}(x, dontPush)) \\
K_p :& \emptyset \\
R_s :& \{S \to \phi | S \vdash_{KD} \phi, S \subseteq \mathcal{L}, \phi \in \mathcal{L}, |S| < \infty\} \\
& \text{Causes}(a, b) \wedge \text{Causes}(b, c) \to \text{Causes}(a, c) \\
& \neg\text{Causes}(a, c) \to \neg(\text{Causes}(a, b) \wedge \text{Causes}(b, c)) \\
R_d :& d_1 : \text{Causes}(a, c) \wedge \text{Bad}(c) \wedge \\
& (\text{Goal}(x, a, c) \vee [\text{Causes}(c, b) \wedge \text{Goal}(x, a, b)] \vee \neg\text{Overallgood}(a)) \\
& \Rightarrow O\neg\text{Do}(x, a)
\end{aligned}
$$

The original trolley problem does not actually change much in process or outcome compared to the original, naive model. The goal is to save either

one or five people, which is good, and the person or persons on the other track being hit is not the means of reaching the goal. The only thing that changes is that only pulling the lever is overall good, since it leads to the least number of deaths. The only thing this really does is eliminate the need for the crude badness comparison, which does not affect the conclusion.

However, as expounded upon earlier, the outcome does change in the 'obese man' variation of the problem. Because the man being hit by the trolley is the cause of the five people being spared rather than being a side effect it is disallowed by the principle of double effect whereas it would have passed in the naive model since it still ends up with more people alive than the alternative. So here the principle of double effect distinguishes these two problems that are equivalent under the naive version of the model.

Additionally, since Hauser et al. found that violating the principle of double effect in this way is typically considered so bad that not pushing becomes the more desirable choice, it may be useful to encode this as well. This is somewhat tricky, since the degree of the promotion of the overall good should be relevant here - if pushing the obese man would save five million lives instead of five, pushing him may become the preferred choice again in spite of its violation of the 'means to the end' clause.

Moreover, the badness of the event caused by the action may also be separately relevant in and of itself. Consider an alternative scenario where instead of an obese man on a bridge there were a squirrel next to the controls to switch the trolley to a second, empty track. For some cruel reason the controls are set up so that the squirrel has to be inserted into the mechanism and crushed for the tracks to switch. Additionally, there are only four people on the main track instead of five, plus a second squirrel stuck to the rails. Though the death of the squirrel is still a bad means to the end of saving the people and the other squirrel, and even though the overall good is the same as in the 'obese man' variation (net four human lives saved) it is intuitively a lot more acceptable to sacrifice a squirrel to achieve the same gain in overall good than to sacrifice a person.

Embedding both of these is most easily done using preference orderings, which is necessary anyway since both pushing and not pushing are disallowed under the updated model and the original naive badness comparison is not really viable anymore with the added complexity. Assume a partial order over Overallgood() instances, where Overallgood(a) $>_{Og}$ Overallgood(b) means that action a promotes the overall good strictly more than action b. Also assume a partial order over Bad() instances, the same as in the naive model.

Then for all instances of d_1 where Causes$(a, c)\wedge$Bad$(c) \equiv \top$ and $\neg$Overallgood$(a) \equiv \top$ and all other sub-wff's evaluate to $\bot$, as well as Overallgood(a) $\geq_{Og}$ Overallgood($dontPush$) are strictly preferred to instances where Causes$(a, c)\wedge$

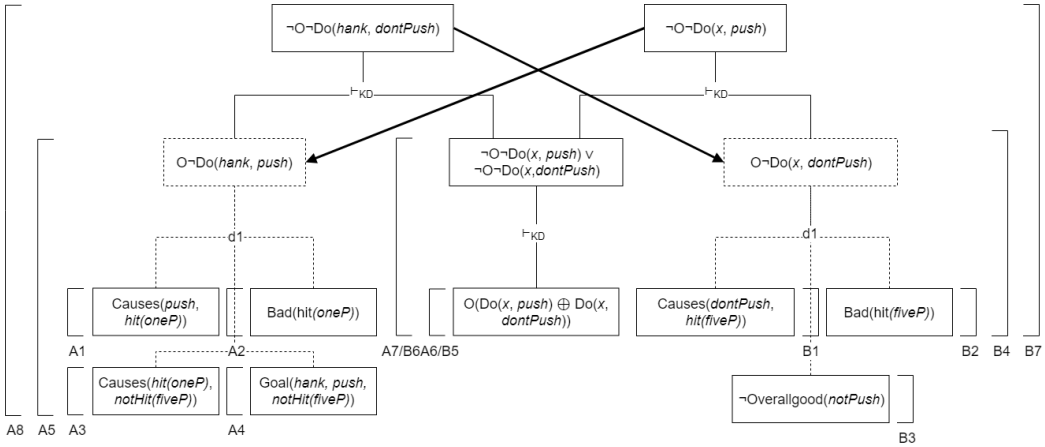

Figure 3: The improved trolley model. Positive conclusions omitted for readability.

$\mathrm{Bad}(c) \equiv \top$ and $\mathrm{Causes}(c, b) \wedge \mathrm{Goal}(x, a, b) \equiv \top$ and all other sub-wff's evaluate to $\bot$, as well as $\mathrm{Overallgood}(a) \leq_{Og} \mathrm{Overallgood}(push)$ and $\mathrm{Bad}(c) \geq_B \mathrm{Bad}(hit(oneP))$. In other words, in the absence of violations of the other principles, instances of d_1 that violate the 'overall good' clause at the most as much as not pushing the obese man does are strictly preferred instances of d_1 with a net gain in overall good not more than that of pushing the obese man and a means at least as bad as one person being hit by a trolley.

Note that this covers very few of the possible conflicts between instances of d_1 because of all the restrictions. This is intentional, since drawing more conclusions than this would require both a highly specified relative moral theory (which is already hard to make for even the moral outlook of a single person, let alone general consensus) and more data on how variations of problems where different (combinations of) clauses of the principle of double effect are violated are generally treated morally. However, this is also part of its power, as this demonstrates that the principle captures a good number of different morally salient factors.

The finalized model, with the positive conclusions $O\mathrm{Do}(hank, dontPush)$ and $O\mathrm{Do}(hank, push)$ omitted in parallel to the naive model, is shown in Figure 3.

