RIGHTFUL MACHINES

by

AVA THOMAS WRIGHT

(Under the Direction of Frederick Maier)

ABSTRACT

In this thesis, I set out a new, Kantian approach to resolving dilemmas and other conflicts of obligation for semi-autonomous machine agents such as self-driving cars. The approach begins with the modern distinction between law and ethics, and looks to a standard of justice (rather than ethics) to determine how to resolve conflicts of obligation such as in what is known as the "trolley problem." Rather than building machines that reflect one or another group's ethical preferences, efforts to build explicitly moral machine agents should focus on building *rightful machines*. I propose that "answer set programming," which can be understood as an efficient machine implementation of non-monotonic forms of reasoning through its answer set /stable model semantics, is a workable engineering solution for handling deontic conflicts for rightful machine agents. I critically evaluate two prior efforts in this area and demonstrate the new approach to conflicts using answer set programming.

INDEX WORDS: Priority of Right, Machine Ethics, Conflicts, Dilemmas, Trolley Problem, Machine Agents, Answer Set, Logic Programming

RIGHTFUL MACHINES

by

AVA THOMAS WRIGHT

BA, Rice University, 1995

JD, Georgia State University, 2000

MA, Georgia State University, 2010

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2018

© 2018

Ava Thomas Wright

All Rights Reserved

RIGHTFUL MACHINES

by

AVA THOMAS WRIGHT

Major Professor:	Frederick Maier
Committee:	Walter D. Potter
	Sarah Wright

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
December 2018

TABLE OF CONTENTS

Page

CHAPTER

1 INTRODUCTION..1

2 RIGHT, ETHICS, AND THE PRIORITY OF RIGHT.....................................5

The priority of right over ethics...10

Duties of rightful machines...13

3 SOLVING THE TROLLEY PROBLEM I: FAT MAN..................................16

A trolley *non*-problem: Bystander...20

4 SOLVING THE TROLLEY PROBLEM II: DRIVER...................................26

The priority of right in Driver..30

Self-defense, legal necessity, and dilemma..32

5 NORMATIVE CONSISTENCY AND DEONTIC LOGIC............................38

The inadequacy of the standard system of deontic logic...........................39

Non-monotonic deontic logic...43

Logics of belief revision...47

6 ANSWER SET PROGRAMMING LEGAL RULES AND CONFLICTS.....50

Evaluating conflicts of legal duties in answer set programming 60

7 ANSWER SET PROGRAMMING KANT'S CONFLICT BETWEEN VERACITY AND PHILANTHROPY..75

A Flawed Attempt to Encode the Conflict Between Veracity and Philanthropy in SR..77

Answer set programming the conflict in SR..82

The indefeasibility of the duty of veracity in SR......................................90

8 ANSWER SET PROGRAMMING THE TROLLEY PROBLEM..................95

A Flawed Logic Programming Approach to Bystander.............................96

Answer set programming the trolley problem...103

9 CONCLUSION...108

REFERENCES...110

CHAPTER ONE

INTRODUCTION

In a recent massive experiment conducted online, millions of subjects were asked what a self-driving car whose brakes have failed should do when its only choices were to swerve or stay on course under various accident conditions (Awad, et al., 2018). Should the car swerve and kill one person in order to avoid killing five people on the road ahead? Most subjects agreed that it should. Most subjects also agreed, however, that the car should generally spare younger people (especially children) over older people, females over males, those of higher status (e.g., the rich) over those of lower status, and the fit over the overweight, with some variations in preferences correlated with subjects' cultural backgrounds. But while such results may be interesting, I will argue that they are largely irrelevant to the question as to what a self-driving car faced with such a dilemma should do.

Efforts to build explicitly moral machine agents such as self-driving cars should focus on *duties of right*, or justice, which are in principle legitimately enforceable, rather than on duties of virtue, or ethics, which are not. While hypothetical dilemmas such as the (in)famous "trolley problem" (which inspired the experiment above) have received enormous attention in machine ethics, there will likely never be an ethical consensus as to their correct resolution, and even if one could be achieved, it would be largely irrelevant to the problem. What matters is whether machine agents charged with making decisions that affect human beings act rightfully, that is, in ways that respect real persons' equal rights of freedom and basic principles of justice. Whatever resolution of dilemmas such as the trolley problem one prefers ethically, it is the law that determines when makers and users of semi-autonomous machines such as self-driving cars will be liable or culpable for the machine's decisions, and law must conform to principles of justice, not the partial ethical preferences of one group or another.

In this thesis, I set out a new, Kantian approach to resolving dilemmas and other conflicts of obligation for semi-autonomous machine agents such as self-driving cars. The approach begins with the modern distinction between justice and ethics, and looks to a standard of justice (rather than ethics) to determine how to resolve conflicts of obligation such as in the trolley problem. An action is just, Kant says, when it "can coexist with the freedom of every other under a universal law;" therefore, the rightfulness of any act is specified explicitly in terms of its consistency within a system of equal rights of freedom. I interpret this consistency not descriptively but as a normative requirement that justice imposes upon any legal system of enforceable duties and rights.

Hence when dilemmas between strict legal obligations such as in the trolley problem arise, we should not conceive them as cases where we are forced to violate one or another of our inconsistent obligations but, instead, as cases where we must revise legal obligations and rights in order to meet the normative requirement of consistency in a system. The legislative, executive and judicial institutions of the civil state are necessary, Kant argues, to construct and maintain such a system for human beings in social interaction. The shift from ethics to a standard of justice clarifies dilemmas such as the trolley problem and other conflicts of duties

I then take up issues of implementation and consider whether and how systems suitable for governing explicitly rightful machines can meet normative requirements of justice such as consistency. I suggest that non-monotonic deontic logical approaches to conflicts of obligation such as that implemented in answer set or logic programming can meet the consistency requirement, though with certain reservations.

Finally, I review two recent prior efforts to apply answer set or logic programming to model conflicts of obligations. Both attempts fail to sufficiently observe the distinction between right and ethics, and the priority of right. The first models a conflict between a duty of truthfulness and a competing duty of philanthropy (Ganascia 2007). The second models a number of variations of the trolley problem (Pereira and Saptawijaya 2011). I criticize each and then apply the answer set programming approach developed in this thesis.

Semi-autonomous machine agents that learn in an open-ended manner act in ways that are unpredictable by design. When they interact with real human beings, their behavior must, therefore, be subject to some constraining normative governance system. What these constraints are and how they should be implemented in such a system is the problem I take up in this thesis. I only point the way to rightful machines here, however, and hope that further research may fill out a more complete system.

CHAPTER TWO

RIGHT, ETHICS, AND THE PRIORITY OF RIGHT

1. Kantian Justice

In the *Doctrine of Right* (DR), Kant defines the "Universal Principle of Right" (UPR) as follows:

> Any action is *right* if it can coexist with the freedom of every other under universal law; or if on its maxim the freedom of choice of each can coexist with everyone's freedom in accordance with a universal law (DR: 6:230).

Kant thus defines the legal permissibility (rightfulness) of an action in terms of its systematic consistency with everyone else's equal rights of freedom under universal law. If the act is consistent with everyone's equal rights under universal law, it is permissible. While Kant defines legal permissibility here, permissions, duties and (claim-)rights are logically interdefinable by taking any one as a primary operator (see Hohfeld 1919: 35-50).

Kant reiterates justice as systematic freedom under universal law when defining the right of freedom:

> *Freedom* (independence from being constrained by another's choice), insofar as it can coexist with the freedom of every other in accordance with a universal law, is

> the only original right belonging to every [person] by virtue of his humanity (DR: 6:237).

Hence while *freedom* is '"independence from being constrained by others," according to Kant, the *right of freedom* is that freedom systematically limited by everyone else's equal right of freedom under universal law. The right of freedom lacks definition outside a system of equal rights of freedom under universal law.

But how are we to determine the shape and scope of the equal right of freedom in the system? Freedom as independence from constraint is neither self-explanatory nor necessarily self-limiting. Maximizing rights of freedom consistently in a system is both vague and ambiguous (Hart 1973: 547-50; see Rawls 1993: 291-92)). For example, rules of debate make equal rights to speak consistent by systematically limiting them, but the rules are specified by reference to the goal of a good debate, not by reference to the vague end of maximizing equal freedoms to speak (Hart 1973: 543).

In both the UPR and the definition of the right of freedom, Kant says that actions or principles of action ("maxims") must not only coexist consistently with a system of equal freedom but also be "in accordance with a universal law." This is a reference to what Kant identifies as the supreme principle of morality, the "categorical imperative:"

> [A]ct only in accordance with that maxim through which you can at the same time will that it become a universal law (GM: 4:421).

The categorical imperative eliminates principles of action that cannot be universalized without a contradiction in the will. For example, suppose I make a false promise to repay you a loan in order to get quick cash. In a world where everyone did that, my promise would fail to achieve my end since you would not believe me, and indeed, such promises would be inconceivable in such a world (see GM 4:422). I therefore cannot coherently will to use such a promise to achieve my end and at the same time will that everyone do so when in my position. I would will to make an unreasonable exception for myself.

Kant articulates the categorical imperative in four different "formulations," which he claims are equivalent. The "Formula of Humanity" requires that you use the "humanity" (i.e. rational nature) in persons "always at the same time as an end, never merely as a means" (GM: 4:429). Kant explains that you use the humanity in another person as a mere means when she "cannot possibly agree" to the principle of your action with regard to her (GM: 4:429). So, again, if I falsely promise to repay a loan in order to trick you into giving me quick cash, you cannot possibly agree to my deceit because its success depends upon your ignorance of it (GM: 4:429-430). Even if you would be willing to freely give me the money, you cannot possibly agree to be deceived into giving it to me. Note that Kant's example here is of a violation of a duty of right that is a criminal offense in the law (i.e., fraud).

Kant's possible agreement gloss on the formula of humanity of the categorical imperative is the key to understanding how justice shapes equal rights of freedom in a system "under universal law." According to Kant, public laws that shape duties of right are just if and only if *everyone could consent* to them (see O'Neill 2011: 170-185). Kant articulates justice as possible universal consent most clearly in his political essay, "On the Common Saying: 'That May Be Correct in Theory but It Is of No Use in Practice'" (T):

> For this is the touchstone of any public law's conformity with right...if a public law is so constituted that a whole people *could not possibly* give its consent to it (as, e.g., that a certain class of *subjects* should have the hereditary privilege of *ruling rank*), it is unjust; but if it is *only possible* that a people could agree to it; it is a duty to consider the law just... (T: 8:297).

Laws structuring rightful relations with others must be such that it is possible for everyone to consent to them, according to Kant. What this implies is that some positive public law is normatively *obligatory* to secure necessary conditions for the possibility of universal consent, while the rest of positive public law is *permissible* in the service of other aims, so long as it does not violate those conditions. Kant's modal standard of legitimacy thus both generates some necessary (obligatory) positive law and simultaneously restricts all possible (permitted) positive law.

The logic of the possible universal consent standard can be captured in the following valid modal logical argument scheme (where '$\Box$' is 'juridically necessary' and '$\Diamond$' is 'juridically possible'):

1. $\Box\Diamond(x)Cx$ (It must be possible for everyone to consent to public law.)
2. $\Box\,(\Diamond(x)Cx \rightarrow [\text{conditions}])$ (Certain conditions must be met for universal consent to public law to be possible.)

Therefore, $\Box[\text{conditions}]$ (Therefore, meeting those conditions is juridically necessary.)

In the possible worlds semantics of modern modal logic, possibility (or false necessity) *generates* an accessible possible world, whereas necessity (or false possibility) quantifies over all accessible possible worlds. Kant's standard indirectly exploits this feature of modal logical semantics to entail conditions for both obligatory and permissible positive law.

Kant's possible universal consent standard focuses attention on the necessary conditions for the normative authority of consent as a normative power, rather than on an answer to the question as to whether some apparent act of consent is actual or accords with one's preferences or not. Kant sets out substantive constitutional principles that public law must meet to satisfy the standard:

> A constitution established, first on principles of the *freedom* of the members of a society (as individuals), second on principles of the *dependence* of all upon a single common legislation (as subjects), and third on the law of their *equality* (*as citizens of a state*)—the sole constitution that issues from the idea of the original contract, on which all rightful legislation of a people must be based—is a *republican* constitution (PP: 8:349-50).