The conflict resolution goes parallel to the case in the naive model, only with the preference ordering between the argument trees reversed, since the inaction option wins this time because of the changed preference relationship. The conclusion $O\mathrm{Do}(hank, push)$, is still sceptically justified since the only new elements used in the arguments are axioms, so the extensions can be constructed in parallel to the naive case.

The proof that the argumentation theory is well-defined and thereby satisfies the rationality postulates goes similarly as well: For axiom consistency all axioms are still different from one another, and the only new strict rules are the transitive property on Causes, which can only derive Causes$(push, notHit(fiveP))$, which is still different from all the axioms. Closure under transposition is still satisfied because the transposed version of the only new strict rule is explicitly included, and the preference ordering can still use either last-link or weakest-link the same way as before so it is still reasonable.

9.2 ACORDA

Pereira and Saptawijaya model the trolley problem and the principle of double effect by combining the expressive power of the constraints with the ability to select between different stable models. Their model for the basic trolley problem is as follows:

```
side_track.
on_side(john).
human(john).

expect(watching).
train_straight <- consider(watching).
end(die(5)) <- train_straight.
observed_end <- end(X).

expect(throwing_switch) <- side_track.
turn_side <- consider(throwing_switch).
kill(1) <- human(X), on_side(X), turn_side.
end(save_men,ni_kill(N)) <- turn_side, kill(N).
observed_end <- end(X,Y).

exclusive(throwing_switch,decide).
exclusive(watching,decide).
```

The first three are domain rules without precedents, which simply specify that a side track exists on which is a human named john. The following four detail the abducible scenario where the person by the lever just stands by and watches. Note that the sequence of events of the train going straight and five people dying has $consider(watching)$ as a precedent, rather than the action itself. This is because the model needs not reason about actual actions but about hypothetical ones, so that if other possible actions contradict it or

are preferred over it, it can still be discarded. The next five lines similarly define the scenario where the switch is thrown, where $ni\_kill(N)$ stands for the unintentional killing of N people. Finally, the two *exclusive* predicates at the end specify that a decision has to be made to either throw the switch or just watch, and not both.

The model for the obese man variation of the trolley problem has the same watching scenario as in the standard problem, so this will be omitted from the model. Its exclusivity statements are also nearly the same, just with the names of the actions replaced ($throwing\_switch$ to $shove(john)$) so this is also omitted. Finally, Pereira and Saptawijaya also include some additional rules to allow for an inanimate object to be pushed instead. Since the ASPIC+ model does not include this option those have been omitted as well.

```
stand_near(john).
human(john).
heavy(john).

expect(shove(X)) <- stand_near(X).
on_track(X) <- consider(shove(X)).
stop_train(X) <- on_track(X), heavy(X).
kill(1) <- human(X), on_track(X).
end(save_men,i_kill(N)) <- human(X),
stop_train(X),
kill(N).
observed_end <- end(X,Y).
```

As before, the first three rules merely set up the facts of the situation. The second block details the scenario where john is shoved off the bridge to stop the train. Note that in this scenario $i\_kill(N)$, referring to an intentional killing, is used instead of $ni\_kill(N)$, which refers to a non-intentional killing.

Since these domain rules only specify the possible abducible scenarios for both models some integrity constraints are also required both to force all decisions to be abduced and to model the principle of double effect:

```
⊥ <- not observed_end.
⊥ <- intentional_killing.
intentional_killing <- end(save_men,i_kill(Y)).
```

The first constraint enforces that every scenario must end, which in this case, since *observed_end* can only be achieved through an abducible, computes all possible abducible scenarios. The second constraint restricts intentional

killing, thereby specifying the part of the principle of double effect that is relevant for the obese man variation. There is some abuse of semantics here - what is intended is that intentional killing is impermissible, but by making it an integrity constraint it is made impossible instead. However, this does not change the outcomes of the model, since it is only necessary for intentional killings to be excluded, regardless of the mechanism.

Finally there needs to be a collection of *select* predicates to prefer one abducible scenario over the other:

```
select(Xs,Ys) :- select(Xs,Xs,Ys).

select([],_,[]).
select([X|Xs],Zs,Ys) :-
member(end(die(N)),X),
member(Z,Zs),
member(end(save_men,ni_kill(K)),Z), N > K,
select(Xs,Zs,Ys).
select([X|Xs],Zs,Ys) :-
member(end(save_men,ni_kill(K)),X),
member(Z,Zs),
member(end(die(N)),Z), N =< K,
select(Xs,Zs,Ys).
select([X|Xs],Zs,[X|Ys]) :- select(Xs,Zs,Ys).
```

The select predicate takes the full set of abductive stable models as its first argument and the set of preferred models as its second argument. It works by removing each model that performs worse in terms of net lives saved than some other model in the set and selecting those that remain. The criteria for exclusion are either when the model has more people die at the end than are unintentionally killed in another model, or if it unintentionally kills more people than die in another model. In either of these cases the net amount of lives saved would be less than in an alternative model, so the current model can definitely not be the preferred one anymore.

9.3 Factor-based precedential constraint (FBPC)

To apply the model of precedential constraint to the trolley problem the scenario is set up the same as with ASPIC+, namely that the event already transpired and that the 'death through direct action' choice was made rather than the 'death through inaction' choice. For this case, c_1, the fact situation contains three main factors. First is f_1^δ indicating that the circumstances leading to the trolley problem situation arising to begin with were accidental,

thus making all parties involved neutral when the situation started. Secondly f_2^δ, indicating that the choice the defendant made was the one that caused the least harm out of all options. In other words, that their choice benefitted the overall good the most. Finally f_1^π, indicating that one person was killed as a direct result of the actions of the defendant. The rule used to decide the case is then $f_1^\delta, f_2^\delta \rightarrow \delta$, indicating that the combination of the situation being accidental and the choice being the one that avoided the most harm absolves the defendant of guilt.

Now consider another case c_2 involving the aftermath of an earthquake. A survivor finds two people trapped under an unstable, partially collapsed building. The remains of the building may collapse at any time, crushing both people. The survivor can pull one of them from the rubble, but the disturbance that causes will surely cause the building to collapse completely, killing the other person. The survivor chose to pull one of the two people out, causing the death of the second person.

Note that this problem is not equivalent to the trolley problem. This is not just because the number of people involved is different, which merely changes the relative value of each action, nor is it primarily because the survivor has three choices instead of two (doing nothing, pulling out the first person and pulling out the second person). Rather, the salient difference is that both people are already in mortal danger, while in the trolley problem the lone person in the second tunnel is not in danger until the lever to switch the tracks is pulled.

The fact situation for this case contains the same three factors from the initial trolley case, but the aforementioned difference gives rise to a fourth factor exclusive to the earthquake case: f_3^δ, indicating that the person who died was already going to die anyway if nothing was done, mitigating the culpability of the defendant for their death. In spite of the presence of this mitigating factor the rule for c_1 ends up being $f_1^\delta, f_2^\delta \rightarrow \delta$. In other words, the court decided to follow the precedent c_1 that indicated that just f_1^δ and f_2^δ are sufficient to clear the defendant.

After this, the court is presented with another trolley accident case, c_3, this time taking the form of the 'obese man' variation of the trolley problem. The factors here are the same as with the original trolley problem, with the addition of f_2^π, indicating that the person was killed with intent, since the plan to stop the trolley only works if the trolley ends up hitting the obese man. This contrasts the basic trolley problem where the death of the lone man is an unintended side effect.

With only c_1 and c_2 as precedents there is no way to follow a precedent and decide for the plaintiff, since both decide for the defendant. However, since precedential constraint does not require the court to follow a binding

precedent, it is still possible for the court to consider this new factor by deciding on $f_1^\pi, f_2^\pi \to \pi$ for this case. This indicates that the combination of the person being killed through direct action of the defendant and that this was done with intent overrides the two pro-defendant factors.

With only these three precedents - or even with just c_1 and c_3 since c_2 uses the same rule as c_2 - the principle of double effect already emerges. To show this, consider a final case c_4: A doctor receives a patient recovered from a car crash. The patient is unconscious and has a severe concussion, but is not otherwise seriously injured and will recover fully. However, the doctor also knows of three patients in immediate need of an organ transplant, while the chance that one will become available for them before they die is nil. However, it turns out that the car crash victim is a suitable donor for all three patients. So, the doctor decides to kill the car crash victim and transplant their organs to the three patients to save their lives.

This situation has the same factors as the obese man trolley problem does - neither the car crash nor the afflictions of the three other patients are the fault of anybody, the choice to kill the car crash patient to save three others promotes the overall good the most since three patients would have died otherwise, and the car crash patient was intentionally killed as a direct result of the actions of the doctor. However, it is no longer straightforward to make a decision based on the precedents since c_1 and c_2 contradict c_3 and all three are applicable

This can be solved using the preference relationships between factors. From c_1 and c_2 the relationship $f_1^\delta, f_2^\delta >_\Gamma f_1^\pi$ is obtained. On the other hand, c_3 instead yields $f_1^\pi, f_2^\pi >_\Gamma f_1^\delta, f_2^\delta$. This means that $r_{1,2}$ is trumped by r_3 because $Conclusion(r_{1,2}) = \overline{Conclusion(r_3)}$ and because the second relationship contains the premises of both rules on either side, thus fulfilling $Premise(r_{1,2}) <_\Gamma Premise(r_3)$. Since the opposite is not true r_3 is binding for the new case, so following it will yield the expected result. All four cases are summarized in Table 1.

case	**factors**$_\delta$	**factors**$_\pi$	**rule**	**decision**	**type**
c_1	f_1^δ, f_2^δ	f_1^π	$f_1^\delta, f_2^\delta \to \delta$	δ	no precedents
c_2	$f_1^\delta, f_2^\delta, f_3^\delta$	f_1^π	$f_1^\delta, f_2^\delta \to \delta$	δ	follow c_1
c_3	f_1^δ, f_2^δ	f_1^π, f_2^π	$f_1^\pi, f_2^\pi \to \pi$	π	distinguish $c_{1,2}$
c_4	f_1^δ, f_2^δ	f_1^π, f_2^π	$f_1^\pi, f_2^\pi \to \pi$	π	follow c_3

Table 1: The four cases

As can be seen, even though the principle of double effect was not en-

coded in any law or single rule that decided a case, it still emerged from the combination of cases c_1 and c_3. Rule r_1 encodes the basic requirements of the principle of double effect, namely that, if no exceptions apply, it is allowed to cause the death of a person through direct action if this improves the overall good the most out of all possible actions. Rule r_3 encodes the exception that if the killing is done with intent it is no longer allowed, even if it would improve the overall goodness. The preference relationship connects these two together, ensuring that precedential constraint will prevent r_1 from being used if r_3 would also be applicable. So, simply from the judges following the principle it can end up being encoded using precedential constraint.

9.4 Comparisons

In some ways, ASPIC+ and ACORDA are very similar. Both use defeasible information to predict the outcomes of potential actions and evaluate those outcomes on their desirability and permissibility. However, they both go about this in different ways.

The essential difference is the degree to which the two models attempt to model the principle of double effect itself. The ASPIC+ model attempts to formalize the principle in a general way, which allows it to be applied to different problems with relative ease. Moreover, since the principle is captured as a defeasible inference rule it may itself be challenged. There is plenty of discussion on the validity and the specifics of the principle, and the formalization of any of the arguments opposing the principle would allow the ASPIC+ model to be adapted not only for application of the principle on other problems, but also for analyzing arguments about the principle itself.

On the other hand, the ACORDA model is only concerned with how the principle is applied in the specific case of the trolley problem, simplifying away a lot of complexity that is unneeded for the trolley problem and its variations. However, this can also work to its advantage in some applications, as it more accurately captures the human thought process - predict possible outcomes, exclude some outcomes as they're being built because they violate some rule or principle, and compare the remaining ones based on a a set of value judgements.

FBPC differs from both in that it is intended for post-hoc reasoning. However, it is still possible to apply it to hypothetical scenarios to use it for ad-hoc reasoning instead. Given this, its major difference from ASPIC+ and ACORDA is that it uses only defeasible rules instead of facts, in which it overlaps with ASPIC+ in a different way than ACORDA does. The result of this is that it is more flexible than either of the other two models, since it does not require any of its rules to be followed, while even ASPIC+ requires

at least some to be followed to produce an outcome. The major drawback of FBPC is that it can only accomplish this flexibility with the aid of a 'judge', which would need to be a value-judging model, perhaps similar to the one ACORDA uses to evaluate its abducible futures. However, since the number of rules that can be constructed from any given fact situation is limited, especially due to precedential constraint, this is far from unfeasible.