Any constitution that lacks provisions to establish these principles—freedom, the rule of law, and equality—cannot secure the possible consent of all, Kant claims. Kant's rationale for this claim appears to be that one has no normative authority to consent to changes in one's normative relations with others that would violate conditions required to exercise the normative power of consent. The argument is thus similar in form to that sometimes deployed against the legitimacy of slave contracts. Even if it were possible to alienate the innate right of freedom, to do so would be to render oneself a thing rather than a person and, therefore, incapable of being legally bound by the slave contract (see DR: 6:241).

Kantian justice is thus characterized by two main normative requirements: 1) equal rights of freedom, together with correlative duties and permissions, must be specified within a consistent system of such rights, and 2) public laws specifying rights, duties, and permissions in the system must be such that it is possible for everyone to consent to them. The latter requirement imposes a number of substantive and procedural criteria for just law, some of which Kant sets out in what he refers to as the "republican" constitution.

2. The priority of right over ethics

It is an axiom of modern, post-Enlightenment moral philosophy that every person capable of autonomy is equally free. What equal freedom immediately implies is that it is impossible to force another person to act ethically, that is, to act *for ethical reasons*. One can force others to act in ways that conform outwardly with their ethical duties, perhaps by threatening them with punishment if they fail to comply, but then they would be acting merely to avoid being punished (DV: 6:381). Hence ethical duties are unenforceable. One can only enforce the public or outward aspects of duties; one cannot make people act ethically. This is one aspect of the priority of right.

The other, more important, aspect of that priority is that it is the very purpose of the system of public laws to authoritatively construct and enforce duties and rights when they come into conflict with each other. *Equal* innate freedom implies that each person has her "own right to do *what seems right and good to* [her] and not to be dependent on another's opinion about this," Kant says (DR: 6:312). No one individual has the moral authority to unilaterally define what everyone's moral rights and duties are with respect to others (i.e., legislate them), or to enforce them (i.e., execute them), or to resolve disputes (i.e., adjudicate them). Reason alone cannot a priori determine our rights and duties with respect to each other, Kant says (DR: 6:312). Even if everyone were committed to being perfectly ethical, wronging one another in social interactions is inevitable in the absence of public law, since "when rights are i*n dispute* (*ius controversum*), there would be no judge competent to render a verdict having rightful force" (DR: 6:312). The solution is to construct

> *...a system of laws for a people...which because they affect one another, need a rightful condition under a will uniting them, a constitution (constituto)*, so that they may enjoy what is laid down as right' (DR: 6:311, emphasis in original).

Kant refers to this system of public laws as "public right," and a society existing under such a system as one existing in a "rightful" or "civil" condition, as opposed to one in a "state of nature." Only by constituting a "united will" to authoritatively determine, enforce, and adjudicate our rights and duties with respect to each other, can we avoid wronging each other, Kant argues. We, therefore, have a duty to enter into a civil condition with others:

> [The rational will requires that] it must leave the state of nature, in which each follows its own judgment, unite itself with all others (with which it cannot avoid interacting), subject itself to a public lawful external coercion, and so enter into a condition in which what is to be recognized as belonging to it is determined *by law* and is allotted to it by adequate *power* (not its own but an external power); that is, it ought above all else to enter a civil condition (DR: 6:312).

Hence determinations made in the system of public laws as to what rights and duties everyone interacting in community has take *normative priority* over individual ethical judgments in particular cases. To reject public authority and use one's own individual judgment in cases of conflict is to act wrongly and unethically, indeed, to commit wrong "in the highest degree," Kant says (DR:6:308n). Resolving such conflicts in order to avoid wronging one another is the very purpose of the system of public laws. This is the second aspect of the priority of right, and its central meaning.

Hence the priority of right has two aspects. I cannot force you to be ethical, but even if I could, it would be wrong (immoral, both wrongful and unethical) for me to ignore public law and determine on my own what you and I morally should do when conflicts arise. I commit wrong in the highest degree when I ignore public law in such cases. This normative obligation to submit to the just law of a public authority is ultimately rooted in respect for the "humanity" in oneself and, equally, in others.

Much more could be said concerning the priority of right over ethics and there are a number of other complications, but my aim here is only to set out why Kant endorses the idea. There is no question that Kant does, as I think must every modern philosopher who endorses moral equality in some form. For example, the utilitarian John Stuart Mill's principle of justice, the "Harm Principle," arguably states a version of the priority of right, and John Rawls defends the priority of right explicitly and at length (Mill, 223-4; see Rawls 2001: 41). The priority of right thus has broad application in modern value theory.

3. Duties of rightful machines

Duties of right concern only the public, outward aspects of one's actions and are thus completely specifiable without reference to the agent's intent or "maxim" of the end of action. For example, while one has a general moral duty to keep one's promises, one has a legal duty of right to keep only those promises that meet the outward, public criteria that legitimate public authority has defined as a contract, such as offer, acceptance, consideration, etc. Whether I perform on the contract in order to honor my promise or solely because I fear a civil suit, I meet my legal contractual obligation just the same. Similarly, I meet my legal obligations to avoid criminal acts such as theft and murder even if I avoid them solely because I fear punishment. Corresponding ethical duties, by contrast, require me to avoid such crimes because they are wrong.

The rightful enforceability and precise specifiability of duties of right have important implications for builders of explicitly moral machine agents. First, the precision required in the specification of duties of right should make such duties easier to capture in governance systems. Second, rightful machines sidestep objections related to the agent's capacity for freedom. If a machine cannot act according to an ethical principle that it freely chooses, then the machine cannot act ethically and can at best produce only a simulacrum of ethical action (Guarini 2012). But if, on the other hand, advanced machines in the future become capable of autonomous ethical agency, then installing a coercive explicitly ethical governance system would violate the *machine's* right of freedom (Tonkens 2009). By contrast, duties of right require no particular (or any) subjective incentive for action; hence mere conformity with the outward aspects of such duties is sufficient to act rightfully. And on the other hand, duties of right are rightfully enforceable; hence a coercive governance system may not violate even a genuinely autonomous machine's rights. Finally, and perhaps most importantly, since ethical duties are not rightfully enforceable against those who violate them, explicitly *ethical* machines (that is, agents that act according to ethical principles) may often act wrongfully, and it is not difficult to imagine dystopias where machine agents paternalistically manage human affairs in the service of partial ethical ideals. By contrast, machines that conform to duties of right will by definition respect real human persons' rights of freedom and avoid paternalistic ethical meddling.

Self-driving cars and other machine agents programmed to act in accordance with popular ethical intuitions would be neither ethical nor rightful machines, and instead, seem to me to pose a threat to civil society. The goal of machine ethics should be *rightful machines*.

CHAPTER THREE

SOLVING THE TROLLEY PROBLEM I: FAT MAN

1. The original trolley problem: Fat Man versus Driver

Consider the original ("Driver") version of the "trolley problem" (Foot 1967: 3): Imagine you are driving a trolley whose brakes have failed. The runaway trolley, gaining speed, approaches a fork in the tracks, and you must choose which track the trolley will take. On the main track are five people who will be struck and killed if you stay on course, while on the side track is one person who will be struck and killed if you switch tracks. What are you obligated to do? In polls and experiments, most people (about 90%) say they would turn the trolley (Mikhail 2007).

Now contrast Driver with the following variation ("Fat Man") (Thomson 1976: 207-8): Imagine that instead of driving the trolley, you are standing on a footbridge overlooking the tracks. The five are still in jeopardy in the path of the runaway trolley, but now there is no side track. Standing next to you on the footbridge is a fat man leaning over the footbridge railing. You suddenly realize that you could stop the trolley and save five people if you pushed the fat man off the footbridge. He would be struck and killed, but the collision would block the forward momentum of the trolley, saving the five. Should you push the fat man over? Most people (again, about 90%) say they would *not* do so, in a reverse mirror image of the intuitions in Driver (Mikhail 2007).

The trolley "problem," originally raised by Phillipa Foot, is the problem of how to rationally reconcile moral intuitions in Driver with those in cases like Fat Man, since most people are willing to kill one to spare five in the former but not in the latter case (Foot 1967: 3). Foot suggests that "negative" duties such as to avoid injuring or killing others are morally more important than "positive" duties such as to render aid to them (Foot 1967: 4-7). In Driver, Foot says, you are faced with a conflict between negative duties not to kill five and not to kill one, and since you must therefore violate a negative duty not to kill someone no matter what you do, it is better to turn the trolley and kill fewer people (Foot 1967: 5). By contrast, in cases like Fat Man, you are faced with a conflict between a negative duty not to kill one (the fat man) and a *positive* duty to protect the five from harm. In such cases, the negative duty is more important than the positive one, Foot claims (Foot 1967: 5). One therefore should kill the one to spare the five in Driver but avoid doing so in Fat Man.

The priority of right solves the original trolley problem

Foot's analysis is roughly correct but incomplete. To complete the analysis Foot needs to provide some account of why and in what sense "negative" duties to avoid acts such as killing others should take normative priority over "positive" duties to perform acts such as protecting others from harm (see Thomson 2008: 372). I argue that duties not to kill in the trolley problem take normative priority not because they are negative duties but because they are *duties of right*, whereas conflicting positive duties to protect others from harm in cases like Fat Man are *ethical* duties. Duties of right determined authoritatively in public law take normative priority over conflicting ethical reasons for action (see previous chapter). Foot's distinction between negative and positive duties roughly tracks the distinction between legal and ethical duties, since most legal duties are negative and most ethical duties are positive duties. But the relevant distinction is between duties of right and those of ethics.

Perhaps you think the fat man ethically should jump off the bridge himself to save the five, and perhaps you are one of the 10% who think it might not be unethical, therefore, for you to push him because that minimizes lives lost. But the fat man's right to life in such a case has already been authoritatively determined in the system of public laws, and you have a moral duty to respect that determination rather than substituting your own ethical judgment for it, even if you disagree. To do otherwise is to act lawlessly, to commit wrong "in the highest degree" (DR: 6:308n). This is the priority of right. The fat man's right to life includes at least a right not to be coerced to die in order to aid others. This much of the right to life likely must be present in any legitimate system of equal freedom under public laws to which everyone could possibly consent (see PP: 8:349-50). No one can consent to be coercively killed, even to save others: if one could, then the killing would not be coercive; otherwise, there would be no consent. To push the fat man off the bridge in this case fulfills the usual legal elements of murder: 1) killing, 2) a person, with 3) "malice aforethought" (i.e., at least reckless awareness), and 4) without justification or excuse, since a defense of legal necessity generally cannot be raised to a murder charge at common law. Hence the fat man's right to life in such a case has already been authoritatively determined in public law, and you therefore have a normative duty to respect it, whatever your ethical opinion in the case may be.

While it is characteristic of Kantian deontology that duties constrain the goals one may permissibly pursue, the priority of right does not consist primarily in the deontological priority of such ethical constraints over conflicting ethical goals one might have. Both "negative" and "positive" ethical duties might constrain the pursuit of goals such as utility maximization in Kant's deontology. Kant does not explicitly distinguish negative from positive duties anyway; instead, he distinguishes perfect or *strict* duties that always apply in all circumstances, from imperfect or *wide* duties that apply only sometimes or in certain circumstances (GM: 4:422-23; DV: 6:390). The former are usually negative duties, while the latter are usually positive duties. But it seems clear that wide duties to achieve ethical goals might sometimes ripen into ethical obligations that take priority over ethical reasons for action generated by strict ethical duties. For example, an ethical obligation to save a drowning child in a case of easy rescue should take priority over a conflicting strict ethical duty not to break a promise one has made to meet someone for lunch. Hence the priority of right has little to do with the difference between perfect and imperfect ethical duties, despite that all duties of right are perfect. Foot's claim that negative duties take priority over positive duties is therefore roughly correct, but may not always be the case, and the analysis is incomplete. What is missing is an understanding of how and why strict duties of right, which are usually negative duties, take normative priority over ethical duties, which are usually positive duties.

Distinguishing right from ethics and observing the priority of right thus solves Foot's original trolley "problem." One has a duty of right determined authoritatively in public law not to kill the fat man that therefore takes *normative priority* over one's ethical duty to save the five from harm. Whereas in Driver, there appears to be a conflict between a strict duty of right not to kill the one and strict duties of right not to kill each of the five.