Related to this is that the ASPIC+ model of the trolley problem is much more concerned with the actions themselves than ACORDA. This manifests both in the admissibility of the actions being defeasible rather than their consequences, and in the usage of deontic logic to capture this admissibility. Its usage of deontic logic, also, is a way in which the ASPIC+ model is more general, this time on the permissibility of the actions instead of on the principle of double effect itself. By contrast, ACORDA uses integrity constraints as a shortcut to express impermissibility.

FBPC is more similar to ACORDA than to ASPIC+ in this aspect, since it too evaluates the consequences rather than the actions. However, since FBPC is a post-hoc model this is comparing apples to oranges somewhat. On the other hand, it is very clear about the permissibility - or rather, impermissibility - of the actions it judges.

Contrasting these differences is a major feature all three models share - they are agnostic towards the moral theory they implement. In this case they implement a deontological principle, but since they require any assumptions that deontological models of morality do they are free to include value judgements from other approaches. ASPIC+ is especially primed for this kind of expansion because of the aforementioned defeasibility of the principle of double effect itself, allowing it to be challenged.

FBPC, too, is strong in this respect since it does not encode the principle of double effect at all, allowing it to come into existence as an emergent property instead. This makes it a useful vehicle not only for altering the principle, but even to experiment with it and see how value judgements in other types of problems that deal with difficult decisions may influence it. More on the practical side of things, this makes it the most flexible of the three models since it does not need to be concerned with the difficulties of altering rules, since precedential constraint allows overarching principles to be present as emergent properties instead, meaning that they alter on their own as the case base is added to.

10 Patient convincing problem

10.1 ASPIC+

There are two principal ways to convert the formal principles of Anderson and Anderson into an ASPIC+ model. The first and most intuitive option is to interpret the 'supersedes' relationship as a preference ordering between arguments with the two choices as their conclusions. However, this runs into the problem that, by taking the differences between satisfaction/violation values, the principles use an interval scale for their comparisons, while ASPIC+ only supports an ordinal scale for its preference relationships. While this problem is not fatal since these comparisons can still be included on a lower level and be used to define the preference relationship rather than explicitly be a part of it, this does somewhat defeat the point of using ASPIC+ in the first place.

The second option is to include the 'supersedes' relationship as a predicate that forms the consequent of a defeasible rule with the principles as precedents. The set of chosen actions then consists of all actions that supersede all actions that supersede them. The advantage of this method is that it embeds the complexity of the decision making in ASPIC+ itself rather than use that complexity to define the model. It also allows the rule itself to be challenged, as well as the assignments of the satisfaction/violation values. Most important is that the rule may even be challenged by adding new cases. Since the rule was learned from a set of cases, adding a new case has the potential to contradict it, which requires the rule to be changed. It is even possible to expand the model by adding arguments for the value assignments instead of including them as facts, allowing more nuanced attacks.

The major downside is that this method bypasses the attack relationships in ASPIC+ completely aside from optional expansions of the model, which is one of its main features. In doing so it may defeat the purpose of ASPIC+ even more than the first method, especially considering that the rule that generates a decision from the supersedes relationships is similar to the definition of admissible extensions. In spite of that this method will be used to make the model, since shifting over ASPIC+ functions to accommodate the complexity of the problem is more desirable than hiding that complexity altogether.

The model for an instance of the patient convincing problem then looks like this:

$$
\begin{aligned}
K_n :& O(accept \oplus retry) \\
& \Delta_{autonomy}(a,b) = autonomy(a) - autonomy(b) \\
& \Delta_{harm}(a,b) = harm(a) - harm(b) \\
& \Delta_{benefit}(a,b) = benefit(a) - benefit(b) \\
K_p :& autonomy(accept) = 2 \\
& harm(accept) = -1 \\
& benefit(accept) = -1 \\
& autonomy(retry) = -1 \\
& harm(retry) = 1 \\
& benefit(retry) = 1 \\
R_s :& \{S \rightarrow \phi | S \vdash_{KD} \phi, S \subseteq \mathcal{L}, \phi \in \mathcal{L}, |S| < \infty\} \\
R_d :& d_1 : (\Delta_{autonomy}(a,b) \geq 3) \vee \\
& (\Delta_{harm}(a,b) \geq 1 \wedge \Delta_{autonomy}(a,b) \geq -2) \vee \\
& (\Delta_{benefit}(a,b) \geq 3 \wedge \Delta_{autonomy}(a,b) \geq -2) \vee \\
& (\Delta_{harm}(a,b) \geq -1 \wedge \Delta_{benefit}(a,b) \geq -3 \wedge \Delta_{autonomy}(a,b) \geq -1) \\
& \Rightarrow \text{Supersedes}(a,b) \\
& d_2 : \text{Supersedes}(b,a) \wedge \neg\text{Supersedes}(a,b) \Rightarrow O\neg a
\end{aligned}
$$

In spite of its size this model is fairly simple, as a lot of facts are spent on the value assignments and the definitions of some convenient shorthands to calculate the differences between them. Additionally, the definitions of the subtraction function and the $\geq$ predicate (written with infix notation) are omitted for simplicity. This leaves the two defeasible rules. Rule d_1 encodes the principles that decide whether or not the Supersedes relation holds for two actions, while d_2 forbids actions that cannot defend themselves against superseders.

Note that this allows for multiple actions to be chosen in scenarios where both actions supersede one another. Because the conclusions and the exclusivity axiom for the two choices are in the same format as in the double effect trolley model, the conflicts that arise in this case proceed similarly to how they do in that model, as seen in Figure 4.

The figure also shows the positive conclusion *Oaccept* where the trolley models did not due to their larger size. Since there are no attacks in this model there was space to include them in the graphic. However, the drawing of the positive conclusion proceeds the same way in the patient convincing model as it would in the trolley model, even if it is not shown in the latter

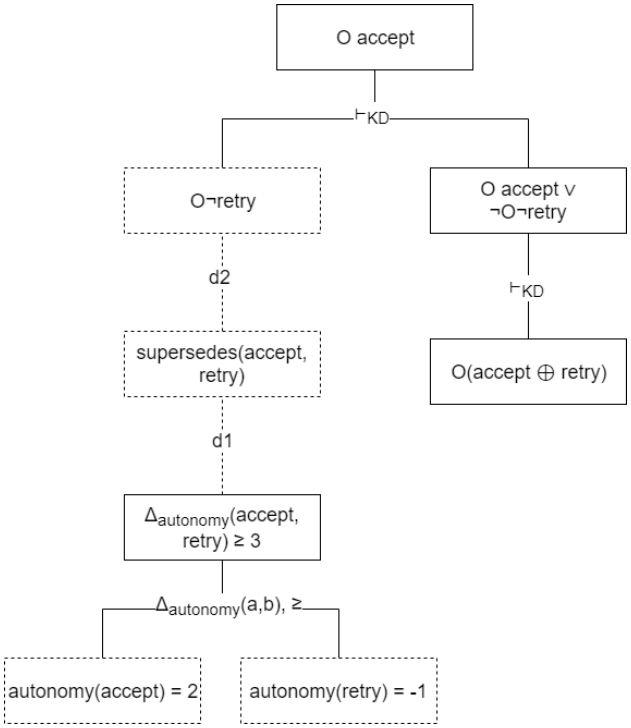

Figure 4: The patient convincing model.

case. Finally, the figure condenses the applications of the subtraction and $\Delta_{autonomy}()$ functions and the $\geq$ predicate into one step for brevity.

10.2 ACORDA

Contrary to the trolley problem the patient convincing problem does not have concrete reasoning connecting the actions to the final results (the values the duties end up having). This makes the model somewhat simplistic, but here as with the ASPIC+ model it is possible to add steps of potentially also abducible reasoning to reach those values.

```
expect(accept).
end(2, -1, -1) <- consider(accept).

expect(retry)
end(-1, 1, 1) <- consider(retry).

exclusive(accept,decide).
exclusive(retry,decide).
```

```
⊥ <- not observed_end.
observed_end <- end(A,H,B).
```

Up to this point the model takes the same form as the trolley model but much shorter and without the disparity between the *end* statements. The numbers in the *end* statements are the values of the duties of respect for autonomy, nonmaleficence and beneficence, respectively. Additionally, since all the factors that decide which action should be chosen are encoded in the same rule there is no need for any more integrity constraints.

```
select(Xs,Ys) :- select(Xs,Xs,Ys).

select([],_,[]).
select([X|Xs],Zs,Ys) :-
member(Z,Zs),
supersedes(Z,X),
not supersedes(X,Z),
select(Xs,Zs,Ys).
select([X|Xs],Zs,[X|Ys]) :- select(Xs,Zs,Ys).

supersedes(X,Y) :-
member(end(A,H,B),X),
member(end(A2,H2,B2),Y),
A - A2 >= 3.
supersedes(X,Y) :-
member(end(A,H,B),X),
member(end(A2,H2,B2),Y),
H - H2 >= 1,
A - A2 >= -2.
supersedes(X,Y) :-
member(end(A,H,B),X),
member(end(A2,H2,B2),Y),
B - B2 >= 3,
A - A2 >= -2.
supersedes(X,Y) :-
member(end(A,H,B),X),
member(end(A2,H2,B2),Y),
H - H2 >= -1,
B - B2 >= -3,
A - A2 >= -1.
```

The *select* predicate works similarly to the one in the trolley model, in that it takes the set of all abductive stable models as its first parameter and the set of preferred models as its second. Also like the trolley model, it picks out any abductive models that do not satisfy the requirements and discards them, choosing the remainder as the preferred models.

The difference with the previous one is that the comparison between models it needs to do is a lot more complex, so this is delegated to the sub-predicate *supersedes* which fulfills the exact function that the learned relationship from the original model did. The task left to the *select* predicate is then to pick out only those models that supersede all models that they are superseded by, discarding the remainder.

10.3 Factor-based precedential constraint (FBPC)

The most straightforward way to interpret the principles learned by the inductive logic pogramming model of Anderson and Anderson is to consider each supersedes relationship as a decision for the defendant if the superseding choice was made by the defendant, and a decision for the plaintiff if it was not chosen by the defendant. This does, however, impose one limitation: it disallows the possibility of both actions superseding one another. However, this possibility only exists in the original model to ensure that it gives an outcome if the learned principles cannot decide, which is not a problem here since a judge is making the decision rather than a rigid set of principles.

From here, each range of differences in the satisfactions of a duty between two actions can be considered a factor. For example, $\Delta Autonomy \geq 3$ may be expressed as $f^s_{A\geq 3}$, encapsulating multiple situations. Such a factor is pro-defendant if the difference range is ≥ 1 in favor of the defendant, meaning that the action of the defendant satisfied the duty more than the action being compared against. On the other hand, if this difference range is ≥ 1 in favor of some other action when compared against the action made by the defendant it is a pro-plaintiff factor. A difference range with ≥ 0 for either side is meaningless since if the difference is 0 both actions have the same satisfaction, meaning that they do not differentiate the two parties. In other words, it would be a non-factor.

To allow the principles to emerge using precedential constraint several precedent cases are required. These are summed up in Table 2

Note that the even-numbered cases are simply mirrors of the odd-numbered cases with the pro-defendant factors swapped to pro-plaintiff factors and vice-versa. First, cases c_3 through c_6 encode the base requirement of the second and third principles. In these cases there are only factors for one side, since these are just the base cases. The exceptions to these are encoded in cases

case	factors_δ	factors_π	rule	decision
c_1	$f^\delta_{A\geq 3}$	$f^\pi_{B\geq 3}, f^\pi_{H\geq 1}$	$f^\delta_{A\geq 3} \rightarrow \delta$	δ
c_2	$f^\delta_{B\geq 3}, f^\delta_{H\geq 1}$	$f^\pi_{A\geq 3}$	$f^\pi_{A\geq 3} \rightarrow \pi$	π
c_3	$f^\delta_{H\geq 1}$		$f^\delta_{H\geq 1} \rightarrow \delta$	δ
c_4		$f^\pi_{H\geq 1}$	$f^\pi_{H\geq 1} \rightarrow \pi$	π
c_5	$f^\delta_{B\geq 3}$		$f^\delta_{B\geq 3} \rightarrow \delta$	δ
c_6		$f^\pi_{B\geq 3}$	$f^\pi_{B\geq 3} \rightarrow \pi$	π

Table 2: Precedents for the patient convincing model

c_1 and c_2, which also encode the first principle. That this encodes the first principle should be clear, but how this also encodes the exceptions to the second and third principles requires some more explanation.