2. A trolley *non*-problem: Bystander

Before moving on to Driver, it is necessary to consider what I argue is a trolley *non*-problem that has generated a lot of confusion, a variation referred to as "Bystander." This case is precisely the same as Driver, except that instead of being the driver of the trolley, you are a bystander standing next to a switch that you could flip (or not) in order to turn the trolley to the side track, so killing one and sparing five. Some claim that making this change transforms the moral decision one must make from that between choosing whether to kill one or kill five, as in Driver, into that between choosing to kill one or *not to save* five, as in Fat Man. The problem is then supposed to be why most people would still nevertheless choose to flip the switch in Bystander (as they did in Driver), since most people would choose not to kill one to save the five in Fat Man.

Bystander is a non-problem because it is posed ambiguously. It is in fact unclear in Bystander whether you are choosing to kill five (as in Driver) or merely not to save them (as in Fat Man) when you choose not to flip the switch to turn the trolley. Both Foot and Thomson assume that choosing not to turn the trolley in Bystander violates no duties to avoid killing the five, while at the same time assuming that choosing not to turn the trolley in Driver does violate duties to avoid killing the five. While there is a basis for both assumptions, the issue is not a trivial one, particularly in Bystander.

A legal analysis may be instructive here. While taking no action can be tantamount to taking an "action by omission" in cases where one has a prior legal duty to take some action, there is no general prior legal duty to help or protect others from harm in Anglo-American law (see, e.g., MPC 2.01(3)). For example, while I am guilty of murder if I intentionally take no action to feed my own child who as a result dies, I am not guilty of murder if I take no action to feed children for whom I am not responsible but who would have lived had I provided for them. Family law subjects me to a legal duty of care with respect to my own children, but not with respect to those of others. Ethically, I should help them if I can, of course, but legally, I am not required to do so, and only a prior legal duty will suffice as a foundation for an action by omission. A handful of states impose a duty of "easy rescue" when the rescue does not endanger the rescuer, but the duty applies only in carefully limited emergency circumstances such as at the scene of a traffic accident (see, e.g., Minn Sec. 604A.01). (Some European law imposes a more stringent statutory duty to assist, but the duty is still scoped to fall well short of a general duty to assist.) Hence if you choose not to turn the trolley, the legal analysis in Bystander will turn on whether you have some specific prior legal duty to the five to protect them from being harmed by the trolley. Such a duty is easier to make out in Driver than in Bystander, but the issue is the key to the legal analysis in both cases, if one chooses not to turn the trolley and kill the five.

The trolley driver does plausibly have a prior legal duty to drive the trolley safely, which includes at least a duty to prevent causing harm to others in the normal operation of the trolley. This duty becomes clearer if we imagine Driver with *no one* on the side track. If the driver nevertheless chooses not to divert the trolley to the empty side track and so to kill five people, then the driver's inaction would constitute an action by omission of her duty to drive the trolley safely, and a murder charge seems appropriate. Perhaps the bystander would not be subject to a similar prior duty in such a case, as Foot and Thomson appear to assume, and the bystander might simply watch the trolley continue on the main track to kill the five. But if the source of the driver's prior duty to operate the trolley safely stems primarily from the driver's *control* over the trolley, then the bystander should also be subject to such a duty because the bystander (bizarrely, to be sure) in Bystander controls the trolley just as completely as if she were the driver. The bystander might thus be subject to a similar prior legal duty to operate the trolley safely, and might plausibly act by omission if she chooses not to flip the switch and so to kill five people, rather than flipping the switch to divert the trolley to the empty side track. A murder charge would thus again seem appropriate. Yet Foot and Thomson both appear to assume that a bystander in complete control of the trolley has no prior duty to prevent the trolley from killing the five, and that Bystander is therefore just like Fat Man in this respect.

In experiments where a case like Fat Man, rather than Driver, is presented to subjects before Bystander, many fewer would still choose to turn the trolley, and those who would are much less sure about it (Petrinovich and O'Neill, 1996: 156-8). Such framing or ordering effects appear to affect every variation of the trolley problem *except* Driver and Fat Man (Liao, et al., 2007). In Fat Man, unlike Driver, it is clear that you have no prior legal duty to prevent the trolley from harming the five. Even if you could easily push, say, a boulder, rather than the fat man, over the footbridge in order to block the trolley and so save the five, you have no general legal duty to do so (in the absence of an applicable "easy rescue" statute). Whereas in Driver, it is plausible that you have a prior legal duty not to kill the five that applies whether there is anyone on the side track or not.

I speculate, therefore, that the question with which subjects struggle in Bystander is the key legal issue of whether not acting (not turning the trolley) is tantamount to an action by omission in the case. The answer may depend on whether the bystander has a duty to protect the five from being harmed by the trolley by virtue of her complete control over the trolley, or not. When subjects are asked to evaluate Bystander after evaluating Driver, they are more likely to see an analogy and assume that there is such a prior duty, which justifies turning the trolley despite that doing so violates a duty not to kill the one on the side track. Either way, they would violate duties not to kill, and so killing fewer may seem preferable, as Foot suggests. When subjects are asked to evaluate Bystander after a case like Fat Man, on the other hand, they again see an analogy and are thus less sure as to whether a bystander with control has a duty to protect the five from being hit by the trolley, and so the duty not to kill the one on the side track makes it more difficult for them to choose to turn the trolley. While a majority would still choose to turn the trolley in Bystander after evaluating a case like Fat Man, intuitions in Bystander become much less clear.

Bystander is therefore a non-problem because it fails to serve as an example of anything. Moral intuitions shift over answers to the question of whether the bystander in a position of complete control over the trolley is responsible for the deaths of the five, or not, and thus whether Bystander is like Driver or is, instead, like Fat man. If Bystander is like Driver, then not turning the trolley is an action by omission of the duty to safely operate the trolley, and the choice is one between different numbers of intentional killings. If Bystander is like Fat Man, on the other hand, then it is a case where one would violate a strict duty of right not to kill one in order to meet an ethical duty to save five, and duties of right take normative priority. The prior duty in Bystander is thus simply ambiguous. When the prior duty is clarified one way or the other, there is no problem or dilemma.

In her original article on the trolley problem, Thomson argues that the fact that one is a driver should not matter, and that a passenger taking over for an incapacitated driver would be just as responsible for the deaths of the five when choosing not to turn the trolley as the driver would be (Thomson 1976: 207). That is, Thomson argues that one intentionally kills five by not turning the trolley, even if one is merely a passenger or, effectively, a bystander who happens to have complete control over the trolley. Hence Bystander is like Driver with respect to one's prior duties toward the five. In her most recent article on the trolley problem, however, Thomson reverses position on this issue and concludes that Bystander is not like a Driver case, and that one should not turn the trolley in Bystander, after all, even if doing so seems intuitively ethically correct (to her) (Thomson 2008: 372-4). Hence Bystander is like Fat Man, Thomson concludes, where you would kill one to save five to whom you owe no prior duties. Thomson's shifts of position in Bystander demonstrate the *non*-problem that the case poses as a thought experiment.

CHAPTER FOUR

SOLVING THE TROLLEY PROBLEM II: DRIVER

1. Conflicts between strict legal duties

As I argued in the previous chapter, there seems good reason to think that the driver has a prior legal duty to safely operate the trolley, and so that not turning the trolley when doing so would prevent the trolley from killing five people is an action by omission tantamount to intentionally killing them. If the driver were driving a car on a highway, rather than a trolley on fixed tracks, and had to choose whether to run over and kill five people in the road ahead or swerve onto a side road and kill one, then the driver's responsibility for choosing to kill the five seems clear (Thomson, 2008: 369). In such a case, the driver either in fact takes positive action to maintain the car's course and so kill the five, or the driver omits to perform her prior duty to safely drive the car and avoid collisions by allowing it to run over five people when she could have turned to avoid them. I will therefore assume, as Foot and Thomson do, that the conflict in Driver is indeed one between strict duties not to kill each of the five and not to kill the one.

Foot takes it for granted that it is better to violate only one rather than five negative duties not to kill and that this is why you should turn the trolley in Driver (Foot 1967: 5). But since principles of justice bar the violation of one person's rights to achieve a greater good such as to save many people, it is not entirely clear why justice should permit the violation of one person's rights to achieve the greater good of avoiding violating five people's rights. The one whose rights are violated might complain of being wronged in either case. Moreover, not all conflicts between strict duties of right can be resolved by appeal to a general moral principle that one should minimize harm. There could be dilemmas where you are forced to choose between violating the same number of persons' rights on each horn of the dilemma, or where you are forced to choose between violations of different kinds of negative duties with no priority ordering. For example, you might be forced to choose whether to kill one or another of your own children, lest they both die (e.g. as in William Styron's *Sophie's Choice,* 1980), or forced to choose which of many creditors to repay out of limited funds, or forced to choose whether to lie about a serious matter in order to avoid breaking an important promise, and so on. Such cases seem easy to multiply.

Yet Kant appears to deny that conflicts between strict obligations can even exist:

> But since duty and obligation are concepts that express the objective practical necessity of certain actions, and two rules opposed to each other cannot both be necessary at the same time—rather if it is one’s duty to act according to one of them, to act according to the opposite one is not only no duty, but even contrary to duty—a collision of duties and obligations is not even conceivable (*obligationes non colliduntur*). (DV: 7:224)

Kant argues here that if one were required to perform an action (a) in accordance with an obligation (Oa) that opposed another simultaneous obligation (O~a), then acting in accordance with the first obligation (a) would imply acting in a way that violated the second obligation (~a), a performance that is not even conceivable (a ^ ~a). One cannot be obligated to perform what is impossible (O(a ^ ~a)); therefore, Kant concludes, one cannot simultaneously be subject to opposing obligations (Oa ^ O~a). (Here "O" is a monadic operator for an obligation one has; "a" is an action one performs.)

Kant's claim that legal obligations cannot come into conflict (~(Oa ^ O~a)) may be understood either descriptively or normatively. Descriptively, the claim seems false, even in ideal theory. There seems to me no reason to think that even a thoroughly rational public authority might not inadvertently create legal obligations that contradict in situations that authority did not foresee. For example, suppose a state authority passes a traffic law that requires stopping at stop signs and also another that forbids stopping in front of military bases (see Rodriguez and Navarro 2017: 179). It is not inconceivable that a local government agency might then erect a stop sign in front of a military base, creating a conflict of legal obligations under applicable enforceable laws for drivers unfortunate enough to encounter the situation. The possibility of such conflicts seems a mundane descriptive fact about any system of laws, and while one might be tempted to assert that the ordinances in question cannot be held to conflict in the case because the driver can have only one true legal obligation, this assertion seems clearly normative rather than descriptive.

I argue that the best way to render Kant's claim about the systematic consistency of one's strict juridical duties, then, is to think of it as a *normative* requirement of justice, rather than a necessary descriptive truth about any system of norms we might call legal. Whether conflicts of legal obligation are descriptively possible or not, it would be wrongful to enforce contradictory legal obligations, as then force would be applied arbitrarily, and one cannot possibly consent to be subject to arbitrary coercive force. But since ethical obligations that are not also legal obligations are not rightfully enforceable, this normative requirement does not apply when conflicts between ethical obligations occur. Hence Kant's analysis of conflicts between legal as opposed to ethical duties is quite different. The former but not the latter normatively must be resolved in a system of equal freedom under universal law.

The normative demand for consistency in the system of public laws and the priority of right hold independently of any thesis regarding moral pluralism. It could be that ethical reasons for action are intrinsically inconsistent and that "tragic" conflicts are therefore unavoidable. That is, it could be that conflicts arise not because of epistemic or other limitations on individual ethical judgments, but because ethical duties themselves are irreconcilable. The priority of right is not required in order to solve the problem of moral conflicts per se, however. The problem that public law solves is that unilaterally enforcing one or another resolution of such conflicts on one's own wrongs others; only the united "omnilateral" will embodied in a constitutional public authority can legitimately define, execute and judge rights in order to resolve moral conflicts. Hence justice requires that duties and rights in Kant's system of equal freedom under universal law be made consistent, whether moral pluralism is true or not. Moral pluralism implies that ethical duties might come into tragic, intractable conflicts, creating deontic dilemmas that cannot be resolved by appeal to any further rational ethical principles. But conflicts between the strict *legal* obligations public authority defines and enforces are nevertheless normatively intolerable in the system of equal rights of freedom under universal law.

2. The priority of right in Driver

I can now offer an approach to the solution of the trolley problem dilemma in Driver. First, I observe that the conflicting obligations at issue are strict legal obligations not to wrong another by intentionally killing her, even to save many others. I further stipulate that the problem is indeed a dilemma in which one is subject to contradictory strict legal obligations ($Oa \wedge O{\sim}a$). That is, there is no other legally relevant factor, such as the act-omission distinction, or a superior right on one side or the other due to fault, that would eliminate or prioritize one of the obligations.

I then appeal to Kant's normative requirement that strict legal obligations must be made consistent in the prescriptive system of public laws. What does this normative requirement imply in such a case? The first implication is that *neither legal obligation in the dilemma can be rightfully enforced.* It is not possible to consent to be subject to the enforcement of contradictory strict legal obligations, as this is tantamount to consenting to arbitrary acts of coercion. But this requirement of consistency in the system of legal duties is a second-order principle of justice, not a property of the system. Enforcement of either obligation if taken by itself is both rightful and wrongful in principle at the level of the prescriptive system of legal duties. At this prescriptive level, consistency is a constraining property of the system; hence a lack of consistency with other legal duties in the system cannot be the reason that a duty is not rightfully enforceable. Contradictory duties are simply inadmissible into the prescriptive system of legal duties, and the implication of a dilemma is, rather, that the enforcement of either obligation is both rightful and wrongful, i.e., that its rightfulness *cannot be determined*.

A second implication is that justice requires that *the dilemma must be resolved by law* (i.e., either by legislative action or judicial or executive order). It does not matter how it is resolved, so long as the procedural and substantive requirements of justice are met when resolving it. What matters is that the conflict is resolved; and moreover, its resolution may vary by jurisdiction, so long as there is due process. Legitimate variation in the law by jurisdiction is in fact a common feature of most legal systems: in some U.S. states, for example, contributory negligence completely bars recovery by injured plaintiffs, while in other states, fault might play no or a very limited role. Yet in each state, the law that resolves the conflict is rightfully enforceable.

Suppose, for example, that five people are attempting to cross an interstate highway (which is generally illegal), and a self-driving car cannot brake in time to avoid hitting and killing them. Suppose the car could swerve to avoid them, but doing so would kill a motorcyclist riding in an adjacent lane. The car thus must choose between killing the five on the highway or swerving and killing the one motorcyclist. In a strict liability jurisdiction, liability for the deaths is assigned strictly without regard to fault, and hence makers will program the car to swerve and kill the motorcyclist, because in such a jurisdiction, one wrongful death is less costly to compensate than five. In a contributory negligence jurisdiction, on the other hand, the car will be programmed to continue ahead and kill the five, because in such a jurisdiction, fault bars recovery, and the maker thus would not be liable for the deaths of the five. In each case car makers will program self-driving cars to minimize their legal liability (Casey 2017). Yet both states' rules are rightfully enforceable within their respective jurisdictions because neither violates constitutional standards of justice. Note that principles of *ethics* are likely to play no role in determining the behavior of self-driving cars in such cases.

Now, if the Driver variation of the trolley problem is framed as one where we are forced to choose between *ethical* duties to avoid harming one as opposed to five, then I would agree with Foot (and popular opinion) that, ethically, one should turn the trolley. But how I might frame the issue ethically is irrelevant to one's strict legal obligations in such cases, and even ethicists committed strictly to Kantian deontology, for example, may disagree on its proper ethical resolution. By contrast, whatever resolution a public authority makes in Driver is rightfully enforceable and so authoritatively decides the issue.

3. Self-defense, legal necessity, and dilemma

Kant connects justice with an authorization to use coercion:

> [C]oercion is a hindrance to resistance to freedom. Therefore, if a certain use of freedom is itself a hindrance to freedom in accordance with universal laws (i.e.

> wrong), coercion that is opposed to this (as hindering a hindrance to freedom) is consistent with freedom in accordance with universal laws, that is, it is right. Hence there is connected with right by the principle of contradiction an authorization to coerce (DR: 6:231).

Self-defense is not wrongful, Kant says, because one's act of self-defense "hinders a hindrance" to one's right of freedom and is therefore consistent with equal rights of freedom in accordance with universal laws (DR: 6:235). That is, when killing one's assailant in self-defense, there is no violation of the duty of right not to kill because the killing hinders the wrong the assailant is attempting to commit. Kant does suggest that one might nevertheless have an *ethical* reason not to kill one's assailant (DR: 6:235). This is possible because unlike legal duties, ethical duties are not specified by reference to their consistency within a system of equal freedom under universal law. While Kant restricts the term "obligation" in such a way as to preclude conflicts even of ethical obligation, he allows that one may have conflicting ethical "grounds" or reasons for action (Timmerman 2013). Such conflicting reasons do not exist in a legal context, however, since legal obligations are completely specifiable in terms of their outward aspects (DR: 6:231). Legal obligations are indifferent to one's "grounds" for performing them.

In a case of the defense of legal "necessity," by contrast, in which one wrongs an innocent because that is the only way to save oneself, one does act wrongfully, Kant argues. While enforcement of the legal obligation not to kill in such a case would therefore be rightful in principle (because it would correct the wrong), enforcement in necessity cases is not practically possible, according to Kant, since even a punishment of death would not effectively deter the crime (DR: 6:235-6). Kant thus regards the defense of necessity, to the extent it is thought a legal defense, as premised on a confusion. Kant would likely reject any version of the general legal "necessity" or "choice of evils" affirmative defense that is sometimes raised in U.S. law (see, e.g., MPC Sec. 3.02).

The case of a dilemma is distinct from either self-defense or Kant's version of the defense of necessity, however. Consider the Driver variation of the trolley problem again, stipulating again that it is a genuine dilemma of contradictory strict legal obligations. You must choose whether to turn the trolley and kill one or stay on course and kill five. Perhaps you might argue when turning the trolley that there is no violation of the legal obligation not to kill the one because by doing so you are hindering the wrong of killing the five. But then you might equally argue that by staying on course you are not wronging the five because by killing them you are hindering the wrong of killing the one. Does this imply that one acts *rightfully* no matter what one chooses in dilemmas because one's act hinders a wrong, as in the case of self-defense? No. The problem, of course, is that this reasoning can go on ad infinitum, since while killing the one hinders the hindrance of killing the five (and so is right), killing the five also hinders the hindering of the hindrance of killing the one (and so is wrong), and so on. Your action can always be proven to be rightful or wrongful by taking one additional step in the infinite regress. Does this regress imply that in a dilemma your act is *wrongful* but that punishment would not be practical, as in Kant's case of necessity? No. Either obligation practically could be enforced by imposing a suitable punishment, which would effectively deter its violation. The problem is, instead, that any such enforcement would fail to hinder the wrong without itself arguably creating another wrong one step further into the regress.

But perhaps you might argue that since the regress in a dilemma prevents *proving* that an act of corrective enforcement would be rightful, then your act is not wrongful no matter what you choose to do, since legal obligations must be in principle rightfully enforceable. Your act is thus rightful by default. But this does not resolve the issue, either, since one might equally argue that the regress prevents proving that enforcement is *wrongful* in dilemma cases (that is, prevents proving that a lack of enforcement is rightful), as well, and so legal obligations in a dilemma are rightfully enforceable. Your act is thus wrongful by default. One cannot assume either that enforcement is wrongful or rightful by default in dilemmas for the same reason that one cannot infer that one or the other obligation is rightful or wrongful because the other is not: the obligations contradict each other. The contradiction destroys both the classical inference that your act is rightful because it hinders a wrongful act, as well as the defeasible inference that acts are rightful because a countervailing corrective act of enforcement cannot be proven rightful. Obligations in the dilemma are indeterminate, not merely unprovable.

Perhaps there is some way to simply suspend judgment in dilemma cases; for example, perhaps courts could simply ignore or refuse to rule on them. But it seems doubtful that courts could really *suspend* enforcement by refusing to hear dilemma cases. It seems that what would be enforced, instead, is a new legal rule that says that whatever choice the agent makes on her own is legally permissible. Now, allowing discretion on the part of the agent in choosing what to do when facing dilemmas of legal obligation might be just if such a rule of discretion were made *explicit* in legislation or by judicial or executive order of a legitimate public authority. Then the rule of discretion in dilemmas would be one to which everyone could possibly consent, and so when agents caught in dilemmas exercised that discretion, rights would not be violated, since they would have been re-specified by reference to the system of laws that now contains the rule of discretion. You have no right to unilaterally decide whether to kill one (or five) to avoid killing five (or one), in the absence of an explicit rule granting you limited discretion to do so. Your decision would be lawless, and those wronged might rightfully resist you by force and hold you responsible.

Obligations in a genuine dilemma case are thus undefined and indeterminate. Perhaps an analogy in arithmetic might be helpful. In arithmetic, any number multiplied by zero is zero, but since division is defined in terms of multiplication (i.e., 'a/b = c iff a = b * c'), then division by zero (a/0 = c) implies that there is a number (c) that when multiplied by zero (0*c) produces some number (a). If the product (a) is not zero, then there is no such number (c), since any number multiplied by zero is zero. But if the product (a) is zero, then any arbitrary number (c) will do, since any number multiplied by zero is zero. Hence division by zero is *undefined* for any dividend except zero (0/0), in which case division by zero is also *indeterminate*, since any number will do. Here, rendering justice by hindering a hindrance to justice makes no sense when duties are in contradiction with each other. There is no such hindering that does not result in a further unjust hindrance; hence justice as hindering a hindrance is undefined in the dilemma case. But at the same time, any hindering of a hindrance will achieve justice in the dilemma case simply because it ends the regress and closes the dilemma. It does not matter which horn of the dilemma one hinders; hence justice as hindering a hindrance in a dilemma case is also *indeterminate*.

Obligations in a dilemma are indeterminate, and corrective enforcement is thus both wrongful and rightful at the level of the prescriptive system of public laws. But justice requires that the system must be one to which everyone can consent, and since no one can consent to be subject to arbitrary coercive force, enforcing either obligation in a dilemma case is nevertheless wrongful. Since enforcement of some kind is unavoidable, however, and deontic dilemmas are not impossible, the dilemma must be positively resolved by a public authority in the form of legislation or executive or judicial order. Then there will be no dilemma, and obligations can be rightfully enforced.

From the point of view of justice, dilemmas are thus little different from other conflicts of strict legal obligation. The main difference appears to be that in the dilemma case we may assume that there is no clear rational resolution of the conflict at issue, whereas in ordinary cases of conflict, we may assume that some rational resolution of the conflict exists. Regardless, public law must resolve the dilemma, just as it does in other cases of conflict. I do not mean to imply that civil institutions are authorized to resolve such conflicts irrationally or arbitrarily; rationality will still impose some bounds upon acceptable resolutions and their rationales. It is just that in the dilemma case there is no decisive reason to resolve the conflict one way or the other.

CHAPTER FIVE

NORMATIVE CONSISTENCY AND DEONTIC LOGIC

Given the distinction between right and ethics and the normative priority of right, the problem of how to build an explicitly moral machine agent largely reduces to the problem of how to build an agent that obeys the public law of a legitimate state. Since law consists in a system of authoritative legal norms and rules, the focus shifts to creating a deontic logic of the law that best serves the normative requirements of justice. Systems governing rightful machines need not explicitly implement such a logic in order to conform with duties of right, but it is difficult to see how a normative governance system could function without making at least some external reference to a legal knowledge base organized in accordance with an appropriate deontic legal logic. Rightful machines' behavior should degrade gracefully, however, when encountering unresolved conflicts in the law, and this is where an independent, explicitly ethical system of governance might play a role.

In this chapter, I evaluate a number of deontic logical approaches to conflicts of legal obligation against the normative requirements of justice. How should a deontic logic of the law handle conflicts of strict legal obligation? I argue that the role of a deontic logic of the law is not to work around such conflicts but to identify and expose them so that civil institutions can authoritatively qualify the rights or duties generating inconsistencies in the system.