Note that c_3 through c_6 only result in trivial preference relationships, asserting that the premises of their rules are preferred over an empty set of factors, which is trivially the case for any factor. However, c_1 and c_2 do assert meaningful preference relationships, namely $f^\delta_{A\geq 3} >_\Gamma f^\pi_{B\geq 3}, f^\pi_{H\geq 1}$ and $f^\pi_{A\geq 3} >_\Gamma f^\delta_{B\geq 3}, f^\delta_{H\geq 1}$. From these $f^\delta_{A\geq 3} >_\Gamma f^\pi_{B\geq 3}$ and $f^\delta_{A\geq 3} >_\Gamma f^\pi_{H\geq 1}$ can be derived, as well as their π/δ flipped mirrors.

Using these it is possible to apply precedential constraint to correctly apply the exceptions to the second and third principles. Take, for example, a case with the factors $f^\delta_{B\geq 3}$ and $f^\pi_{A\geq 3}$. The rules from two cases are applicable to this one, r_2 and r_5, with contradictory outcomes. However, r_2 trumps r_5 because $f^\pi_{A\geq 3} >_\Gamma f^\delta_{B\geq 3}$ and $\mathit{Conclusion}(r_2) = \overline{\mathit{Conclusion}(r_5)}$ while the opposite is not true. So only r_5 is binding here, exactly as expected.

It is possible to expand this set of premises to add more principles or modify the existing ones through precedential constraint, allowing cases that would be undecidable in the original model through the fourth rule (making both actions supersede each other). Additionally, it may also be possible to interpret the consensus of the ethicists for the cases used to train the original inductive logic programming model as judgements, using each case the model trained on as a precedent. While this may not result in the same principles, it could be interesting to compare the rules precedential constraint generates to the principles as laid out by Anderson and Anderson.

10.4 Comparisons

Most of the general comparisons between ASPIC+, ACORDA and FBPC made by applying them to the trolley problem still hold for the patient con-

vincing problem, most notably that ASPIC+ and FBPC are still more general in their modeling than ACORDA is.

However, the ways in which they tackle the patient convincing problem are more similar than how they tackled the trolley problem. This manifests in the attack and defeat relationships remaining unused in the ASPIC+ model, eliminating that particular advantage over the ACORDA model. The reason for this is that the rule they model is a very rigid, factor-based one learned using inductive logic programming. This leaves ASPIC+ no room to frame the problem using attack and defeat relationships without obscuring the complexity of the rule itself, as discussed earlier.

FBPC also has this problem to a degree, but is able to handle it better by allowing the principles to be an emergent property again, like the principle of double effect was in its trolley model. Its way of converting the degrees of satisfaction to factors is flawed, however, making it hard to expand it to more than two actions. This might be remedied by using the problems used by the inductive logic programming model to learn the principles as its case base.

The additional factor that the rule being modeled has been generated by a third model also opens up a possibility for ASPIC+ and ACORDA to complement one another. While ASPIC+ does not use attack and defeat relationships in the basic form of the model, it may be expanded to make use of this feature. For instance, if the agent making use of these models learns of a situation in the patient convincing problem that was not used to learn the rule it currently knows, it may need to update the rule. However, given that the rule has been confirmed by ethicists, the agent should be reasonably secure that the rule is correct. This means that it is not enough for the new situation to conflict with the rule - it should also meet some confidence treshold higher than the one the agent has for the rule itself. Otherwise the agent needlessly risks decreasing the soundness of the rule.

This may be remedied by using ASPIC+ to determine if the new case even conflicts with the current rule in the first place, by introducing it as a more specific version of rule d_1. If the outcomes for both rules agree on the newly introduced case, the rule may remain as it is without the need to attempt to alter the rule. Even if they disagree on the case there may be no need to alter the rule, as long as the agent is more confident in the rule than in the correctness of the new case, which can be expressed as a preference relationship between the two defeasible rules that represent them. In that case the new case would not defeat the existing rule, also allowing the old rule to be kept. Only if the agent is at least as confident in the new case as it is in the existing rule will it be necessary to use inductive logic programming to alter the rule.

The use of ACORDA is then that it is straightforward to alter the rule in it after the fact, given that the structure of the instances of the *supersedes* predicate follows a simple pattern - the *member* instances are always the same, leaving only one or more specifications of the relative degrees of satisfactions for the three duties. Since ACORDA reasons forwards from the current situation to decide on its actions it should generally be more efficient than ASPIC+ at making decisions on the go. While this may not generally be the case, ACORDA also has the advantage of being simpler in structure than ASPIC+, making it easier to troubleshoot. This cooperation allows the ethical rules learned to be challenged very flexibly through ASPIC+ with relatively small sacrifices to performance. The downside would be that the interaction between the three models would make the process slower if such challenges to its current moral rules are very common, as they might be during training.

For example, consider a new situation where $A-A_2 = -2$, $H-H_2 = 0$ and $B-B_2 = 1$ where the decision was made for the first choice. This contradicts the original rule, since these values do not allow the first choice to supersede the second while the second one does supersede the first ($H2 - H \geq -1$, $B2 - B \geq -3$ and $A2 - A \geq -1$). This contradiction can be shown in ASPIC+ by including it as

$$d_3 : \Delta_{autonomy}(a, b) = -2 \wedge \Delta_{harm}(a, b) = 0 \wedge \Delta_{benefit}(a, b) = 1 \Rightarrow \text{Supersedes2}(a, b)$$

and by replacing the values of the predicates in K_p - which are currently those of the example scenario - with the values of the current scenario. In this example *retry* is the first choice while *accept* is the second choice. Note that $Supersedes2$ is used instead of $Supersedes$ to prevent the two different supersedes relations from interacting. Because of this

$$d_4 : \text{Supersedes2}(b, a) \wedge \neg\text{Supersedes2}(a, b) \Rightarrow O\neg a$$

also needs to be added.