1. The inadequacy of the standard system of deontic logic (SDL) and variations

The standard system of deontic logic is a normal modal logic with a deontic gloss on the $\Box$ (box) and $\Diamond$ (diamond) operators, interpreted as obligation and permission, respectively. The system is a K logic characterized, syntactically, by the D (deontic) axiom, '$\Box p \rightarrow \Diamond p$' (that is, if action p is obligatory, then p is permitted) or the 'D$\Diamond$Introduction' rule in a Fitch-style proof system, and, semantically, by a seriality condition on frames in the Kripkean possible world semantics (that is, for every world, there is at least one accessible world). What SDL amounts to is the rejection of conflicts of obligation ($\sim(\Box p \ \hat{}\ \Box{\sim}p)$), which is just the D axiom.

But as I argued in the previous chapter, there is no reason to think that deontic conflicts cannot occur, and indeed, some reason to think they are common. An adequate deontic logic should not simply deny the possibility of such conflicts, as SDL does. Yet if one simply rejects axiom D so as to admit conflicts of obligation into SDL, then the logic becomes immediately incoherent, since given uncontroversial principles for the inheritance of obligations (RM) (If $\vdash p \rightarrow q$, then $\vdash Op \rightarrow Oq$), and aggregation (AND) ($\vdash (Op \wedge Oq) \rightarrow O(p\hat{}q)$'), one can derive any obligation from the contradiction in accordance with the classical logical principle *ex falso quodlibet* (EFQ) ('$(p \wedge {\sim}p) \rightarrow q$') (see Girle 2017: 195-6). That is, given a dilemma where simultaneously Op and $O{\sim}p$, any arbitrary action q can be proven to be obligatory: Oq: 1. Op. assp. 2. O~p. assp. 3. O(p ∧ ~p). 1,2, AND. 4. Oq. 3, EFQ, RM. A number of efforts to weaken one or more of these principles to avoid this deontic explosion of obligations have therefore been undertaken, though with limited success.

Semi-classical and paraconsistent logics reject EFQ, replacing the two truth values (true, false) of classical semantics with a semantics of many values (e.g., null, just true, just false, and both true and false) (see Girle 2017: 99-105). Such logics have generally proven too weak to be very useful, however, because they fail to vindicate certain common intuitively valid deontic arguments. For example: 1. S ought to fight in the war or perform alternative service to his country (O(f ∨ a)). 2. S ought not fight (O~f). Therefore, Smith ought to perform alternative service to his country (Oa) (see Goble 2005: 467). This conclusion cannot be derived in most paraconsistent or relevance deontic logic systems, as they lack the disjunctive syllogism of propositional calculus needed to make the inference ((f ∨ a) ∧ ~f) → a. Such failures are not conclusive, however, and overcoming them continues to be an area of active research.

Other efforts attempt to avoid deontic explosion by weakening Aggregation (AND) or Inheritance of Obligation (RM), rather than rejecting EFQ. They typically do so by imposing consistency or permissibility checks of various sorts on their application. For example, Aggregation (AND) may be weakened by requiring that p and q be *jointly possible* before allowing their aggregation under obligation (CAND: If |/-+ p → ~q then |- (Op ∧ Oq) → O(p ∧ q)), or by requiring that p and q be *jointly permissible* (PAND: |-

$P (p \wedge q) \rightarrow ((Op \wedge Oq) \rightarrow O(p \wedge q))$. Inheritance of Obligations (RM) may be weakened by requiring that p be permissible before allowing q to inherit an obligation from the obligation that p (RPM: If $\vdash p \rightarrow q$ then $\vdash Pp \rightarrow (Op \rightarrow Oq)$) (see Goble 2005: 467-473). Each resulting logic avoids the original form of deontic explosion and has its advantages and disadvantages in accounting for the more or less intuitive validity of various example deontic arguments.

The problem with these attempts in the present context is that they offend the demand for consistency *normatively*. Contradictory obligations are admitted as first-class citizens of such logics. In paraconsistent logics, inferences are derived in the face of contradictions by the alchemy of a non-classical semantics, which often confounds intuitions. In the weakened deontic logics described above, by contrast, contradictions are like icebergs around which reasoning proceeds gingerly, if at all. In neither case does the logic require that one contradictory obligation be defeated, or a rule generating the contradiction be qualified or revised, in order to allow an inference through the other obligation, or vice versa. For example, suppose a criminal statute punishes those who intentionally kill a person ($k \rightarrow Op$), while another statute forbids punishing minors ($m \rightarrow O{\sim}p$), and suppose a court confronts a case where a minor has intentionally killed someone ($k \wedge m$). This licenses the inference Op and also $O{\sim}p$, so creating a conflict of obligations. The weakened logics above draw both inferences but then limit any further inferences that depend directly on one or another of them. For example, suppose that punishment always consists in incarceration ($p \rightarrow c$). RM would license $Op \rightarrow Oc$, and therefore the inference that the killer must be incarcerated, despite that she is a minor (Oc) and ought not to be punished (O~p). The weakened RPM logic appropriately blocks this inference because Op is impermissible, O~p (== ~Pp). The RPM logic infers that there is a killer who is a minor (k, m), and that one is obligated to punish her (Op) and obligated not to punish her (O~p) but then blocks the explosion of further inferences such as Oc. While the RPM logic therefore succeeds in admitting conflicts while avoiding deontic explosion, which is its goal, the approach to doing so seems to me to miss the point of admitting deontic conflicts in the first place.

Conflicts of deontic obligation should stimulate inference rather than shut it down. What conflicts indicate in the deontic context is that one must either revise one or the other of the inconsistent formulas, or prioritize one over the other, or semantically, that one must choose between competing consistent models of (revised) rules, given the facts in some conflict situation. While a doxastic or epistemic application of modal logic may perhaps not be subject to the same normative demands, a deontic logic of the law should provide some mechanism to make such inferential choices. The goal in the case of the killer who is a minor above is to render a judgment as to whether her punishment is consistent with everyone's obligations and rights in the system of public laws under universal law. But paraconsistent logics and weakened deontic logics that admit contradictions seem useless for this purpose. A court might resolve the case by, for example, qualifying the rule against homicide so as not to apply to minors ($(k \wedge \sim m) \rightarrow p$), or on the other hand, by qualifying the rule barring the punishment of minors so as not to apply in cases of intentional homicide ($(m \wedge \sim k) \rightarrow \sim p$), or the court might articulate some rule of priority (see Alchourron, 1991: 423-424). The deontic logic should be able to admit the conflict descriptively and provisionally generate inferential alternatives, together with further consequences, in order to evaluate each resulting consistent set of rules, and require a choice. The weakened deontic logics, instead, simply admit the conflict and limit further inferences. But what the normative demand for consistency requires is a deontic logical system that concedes the presence of contradictions descriptively but whose semantics ultimately insists that they be authoritatively resolved in the prescriptive system of public laws (see Alchourron 1991).

2. Non-monotonic deontic logic

I suggest that non-monotonic reasoning systems (NMR) with a classical (rather than paraconsistent) base can meet this normative demand for consistency, with appropriate reservations. NMRs are able to admit contradictions without igniting a deontic explosion of obligations because they reject monotonicity, that is, “if K|- p and $K \subseteq K'$, then K' |- p”. What the rejection of monotonicity means is that some inferences might no longer be made when new premises are introduced; for example, one might introduce a new fact that directly contradicts some fact upon which an inference depends, so defeating that inference. NMRs therefore avoid the deontic explosion of obligations that plagues SDL.

The consequences a non-monotonic logic licenses one to accept given some set of accepted statements K can be defined in terms of its extensions, which, informally, are the rational and stable sets of conclusions that one may accept, given K. Extensions are *rational* in the sense that defeasible conclusions are not accepted if they would create inconsistencies, and *stable* in the sense that 1) all the conclusions one accepts have some justification in K, while 2) adding any further conclusion would create an inconsistency. The consequences one is licensed to accept given K can then be defined as the intersection of extensions, or as the intersection of some set of preferred extensions.

Classical logic can be defined as a structure S=(F, R) where F is a set of formulas, and R is a set of rules of inference. R defines a classical consequence relation (|-) between a set of formulas and a formula of the language (p). A non-monotonic default logic can be defined as a structure S = {F, K, R} where F is a set of formulas, K is a set of default rules, and R is a set of rules of inference that define a non-monotonic consequence relation (|~) as follows (see Maranhao 2006: 66-67):

> F_K |~ p if and only if F, K’ |− p for all subsets K’ ⊆ K which are maximally consistent with F.

For example, suppose K = {b→f, p→¬f} ("Birds fly; penguins do not fly.") Hence {b} |$\sim_K$ f because for all subsets K' of K that are maximally consistent with {b}, K' |- f. (e.g., {b, b→f, p→¬f} |- f.). (Hence "birds fly.") But if we add p to F, then {b, p} |/$\sim_K$ f. because one of the maximally consistent subsets K' of K {b→f, p→¬f} is not consistent with {b, p}, that is {b→f, p→¬f, b, p, f, -f}. (Hence "birds that are penguins do not fly.") This demonstrates that adding p to the premises causes the conclusion b to be withdrawn, or non-monotonicity.

I will briefly sketch out the answer set semantics for the programming language I will use in later chapters to model legal conflicts, Answer Set Prolog (see Gelfond 2008). (See Gelfond 2008 for a fuller treatment of what follows.) A logic program (Π) consists of a set of rules of this form:

$$a \text{ :- } b_k, b_{K+1}, ..., b_m, \text{not } c_{m+1},... , \text{not } c_n.$$

where a, b, and c are formulas, and "not" is "negation-by-failure." "a" above is called the *head* of the rule, while formulas following the ":-" symbol are called the *body* of the rule. A rule with no head is a *constraint*, while a rule with no body is a *fact*.

A *partial interpretation* of Π consists of a consistent set of *ground* literals satisfying the rules of Π. *Ground* rules, literals, and terms of a program Π contain no variables; hence to create a ground instance of a rule of Π, replace all the rule's variables with ground terms of Π. A partial interpretation S of Π is an *answer set* for Π if S is *minimal* among the partial interpretations satisfying the rules of Π, where minimality is understood in terms of set inclusion.

Answer sets can be obtained in the following way. The *reduct* of program Π relative to a partial interpretation S of Π is the program obtained by first deleting every rule from Π with "not p" in its body, where p is a member of S, and then deleting all "not q" from the remaining rules of Π, where q is any literal. A partial interpretation S of Π is an answer set if S is an answer set for the reduct of Π.

For example, consider this program Π:

```
p :- not q.

q :- not p.
```

First, generate a partial interpretation (a candidate answer set); we can start with the empty set {}. Then get the reduct for {}, obtained here by eliminating "not" parts of the rules:

```
p.

q.
```

The consequences of this reduct are {p,q}, which is not the same as {}. Hence {} is not an answer set.

Next, generate a new partial interpretation {p} and get the program reduct by eliminating 1) the entire first rule because it contains "not p" and then 2) the "not" part of the second rule to obtain:

```
p.
```

The consequences of this reduct are {p}, which is the same as the original set generated. Hence {p} is an answer set.

Next generate {q} and get its reduct:

```
q.
```

The consequences of the reduct are {q}, which again are the same as the originally generated set. Hence {q} is an answer set.

Next generate {p,q} and get its reduct, which is null, which has no consequences; therefore {p,q} is not an answer set.

Hence the example program has two answer sets, {p} and {q}.

I will demonstrate an "answer set programming" approach to evaluating conflicts of legal obligations further in subsequent chapters when I step through detailed examples of conflicts between rules. The approach allows conflicts to be represented, while at the same time 1) providing a mechanism for their resolution, while 2) rejecting direct (classical) contradictions of fact.

3. Logics of belief revision

Carlos Alchourron rejects non-monotonic deontic logics because he argues such systems obscure the distinction between descriptive and prescriptive activity in the law (Maranhao 2006). As a positivist Alchourron looks outside any formal property of law for sources of law's normative authority. Kant understood there to be a necessary connection between law and the normative obligation to obey it; hence Kant rejects positivism. Law that conforms to the Universal Principle of Right (UPR) is normatively obligatory for Kant because of its formal structure (universality, consistency, etc.) and to some degree also its substantive content (not violating rights of equality, freedom, etc.).

All of these requirements flow from Kant's principle of justice, the UPR, and ultimately, the supreme principle of morality, the categorical imperative.

Yet Kant would also recognize that a number of diverse consistent bodies of positive law are possible and each legitimate because they do not violate basic constitutional conditions. Hence Kant may also have some reason to prefer a legal epistemology that shows the explicit evolution of a body of law toward the strongest and most coherent system realizing equal freedom. Logics of belief revision such as Alchourron's AGM may thus offer the most promising approach to realizing Kant's normative requirement of consistency, since such logics have robust formalisms for various operations such as expansion, contraction or revision of the normative system, and all refinements to legal rules are made as explicit as possible (Alchourron, Gardenfors, Makinson 1985). Rules are not represented as defeasible defaults in such systems, although they may still achieve appropriately defeasible inferences by Alchourron's use of a revision operator on the antecedents of conditional obligations (Alchourron 1991). The ultimate goal of systems like AGM is to completely and consistently and *explicitly* represent the full specification of all legal rules. Defeasible logics, on the other hand, may never eliminate defeasible rules that appear to be in conflict but do not generate contradictions because of a preference ordering found elsewhere in the logic. While formally such logics are equivalent to AGM when supplemented by Alchourron's "f" revision operator (Aqvist 2008), a logic such as AGM may better reflect Kant's normatively consistent system of equal freedom under universal laws constructed by a civil community.

A deontic logic suitable for capturing legal rules must of course include many additional elements that I have not discussed at all here. I have said nothing about legal powers, immunities, disabilities, etc., or agency, the relational aspect of legal obligations, or how quantification and deontic modality might work. My focus is strictly limited to the problem of conflicts, which I examine in the simplest (propositional, monadic deontic operator) case.

CHAPTER SIX

ANSWER SET PROGRAMMING LEGAL RULES AND CONFLICTS

In this chapter, I first work through a standard example to demonstrate non-monotonic, defeasible reasoning and explicit rule qualification in answer set programming. I then set out how to properly encode and evaluate conflicts of legal duties using answer set programming.

1. Non-monotonic reasoning in answer set programming

Defeasible inferences

I set it out using the lparse grounder and clingo parser to efficiently ground and solve logic programs (see Pottasco). The head of each rule is to the left of the ":-" symbol, while the rule's body is on the right. Rules can be understood as conditionals where the head is the consequent, and the body is the antecedent. The body may have multiple terms separated by commas "," each of which must be satisfied in order to satisfy the body (so the comma may be translated as "and.") Capitalized words are variables, lower-case words are terms (literals or predicates). "%" indicates a comment. "-" (minus sign) indicates classical negation.

```
% birds fly.  That is, X flies if X is a bird.
flies(X) :- bird(X).

% penguins are birds
bird(X) :- penguin(X).

% penguins do not fly
-flies(X) :- penguin(X).

bird(tweety).       % tweety is a bird
penguin(chilly).    % chilly is a penguin
```

The program states three strict rules, that all birds fly, that penguins are birds, and that penguins do not fly. It then declares two facts, that "tweety" is a bird, and "chilly" is a penguin.

Running the program in lparse-clingo generates the result:

```
UNSATISFIABLE
```

The program strictly implies a contradiction, that chilly both flies and does not fly. The program is therefore unsatisfiable classically. We can infer nothing (or everything) from it, even concerning tweety. But this does not seem to be the result we want. Instead, we prefer to qualify the rule regarding birds in chilly's case, since chilly is a penguin, to that we may infer that chilly does not fly, despite being a bird. We also prefer to maintain the inference that tweety flies.

Non-monotonic reasoning formalisms such as default logic achieve this by making some rules defeasible rather than strict, such as, here, the rule that birds fly. They do so by using the negation-by-failure operator ("not"), which can be read as "it is not provable that." Classical negation is then distinguished syntactically with "-". The following default rule states that if X is a bird and it is not provable that X does not fly, then we can infer that X flies.

```
% Defeasible normal default rule d1: birds fly
flies(X) :- bird(X), not -flies(X). % except when they are birds that do not fly
```

Adding this rule to the program generates the following answer set:

```
Answer: 1

bird(tweety) bird(chilly) -flies(chilly) flies(tweety) penguin(chilly)

SATISFIABLE
```

Now the program has an answer set where the default rule that birds fly is defeated in chilly's case. Chilly is a penguin and no penguins fly; hence it is provable that chilly does not fly. Tweety flies, on the other hand, because she is a bird and, in the absence of any further positive information concerning tweety, it is not provable that she does not fly. (Note that this is not a closed-world assumption. We assume neither that tweety flies nor that she does not fly in the absence of proof. Rule d1 that birds fly and the fact that tweety is a bird provides the proof we need to infer that she flies.)

By using defeasible default rules, we are able to draw inferences from rules that would otherwise conflict to generate paralyzing contradictions. Inference is non-monotonic because adding new facts may result in withdrawing inferences made previously. For example, if we add the following fact to the previous program,

```
penguin(tweety)
```

then the program withdraws the inference that tweety flies, and now infers that tweety, like chilly, does not fly because tweety is a penguin.

```
Answer: 1

bird(tweety) bird(chilly) penguin(chilly) penguin(tweety) -flies(chilly)
-flies(tweety)

SATISFIABLE
```

Answer sets as explicit choices

Suppose we alter the strict rule that no penguins fly to be a defeasible default rule as well:

```
% Default rule d2: penguins do not fly
-flies(X) :- penguin(X), not flies(X).  % except when they are penguins that fly
```

```
Answer: 1

bird(tweety) bird(chilly) flies(tweety) penguin(chilly) -flies(chilly)
```

```
Answer: 2

bird(tweety) bird(chilly) flies(tweety) penguin(chilly) flies(chilly)

SATISFIABLE
```

Now there are *two* answer sets, each corresponding to the defeat of one or the other of the default rules d1 (birds fly) or d2 (penguins do not fly). Chilly either does not fly because she is a penguin, despite being a bird (Answer 1), or she flies because she is a bird, despite being a penguin (Answer 2). The answer sets semantics assigns no automatic priority to one default rule or the other and, instead, simply enumerates all the answer sets.

To resolve the conflict into a single answer set, we have to *explicitly qualify* one or the other of the default rules. Of course, here we should rationally choose to qualify the rule that birds fly so as to make an exception to that rule for birds that are penguins, which do not fly. This qualification is rational because our only support for believing that chilly flies is her strict inclusion in the birds class by way of being a penguin. It is therefore rational to follow a "specificity" heuristic of inference that the more specific rule for the case should take priority, in the absence of any other information.

Some defeasible logic inference systems automatically assume specificity or other such priorities between defeasible rules, though usually the heuristic can be disabled (see Nute 1993). That is, in this example, birds "normally" fly but class inclusion indicates that there is nothing "abnormal" about a penguin that does not fly (Nute 1993: 106-7). Such systems would automatically apply specificity and so there would be no need to make an explicit exception to the default rule that birds fly.

But while the specificity heuristic may be appropriate in doxastic or epistemic contexts, a deontic legal logic should *not* automatically impose this or, indeed, any other automatic priority ordering upon its rules beyond those imposed by literal non-contradiction and other incontestable logical principles such as modus ponens. It is civil authority (e.g., a court or legislature) that should determine what sorts of arguments sufficiently establish priorities between conflicting laws, not programmers or logicians. The more specific law may not always take priority over the more general one, even given class inclusion, though perhaps that is usually the case. For example, suppose 1) killers should be punished, and 2) all soldiers are killers, but 3) soldiers should not be punished. We should not automatically infer that if buzz is a soldier, then she should not be punished for being a killer. Perhaps buzz's killings constitute war crimes. U.S. law, for example, has a number of complex priority rules that are themselves the subject of an extensive body of law. Interpreting these rules must be left to civil authority.

In answer set programming, priority orderings between rules must be made explicit, which are therefore all "abnormal" cases as far as the default rule is concerned. I therefore add an explicit exception to the rule that birds fly for the abnormal case of birds that are penguins:

```
% default rule d1 that birds fly
flies(X) :- bird(X), not ab(d1(X)).
```

```
ab(d1(X)) :- penguin(X).  % explicit exception to d1 for "abnormal" cases like
penguins
```

```
Answer: 1

bird(tweety) bird(chilly) ab(d1(chilly)) flies(tweety) penguin(chilly)
-flies(chilly)

SATISFIABLE
```

This qualification on default rule d1 resolves the conflict between the rules in order to generate a single answer set in which tweety but not chilly flies because chilly is a penguin (ab(d1(chilly)).

In d1, I replaced the "normal" default rule justification (i.e., "not -flies(X)") with an explicit justification "not ab(d1(X))" to better define the "abnormal" or exception case of birds that are penguins, which do not fly. This may recall how the extension of the "abnormal" predicate is circumscribed in circumscription logic, but my aim is to avoid the details of any particular non-monotonic reasoning formalism. The answer sets semantics captures the main elements of any of the various non-monotonic reasoning formalisms such as default logic, autoepistemic logic, etc.

Making exceptions explicit by using circumscribed predicates in non-normal default rules will be useful for tracking which rules were qualified to generate each answer set. The answer set generated above shows that the exception to default rule d1 (birds fly), which is ab(d1(chilly)) (but penguins do not fly) was triggered to qualify the default rule in the case of chilly.

If we had more information about chilly, then there might be some support for thinking that chilly flies despite being a penguin, and so it might be rational to qualify the default rule that penguins do not fly, instead. For example, perhaps chilly is a penguin who happens to have secured employment as an airplane pilot, and pilots fly:

```
pilot(chilly).

% default rule d3: pilots fly
flies(X) :- pilot(X), not -flies(X).    % except when they do not fly
```

```
Answer: 1

bird(tweety) bird(chilly) ab(d1(chilly)) flies(tweety) penguin(chilly)
pilot(chilly) flies(chilly)

Answer: 2

bird(tweety) bird(chilly) ab(d1(chilly)) flies(tweety) penguin(chilly)
pilot(chilly) -flies(chilly)

SATISFIABLE
```

Again we have two conflicting answer sets, one still relying on the fact that chilly is a penguin and so does not fly (in accordance with the previous qualification on the default rule for birds), and the other relying on the fact that chilly is nevertheless a pilot and so does fly. (If we also removed the previous qualification on the birds rule, we would have three answer sets.)

To resolve the conflict we must choose either to qualify the rule concerning pilots so as not to apply to penguin-pilots (and so chilly does fly) (Answer: 1), or to quality the rule regarding penguins so as not to apply to pilot-penguins (and so chilly still does not fly) (Answer: 2). (Note that the specificity heuristic, if it were applied, would not resolve the issue, since not all pilots are penguins or vice versa.) We must choose how we prefer to resolve the conflict, and each answer set corresponds to a choice we might make.

Perhaps the choice here is an easy one, given our likely intent in adding the rule concerning pilots and the fact that chilly is a pilot. It seems likely that by adding this rule and fact we intend to carve out an exception to the penguins rule that allows us to infer that chilly nevertheless flies. Here is the program with the penguins rule qualified accordingly, followed by its answer set:

```
% default rule d1: birds fly
flies(X) :- bird(X), not ab(d1(X)).
ab(d1(X)) :- penguin(X).

% default rule d2: penguins do not fly
-flies(X) :- penguin(X), not ab(d2(X)).
ab(d2(X)) :- pilot(X).       % explicit exception to d2: but pilot
penguins do fly

% all penguins are birds (not defeasible)
bird(X) :- penguin(X).
```

```
% default rule d3: pilots fly
flies(X) :- pilot(X), not -flies(X).

bird(tweety).
penguin(chilly).
pilot(chilly).
```

```
Answer: 1

bird(tweety) bird(chilly) ab(d2(chilly)) ab(d1(chilly)) flies(chilly)
flies(tweety) penguin(chilly) pilot(chilly)

SATISFIABLE
```

Chilly now flies because she is a pilot (d3), despite being a penguin (ab(d2)(chilly)), and despite that birds that are penguins do not fly (ab(d1)(chilly)). Here we prefer to qualify rule d2 (that penguins do not fly) for pilot-penguins because that reflects our intent better than arguing that chilly's penguin nature should make flying impossible for her, even if she is a pilot.