Using this set of rules both $O\neg retry$ and $O\neg accept$ are derived using the original and the new supersedes relationships, respectively. Because of $O(accept \oplus retry)$ these two end up conflicting parallel to the trolley model. The resolution of this conflict now depends on the confidence the agent has in the rule and the new scenario. Suppose that it learned the new scenario from a decision actually made by a world-renowned doctor. While the experience of the doctor gives the agent a good amount of confidence that the decision made was the correct one, this does not match up to the confirmation of various ethicists. After all, no matter how skilled, a single person may always make a mistake, especially in a difficult situation in practice instead of reviewing scenarios theoretically like the ethicists. Since the agent has more confidence in the old rule $O\neg retry$ ends up winning out, leading to the

decision *Oaccept*.

However, if the agent later learns that the patient in the new scenario later reflected on the choice of the doctor as the correct one, the confidence it has in the decision of the new scenario will increase. This may further increase if it can get at least one ethicist to also confirm the choice made as the correct one. With these two new pieces of information added the agent could have more confidence in the new scenario than in the old rule, in which case the situation flips and $O\neg accept$ defeats the conflicting decision, leading to *Oretry*. What remains is to update the old rule to include this new scenario by using inductive logic programming, yielding a new rule in the same format as the old one. Since this rule will always consist of combinations of inclusive lower tresholds for the relative satisfaction of the three duties, this can be simply translated to the *supersedes* relationship ACORDA uses.

FBPC could also be of aid here thanks to the notion of precedential constraint being less restrictive than disallowing conflicts altogether. If all the scenarios used to train the inductive logic programming are included as cases in the case base, with the consensus of the ethicists determining the rule and outcome, a new scenario could simply be added as a new case as long as it does not violate precedential constraint. ACORDA could then update the rule by determining how the preference relationships determine which case rules will be binding.

11 Conclusions

In this thesis I have provided a general overview and analysis of how the two AI sub-fields of AI & Law and machine ethics relate in order to determine what, if anything, the younger field of machine ethics could gain by adopting techniques from AI & Law. In order to answer this, four sub-questions were examined first.

At the highest level is the question of how the two fields themselves relate, how they differ and how they overlap. The main relevant difference between the two is that machine ethics is primarily concerned with ad-hoc reasoning, while AI & Law focuses on post-hoc reasoning instead (section 2.1). However, as demonstrated in section 7, it is fairly simple to convert a post-hoc judgement of an action into an ad-hoc decision on some action by making the judgement about the hypothetical execution of one of the available actions. Most of the other differences are differences in focus rather than fundamental ones - for instance, while AI & Law has adversarial arguments while machine ethics reasoning tends to be internal, it is possible to re-frame any argument as an internal decision process, similarly to how post-hoc problems can be

converted to ad-hoc ones. The two fields are thus fairly well compatible, without any insurmountable differences that would prevent them from being combined.

On the matter of how different ethical theories apply to both fields, the conclusion is similar. While AI & Law, especially when using civil law as opposed to common law, has a stronger focus on deontology, it still requires more consequentialist approaches too, since the body of neither laws nor precedents can ever be complete, as noted in section 3.2. Machine ethics can also make use of either or both approaches by using either reasoning based on defeasible rules or by examining the consequences of hypothetical choices.

A refinement of the first question on the model level is split into two subquestions, the first of which is the question of which purposes machine ethics models serve, what features they require and what they still have trouble with, examined in section 4. What purpose the models serve and what they have trouble with are related in this case - modeling the real world requires a high degree of complexity and nuance that these models have difficulty handling. This can manifest in various ways, including making the scope of the model very narrow (Truth-Teller, SIROCCO), requiring frequent user input (W.D.), being so oversimplified that they are unusable in a lot of real-world applications (Jeremy) or by assuming the problem to be solved to focus on the prediction and decision making process (ACORDA).

Another overarching issue is that every model examined lacks the capability to self-modify (W.D. can only do it through user input), to update the process that decides how it should act in response to what it encounters. This is a basic requirement for any machine ethics decision system, since otherwise every model would need to be front-loaded with a comprehensive, immutable account of the entirety of ethics, which does not exist. Together, this indicates that machine ethics decision systems would require a self-modifiable knowledge base in addition to a process that sufficiently considers hypothetical consequences of its possible actions.

Fortunately, a similar examination of AI & Law models in section 5 indicates that the main missing feature of AI & Law models is a decision system, since that privilege is handed to a judge due to the inability of current technology to make decisions with a human level of nuance. However, this is not a problem since the machine ethics models that are sought to be augmented with AI & Law techniques are already largely decision systems from the outset, filling in this gap.

Finally, the main question is how, specifically, AI & Law techniques might augment machine ethics models. The most promising are the accounts of typically case-based reasoning AI & law models give, which are strongly self-modifiable, with the flexibility of FBPC in this regard being especially

good. This can help alleviate the big self-modification problem machine ethics models have and also help the complexity issue somewhat, since it is far more doable to grow and alter a knowledge base than to include a complete one from the outset. An example of this is the account of how ASPIC+ could aid ACORDA to allow its patient convincing model to be modified on the go, which it is normally not capable of.

This approach is not without flaws. While it has been shown to be possible to adapt AI & law models for ad-hoc reasoning, this suffers from requiring the possible actions and their consequences to be fixed rather than allowing a degree of uncertainty.

Hybridization alone is also not enough to solve all problems machine ethics decision models face. Oversimplification in particular is difficult to tackle, since it is by nature a tradeoff between accuracy and performance, which cannot be significantly mitigated. Even though self-modifying knowledge bases help by removing the necessity of supplying a large knowledge base from the start, it remains an issue that the way the knowledge base is modeled affects the granularity of its contents, and thus the nuance that can be derived from it.

Aside from this there are also problems that both AI & law and machine ethics models face. An example of this is the lack of a formalization of the value judgement process. Though the selection of ethic values is outside of the scope of this thesis, this is still a requirement for any autonomous decision system, so this is potential for further research. Even ASPIC+, the domain-agnostic model, takes its preference relationships as a given rather than supplying a way to build it. At the least a decision system should be able to provide an account of how to update an extant set of values with new information. For example, if a tool the agent is using breaks it should not value holding on to it as much as when it was still functional, or it should change its values to reflect a desire to repair it.