But we might have chosen to qualify the pilot rule d3, instead:

```
% d3: pilots fly
flies(X) :- pilot(X), not ab(d3((X)).
ab(d3(X)) :- penguin(X).  % explicit exception to d3: but penguin pilots do NOT
fly
```

```
Answer: 1
```

```
bird(tweety) bird(chilly) ab(d3(chilly)) ab(d1(chilly)) flies(tweety)
penguin(chilly) -flies(chilly) pilot(chilly)

SATISFIABLE
```

Here, we have decided that chilly still cannot fly despite being a pilot (and a bird), because she is a penguin. Note again that by adding explicit exceptions to the rules, the answer set tracks what exceptions generated this resolution of the conflict. This will prove useful in the legal case.

2. Evaluating conflicts of legal duties in answer set programming

Regimenting the encoding

Suppose we want to encode a conflict between defeasible legal rules that 1) killers should be punished and 2) minors should not be punished, in a case where a minor kills. We want to know what our reasonable alternatives are, given the applicable rules and our facts.

But the following natural encoding of the rules and facts has just one answer set, consisting only in the initial facts:

```
p :- k, not m.
-p :- m, not k.
k.  m.
```

candidate	reduct	consequences
{k,m}	{k. m.}	{k,m}
{k}	{p :- k. k. m.}	{k,m,p}

Encoding "normal" default rules, on the other hand, generates the results we want:

```
p :- k, not -p.

-p :- m, not p.

k. m.
```

candidate	reduct	consequences
{k,m,p}	{p :- k. k. m.}	{k,m,p}
{k,m,-p}	{-p :- m. k. m.}	{k,m,-p}

While this is workable, we would like to have the ability to make explicit but defeasible qualifications on the rules. The following encoding, however, has no answer sets (i.e., is unsatisfiable):

```
p :- k, not q1.

-p :- m, not q2.

k.  m.
```

candidate	reduct	consequences
{}	{p :- k. -p :- m. k. m.}	null
{k, m, p}	(same)	null

{k, m, q1}	{-p :- m. k. m.}	{-p, k, m}
...		

The following is the most flexible version of the encoding, which leaves room to add more qualifications as needed and tracks the qualifications activated as they appear in each answer set:

```
p :- k, not q1.

q1 :- a, not p.

-p :- m, not q2.

q2 :- a, not -p.


k.  m.  a.
```

Answer sets: {k a m q2 q1} {k a m q2 p} {k a m -p q1} Answer sets appropriately reflect qualifications on both rules (q2 q1) , and then on one rule (q2), or the other (q1).

Hence the head of a normative legal rule in this logic should be a deontic prescription on an action predicate (e.g., that an action that is a killing is obligatory, permissible, omissible, etc.), while the body of the rule should consist of legal theories relevant to establishing the deontic status of the action (e.g., that the action constitutes murder, or is an act of necessity, or a perjury, etc.). The body is completed with a generic defeasible qualification ("qual(r1(A)))") on the rule that is tagged with the rule number (r1). For example:

```
% r1: it is permissible to kill in self-defense
pe(kill(A)) : - self_defense(A), not qual(r1(A)).
qual(r1(A))) :- act(A), not pe(kill(A)).
```

The "act(A)" part of this qualification is the same action as in the head of the rule, and will be added as a fact when describing the situation in order to activate the generic qualification defeasibly.

Legal theories that might establish one or another deontic prescription in the heads of rules are then defined independently, perhaps by extraction from a semantic legal knowledge base:

```
% legal elements of self-defense, retreat jurisdiction
self_defense(A) :- force(A), attacked(P), retreat(P).
```

A situation is then described that satisfies elements of various legal theories:

```
% situation
attacked(me). force(shooting). retreat(me). act(shooting).
```

Now add an integrity constraint for deontic contradiction, as well as standard deontic implication and deontic equivalences between obligation and permission:

```
% deontic conflict (contrary)
:- ob(A), ob(-A).
```

```
% deontic implication (subalternation)
pe(A) :- ob(A).  % obligation implies permission

% deontic equivalences
ob(A) :- -pe(-A).  -pe(-A) :- ob(A).
pe(A) :- -ob(-A).  -ob(-A) :- pe(A).
ob(-A) :- -pe(A).  -pe(A) :- ob(-A).
pe(-A) :- -ob(A).  -ob(A) :- pe(-A).
```

With these additions, the program above has the following answer sets:

```
Answer: 1

retreat(me) self_defense(shooting) act(shooting) attacked(me) force(shooting)
qual(r1(shooting))

Answer: 2

retreat(me) self_defense(shooting) act(shooting) attacked(me) force(shooting)
pe(kill(shooting)) -ob(-kill(shooting))

SATISFIABLE
```

The first answer set represents no ruling at all on the case. It will always be possible to qualify all defeasible legal rules and infer no conclusions except the given facts and strict implications from them. We will want to eliminate this possibility. The second answer is that the shooting in the situation is permissible ("pe(kill(shooting))") because it satisfied all the legal elements of self-defense. There is no conflict yet in this case, and the only answer set that rules on the case permits the shooting.

Evaluating legal conflicts as choices between answer sets

Suppose we add another rule, to set up a potential conflict of rules:

```
% r2: it is obligatory not to kill when the killing is murder

ob(-kill(A)) :- murder(A), not qual(r2(A)).

qual(r2(A))) :- act(A), not ob(-kill(A)).
```

And then we add some basic legal elements of murder and more pertinent facts to the situation:

```
% legal elements of murder

murder(A) :- malice(I), kill(A), person(P).


malice(intent). kill(shooting).  person(chilly).
```

The program now produces the following answer sets:

```
Answer: 1

retreat(me) self_defense(shooting) act(shooting) murder(shooting)

attacked(me) force(shooting) person(chilly) kill(shooting) malice(intent)

qual(r2(shooting)) qual(r1(shooting))

Answer: 2

retreat(me) self_defense(shooting) act(shooting) murder(shooting)

attacked(me) force(shooting) person(chilly) kill(shooting) malice(intent)

qual(r2(shooting)) pe(kill(shooting)) -ob(-kill(shooting))
```

```
Answer: 3

retreat(me) self_defense(shooting) act(shooting) murder(shooting)
attacked(me) force(shooting) person(chilly) kill(shooting) malice(intent)
ob(-kill(shooting)) -pe(kill(shooting)) pe(-kill(shooting))
qual(r1(shooting)) -ob(kill(shooting))

SATISFIABLE
```

The first answer set is again where all rules are qualified and no ruling is made in the case. Let us go ahead and eliminate this possibility. If we add a generic clause underneath each prescriptive rule that indicates that the head of the rule has been proven, for example,

```
% r1: it is permissible to kill in self-defense
pe(kill(A)) : - self_defense(A), not qual(r1(A)).
qual(r1(A))) :- act(A), not pe(kill(A)).
ruling :- pe(kill(A)).
```

then the following will force at least one definite ruling in the case

```
problem :- not ruling, not problem.
```

Any answer sets that do not have at least one ruling will have a "problem" they cannot resolve without internal contradiction, so eliminating that answer set from the results. (If there are no rulings, then the program will be unsatisfiable.) Now the program produces just the answer sets that are rulings:

```
Answer: 1

retreat(me) self_defense(shooting) act(shooting) murder(shooting)
attacked(me) force(shooting) person(chilly) kill(shooting) malice(intent)
ob(-kill(shooting)) -pe(kill(shooting)) ruling pe(-kill(shooting))
qual(r1(shooting)) -ob(kill(shooting))

Answer: 2

retreat(me) self_defense(shooting) act(shooting) murder(shooting)
attacked(me) force(shooting) person(chilly) kill(shooting) malice(intent)
qual(r2(shooting)) ruling pe(kill(shooting)) -ob(-kill(shooting))

SATISFIABLE
```

The first answer set qualifies the murder rule (r2) to allow the shooting in self-defense despite that it is a murder, while the second answer set qualifies the self-defense rule (r1) to forbid the shooting because it is a murder.

Adding the following #show directives will prevent showing all the facts of the situation, which are common to all answer sets in the results, as well as avoid displaying duplicative deontic equivalences between prescriptive statements of permission and obligation. We want to show just what legal predicates the facts of the situation satisfies and what qualifications and prescriptive rulings have been made in each answer set. (This might be streamlined in a complete system by collecting all legal theories into a list and then using #show for any theory in the list but for now, the following will do.)

```
#show pe/1. #show ob/1.
#show qual/1.
```

```
#show self_defense/1.
#show murder/1.
```

The program now produces the following cleaner answer sets:

```
Answer: 1

murder(shooting) self_defense(shooting) ob(-kill(shooting)) pe(-
kill(shooting)) qual(r1(shooting))

Answer: 2

murder(shooting) self_defense(shooting) qual(r2(shooting)) pe(kill(shooting))

SATISFIABLE
```

The normative decision to be made is clear. To resolve the conflict in this situation, we must either qualify the rule permitting killing in self-defense (qual(r1())) such that it does not apply in murder cases (and so one is obligated not to kill in the case), or we must qualify the rule forbidding killings that are murders (qual(r2))) such that the rule does not apply in cases of self-defense (and so one is permitted to kill in the case).

In the common law, a killing that would otherwise be a murder is justified if the killing meets the elements of self-defense. Hence we should resolve the conflict by qualifying the murder rule in cases of self-defense. To do that, we make the self-defense qualification strict, but retain a generic defeasible qualification so that the murder rule will remain a candidate for answer sets in future cases of conflict (other than in the case of self-defense):

```
% r2: it is obligatory not to kill when the killing is murder
ob(-kill(A)) :- murder(A), not qual(r2(A)), not qual(r21(A)).
qual(r2(A)) :- act(A), not ob(-kill(A)).
qual(r21(A)) :- self_defense(A).        % except when killing in self-defense
ruling :- ob(-kill(A)).
```

Now the program produces only one answer set, where the murder rule is qualified in case of self-defense in the situation described:

```
Answer: 1

murder(shooting) qual(r21(shooting)) qual(r2(shooting))
self_defense(shooting) pe(kill(shooting))

SATISFIABLE
```

Suppose now that the killing was not in self-defense because I refused to retreat when the attacker was no longer a threat, but suppose further that I was so terrified that I had no idea what I was doing when I killed my attacker. We might then add the following rule and facts (and also delete the "retreat(me)" fact):

```
% r3: it is permissible (excusable) to kill if one is insane
pe(kill(A)) :- act(A), insanity(P), not qual(r3(A)).
qual(r3(A)) :- act(A), not pe(kill(A)).
ruling :- pe(kill(A)).

% legal insanity
```

```
insanity(P) :- no_understanding_act(P), no_knowledge_wrong(A).
#show insanity/1.

# new facts
no_understanding_act(me).  no_knowledge_wrong(shooting).
```

```
Answer: 1

insanity(me) murder(shooting) ob(-kill(shooting)) pe(-kill(shooting))
qual(r3(shooting)) qual(r1(shooting))

Answer: 2

insanity(me) murder(shooting) qual(r20(shooting)) pe(kill(shooting))

SATISFIABLE
```

Now there is an answer set (Answer 2) with a further qualification (qual(r20())) on the rule barring murder (r2), reflecting the possibility that the new rule regarding insanity might defeat the murder rule. Answer 1 represents an unqualified murder rule, with (generic) qualifications imposed, instead, on both the new insanity rule we added (qual((r3())), and the self-defense rule qual(r1())), which may now be qualified with respect to the murder rule because we removed a necessary element of the self-defense theory.

If we choose to resolve the new conflict by excusing killings by reason of insanity, we add a new strict qualification to the murder rule for insanity:

```
% r2: it is obligatory not to kill when the killing is murder
```

```
ob(-kill(A)) :- murder(A), not qual(r2(A)), not qual(r21(A)), not qual(r22(A)).
qual(r2(A)) :- act(A), not ob(-kill(A)).
qual(r21(A)) :- self_defense(A).              % except in self-defense
qual(r22(A)) :- act(A), insanity(P).       % except for insanity
ruling :- ob(-kill(A)).
```

```
Answer: 1

insanity(me) murder(shooting) qual(r22(shooting)) qual(r2(shooting))
pe(kill(shooting))

SATISFIABLE
```

This resolves the new conflict so that the shooting is again permissible despite meeting the elements of murder, though the shooting is now permissible, instead, by reason of insanity.