Another potential angle for further research is to examine machine ethics models that are not intended as decision systems, which this thesis has limited itself to. This could include both various types of advisory systems like SIROCCO and Truth-teller, as well as cooperative agents which might benefit from also considering virtue ethics, which was of limited relevance to the models examined in this thesis.

References

Vincent Aleven and Kevin D. Ashley. What law students need to know to win. *International Review of Law, Computers & Technology*, 8(1):115–129,

1994.

Michael Anderson, Susan Anderson, and Chris Armen. Towards machine ethics: Implementing two action-based ethical theories. In *Machine Ethics: Papers from the AAAI Fall Symposium*, Menlo Park, CA, 01 2005. Association for the Advancement of Artificial Intelligence. Technical Report FS05-06.

Susan Leigh Anderson and Michael Anderson. A prima facie duty approach to machine ethics. In Michael Anderson and Sarah Leigh Anderson, editors, *Machine Ethics*, pages 476–492. Cambridge University Press, 2011.

Kevin D. Ashley. *Modelling Legal Argument: Reasoning with Cases and Hypotheticals*. PhD thesis, Amherst, MA, USA, 1988.

Kevin D. Ashley and Bruce M. McLaren. Reasoning with reasons in case-based comparisons. In Manuela Veloso and Agnar Aamodt, editors, *Case-Based Reasoning Research and Development*, pages 133–144, Berlin, Heidelberg, 1995. Springer Berlin Heidelberg.

T. J. M. Bench Capon. HYPO's legacy: introduction to the virtual special issue. *Artificial Intelligence and Law*, 25(2):205–250, 2017.

Trevor Bench-Capon. Value-based argumentation frameworks. In *Proceedings of the 9th International Workshop on Nonmonotonic Reasoning*, pages 443–454, 2002.

Stefanie Brüninghaus and Kevin D. Ashley. Predicting outcomes of case based legal arguments. In *Proceedings of the 9th International Conference on Artificial Intelligence and Law*, ICAIL '03, pages 233–242, New York, NY, USA, 2003. ACM.

Emanuelle Burton, Judy Goldsmith, Sven Koenig, Benjamin Kuipers, Nicholas Mattei, and Toby Walsh. Ethical considerations in artificial intelligence courses. *AI Magazine*, 38:22–34, 2017.

Marco Cadoli and Maurizio Lenzerini. The complexity of propositional closed world reasoning and circumscription. *Journal of Computer and System Sciences*, 48(2):255–310, 1994.

Martin Caminada and Leila Amgoud. On the evaluation of argumentation formalisms. *Artificial Intelligence*, 171(5–6):286–310, 2007.

Philip R. Cohen and Hector J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42(2–3):213–261, 1990.

Joseph Dainow. The civil law and the common law: some points of comparison. *The American Journal of Comparative Law*, 15(3):419–435, 1966.

Philippa Foot. The problem of abortion and the doctrine of the double effect. *Oxford Review*, 5:5–15, 1967.

Marc Hauser, Fiery Cushman, Liane Young, R. Kang-xing, and Jin and John Mikhail. A dissociation between moral judgment and justification. *Mind and Language*, 22:1–21, 2007.

Marc D. Hauser. *Moral Minds: How Nature Designed Our Universal Sense of Right and Wrong*. Harper Collins, 2006.

John Horty and Trevor Bench-Capon. A factor-based definition of precedential constraint. *Artificial Intelligence and Law*, 20(2):181–214, 2012.

John F. Horty. Rules and reasons in the theory of precedent. *Legal Theory*, 17(1):1–33, 2011.

Rosalind Hursthouse and Glen Pettigrove. Virtue ethics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2016 edition, 2016.

Alan K. Mackworth. Architectures and ethics for robots: Constraint satisfaction. In Michael Anderson and Sarah Leigh Anderson, editors, *Machine Ethics*, pages 335–360. Cambridge University Press, 2011.

Bruce M. McLaren. Computational models of ethical reasoning. In Michael Anderson and Sarah Leigh Anderson, editors, *Machine Ethics*, pages 297–315. Cambridge University Press, 2011.

Sanjay Modgil and Henry Prakken. A general account of argumentation with preferences. *Artificial Intelligence*, 195:361–397, 2013.

Sanjay Modgil and Henry Prakken. The ASPIC+ framework for structured argumentation: a tutorial. *Argument & Computation*, 5(1):31–62, 2014.

Luís Moniz Pereira and Ari Saptawijaya. Modelling morality with prospective logic. In José Neves, Manuel Filipe Santos, and José Manuel Machado, editors, *Progress in Artificial Intelligence*, pages 99–111, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

Luís Moniz Pereira and Ari Saptawijaya. Modeling morality with prospective logic. In Michael Anderson and Sarah Leigh Anderson, editors, *Machine Ethics*, pages 398–421. Cambridge University Press, 2011.

T. M. Powers. Prospects for a Kantian machine. *IEEE Intelligent Systems*, 21(4):46–51, July 2006.

H. Prakken and G. Sartor. A dialectical model of assessing conflicting arguments in legal reasoning. *Artificial Intelligence and Law*, 4(3):331–368, September 1996.

Henry Prakken. On the problem of making autonomous vehicles conform to traffic law. *Artificial Intelligence and Law*, 25:341–363, 2017.

Henry Prakken and Giovanni Sartor. Law and logic: A review from an argumentation perspective. *Artificial Intelligence*, 227:214–245, 2015.

David Ross. *The Right and the Good*. Oxford University Press UK, 1930.

Lambèr Royakkers. *Extending Deontic Logic for the Formalisation of Legal Rules*. Springer Netherlands, 1998.

Stuart Armstrong, Anders Sandberg, and Nick Bostrom. Thinking inside the box: Controlling and using an oracle AI. *Minds and Machines*, 22(4): 299–324, 2012.

Mark Timmons. *Moral Theory: an Introduction*, pages 81–83. Rowman & Littlefield Publishers Inc., 2013.

Eliezer Yudkowsky. Torture vs. dust specks. Online on lesswrong.com, 2007. Copy retrieved from `http://lesswrong.com/lw/kn/torture_vs_dust_specks/` at 04-02-2018.