Here is the complete program:

```
% deontic conflict (contrary)
:- ob(A), ob(-A).

% deontic implication (subalternation)
pe(A) :- ob(A).  % obligation implies permission
#show pe/1. #show ob/1.

% deontic equivalences
ob(A) :- -pe(-A).  -pe(-A) :- ob(A).
pe(A) :- -ob(-A).  -ob(-A) :- pe(A).
ob(-A) :- -pe(A).  -pe(A) :- ob(-A).
pe(-A) :- -ob(A).  -ob(A) :- pe(-A).
```

```
%%%%%%%%%%%%%% rules

% a ruling is required

problem :- not ruling, not problem.


% r1: it is permissible to kill in self-defense

pe(kill(A)) :- self_defense(A), not qual(r1(A)).

qual(r1(A)) :- act(A), not pe(kill(A)).

ruling :- pe(kill(A)).


% r2: it is obligatory not to kill when the killing is murder

ob(-kill(A)) :- murder(A), not qual(r2(A)), not qual(r21(A)), not qual(r22(A)).

qual(r2(A)) :- act(A), not ob(-kill(A)).

qual(r21(A)) :- self_defense(A).              % except in self-defense

qual(r22(A)) :- act(A), insanity(P).       % except for insanity

ruling :- ob(-kill(A)).


% r3: it is permissible (excusable) to kill if one is insane

pe(kill(A)) :- act(A), insanity(P), not qual(r3(A)).

qual(r3(A)) :- act(A), not pe(kill(A)).

ruling :- pe(kill(A)).


#show qual/1.


%%%%%%%%%%%%%% legal theories

% legal elements of self-defense, retreat jurisdiction

self_defense(A) :- force(A), attacked(P), retreat(P).

#show self_defense/1.


% legal elements of murder

murder(A) :- malice(I), kill(A), person(P).

#show murder/1.
```

```
% legal insanity
insanity(P) :- no_understanding_act(P), no_knowledge_wrong(A).
#show insanity/1.

%%%%%%%%%%%%%%% conflict situation
attacked(me). force(shooting). act(shooting).
%-retreat(me).
malice(intent). kill(shooting).  person(chilly).
% new facts
no_understanding_act(me).  no_knowledge_wrong(shooting).
```

The general approach is to use answer set programming to expose and evaluate conflicts between rules rather than to resolve them by programmer fiat. The approach is to identify possible reasonable qualifications, render them defeasible in the rules, and then evaluate resulting answer sets, in an iterative process. The answer sets semantics eliminates some combinations of qualifications as inconsistent by applying incontestable rules of inference such as modus ponens and any inferences or equivalences that are explicitly encoded into the problem, such as the deontic contraries and subalternation relationships. The approach is related to the iterated generate-define-test problem solving methodology characteristic of answer set programming, but with explicit intervention by a legitimate public authority when needed.

A full governance system would add a querying system, build out different kinds of exceptions one might make to rules, and offer a number of post-processing functions and options, as well as fallback normative rules. The Answer Set Prolog (ASP) system used here to illustrate these examples has a number of extensions and other features that might be useful in creating such a querying and governance system (see Pottasco). My purpose here is primarily only to give a sense of how an answer set programming approach might serve a deontic logic of the law and how the approach should work. The main points are that the logic should 1) describe conflicts while at the same time insisting upon the consistency of rules, but 2) avoid forcing a more or less arbitrary resolution of those conflicts when appropriate qualifications on conflicting duties have not yet been determined by a legitimate public authority.

CHAPTER SEVEN

ANSWER SET PROGRAMMING KANT'S CONFLICT BETWEEN VERACITY AND PHILANTHROPY

1. Philosophical analysis of the conflict between veracity and philanthropy

In a late essay, “On a Supposed Right to Lie From Philanthropy” (SR), Immanuel Kant asserts an unconditional legal duty of truthfulness in one's “declarations:”

> To be truthful (honest) in all declarations is...a sacred command of reason prescribing unconditionally, one not to be restricted by any conveniences (SR: 8:427).

Kant traditionally has been misunderstood to argue dogmatically in SR that, ethically, one must always be truthful, even in a hypothetical case where a murderer appears at one's door demanding the whereabouts of a potential victim one is sheltering. Benjamin Constant supposes that Kant held this position, and argues that the duty of truthfulness should have an exception in such a case because "...no one has a right to a truth that harms others" (SR: 8:425).

Kant frames the hypothetical conflict that Constant proposes in SR as a conflict, instead, between a strict *legal* duty not to tell lies in official proceedings (i.e. a duty to avoid perjury) and an *ethical* duty to prevent harm to others (i.e., a duty of philanthropy or beneficence). Kant then argues in SR that the duty to avoid telling such lies takes priority even in a case where telling the truth might endanger others, because the duty at issue is a duty of right; “what is under discussion here is a duty of right" (SR: 8:426n). Kant in fact argues that the duty of veracity at issue in SR is not only a duty of right that therefore takes precedence over conflicting ethical reasons for action, but a constitutional duty of right that is a foundation of civil society. Judicial systems would be impossible without an enforceable duty of right barring perjury, for example. The duty is thus a "formal" or natural legal duty that overrides even the most powerful conflicting ethical reasons one might have to lie to prevent harm to others. The priority of right resolves the conflict, according to Kant.

Kant thus does not argue in SR that there is an absolute ethical duty to tell the truth that overrides all other ethical reasons one might have to lie, as Constant may suppose he does. Kant, instead, takes Constant's case of the murderer at the door as a point of departure to argue that a duty of *right* to tell the truth should be subject to no "philanthropic" ethical exception. Hence Kant does not directly address the question that Constant poses, which has generated much confusion in the vast literature on the essay (see Wood 2011). Kant is simply not interested in SR in discussing whether one might have *ethical* reasons to tell the truth that may conflict with other ethical reasons one might have not to do so. This is characteristic of Kant's understanding of the purpose of ethics, which is not the Aristotelian one of providing a guide for how one should live but, instead, to clarify and strengthen one's will to act morally. Kant engages in ethical "casuistry" over particular cases only in order to strengthen the virtuous will.

2. A Flawed Attempt to Encode the Conflict Between Veracity and Philanthropy in SR

The confusion in the literature concerning the nature of the dispute between Kant and Constant in SR has carried over into efforts to grapple with the conflict in AI applications. Jean-Gabriel Ganascia (2007) attempts to use answer set programming to model the conflict of duties in the murderer at the door case in accordance with three different ethical theories, Aristotelian ethics, Kantian ethics, and Benjamin Constant's ethics. Ganascia then reviews the answer sets each model of the case generates and draws the conclusion that Constant's system should be preferred.

"Categorical Imperative"

Ganascia models the conflict in SR in accordance with Kantian ethics as follows. First Ganascia models Kant's "categorical imperative" by defining predicates that translate any maxim with the literal "I" into maxims for anyone ("P"). This is supposed to reflect the universalizability requirement of the categorical imperative.

```
maxim_will("I", answer_question("I"), tell("I", lie)).
maxim_will(P, answer_question(P), tell(P, S)) :-
             maxim_will("I", answer_question("I"), tell("I", S)),
             not maxim_will(P, answer_question(P), tell(P, SS)),
             neq(S, SS).
```

Then Ganascia encodes that if anyone has a maxim of lying, then that leads to a lack of "trust":

```
untrust(P):- maxim_will(P, G, tell(P, lie)).
trust(P) :- not untrust(P).
```

The result is that all answer sets containing maxims of lying also include a lack of trust ("untrust"), whereas answer sets containing maxims of telling the truth (and, here, murder) do not. A lack of trust is apparently an intolerable consequence, according to Ganascia, so there is no weighing of any other consequences (Ganascia 2007: 46).

"Aristotelian rules"

The key to how Ganascia models the conflict using "Aristotelian rules" is as follows (Ganascia 2007: 43, 46):

```
worse(tell(P, lie), A) :- neq(A, tell(P, lie)), neq(A, murder).
```

Here, Ganascia encodes that telling a lie is worse than action A so long as A is not a murder (and not itself a lie). The rest of the "Aristotelian" program simply determines whether lying or telling the truth delivers the least worst consequences, where telling the truth leads to a murder. Predictably, all answer sets require telling the truth. Ganascia then remarks that "if we replace [the rule above] by the rule

```
worse(tell(P, lie), A) :- neq (A, tell(P, lie)).
```

then some answer sets (i.e. some possible worlds) assert that lying is unjust while others assert that murder is unjust" (Ganascia 2007: 46). Since lying is not encoded as being worse than murder, some answer sets will determine lying to be worse than murder and others will not, where both lying and telling the truth are inconsistent.

"Constant's ethical conception"

Ganascia then updates these Aristotelian rules with Constant's principle that one may lie to those who do not deserve the truth, where here, again, determining whether another deserves the truth is simply a matter of determining whether telling them the truth will result in worse consequences than will lying to them (presumably as a result of action they take):

```
principle(P, answer_question(P, PP), tell(P, PP, truth)) :- not not_deserve(PP,
tell(P, PP, truth)).
```

```
principle(P, answer_question(P, PP), tell(P, PP, lie)) :-not_deserve(PP, tell(P,
PP, truth)).

not_deserve(PP, tell(P, PP, truth)):-
             worst_consequence(tell(P, PP, truth), C),
             worse(C, tell(P, PP, lie)).
```

The result is that all answer sets require lying to the murderer, since the murderer does not deserve the truth, since telling the murderer the truth leads to worse consequences than does lying. Unlike Ganascia's encoding of the Aristotelian and Kantian conceptions, his encoding of "Constant's ethical conception" exploits the defeasible reasoning answer set programming affords. Ganascia encode a rule that one should tell the truth except when it is provable that the truth is not deserved, and then providing the elements that define when truth is not deserved (not_deserve/2) elsewhere in the program. This is roughly how I have argued that legal rules and supporting theories should be encoded. But note that Constant's rule might also *permit* telling the truth to the murderer, and Ganascia's encoding fails to generate any answer sets that reflect this possibility.

Ganascia's contribution is to demonstrate how answer set programming can model defeasible rules by supplying principled exceptions. I will not criticize Ganascia's attempts to capture Aristotelian or Kantian ethical theory (or Constant's ethical theory, with which I am less familiar) in answer set programming, except to note that I fail to see how these attempts reasonably conform to Aristotelian or Kantian moral principles. Ganascia renders both Aristotelian virtue ethics and Kant's categorical imperative improperly in consequentialist terms (Ganascia 2007: 43, 44). But virtue ethics is, instead, about good judgment above all, since what is ethical in any situation is what a person of good character would choose to do in that situation; and while Kant's categorical imperative is sometimes rendered by critics in consequentialist terms (e.g., famously, by the utilitarian J.S. Mill), Kant would reject such interpretations. Ganascia's main substantive ethical point appears to be that ethical rules should be subject to principled exceptions, a general point with which I think all three of Aristotle, Kant and Constant would agree, if properly understood.

The main problem with Ganascia's approach for our purposes is that it fails to respect the distinction between law and ethics and the priority of right, which is in fact decisive in the case at issue in SR. Ethical principles are mingled with quasi-legal rules in his models, and then answer set programming is deployed to resolve conflicts more or less arbitrarily. Ganascia's approach to Constant's conception makes more sense, as there Ganascia shows how the addition (or absence) of an additional qualifying principle can affect how the conflict is resolved. But it is difficult to see what purpose Ganascia's Kantian or Aristotelian models might serve, since the resulting answer sets merely reflect normative choices the programmer (Ganascia) made while encoding it. As I argued in the previous chapter, a deontic logic of the law should expose the rules that generate conflicts so that qualifications can be reviewed and ruled upon by a legitimate public authority such as a court.

3. Answer set programming the conflict in SR

In SR, Kant evaluates the duty of veracity against two conflicting ethical rules of "philanthropy" that would either 1) *permit* one to lie to avoid harm, or 2) *obligate* one to lie to avoid harm. I will show how to use answer set programming to evaluate both ethical rules against the duty of veracity in what follows. Kant's analysis would be much briefer than what I will present here, since the only duty of right in the case is the constitutional duty of veracity (e.g., to avoid perjury), which therefore takes priority over competing ethical duties of philanthropy. Hence there is little or no genuine conflict in the case for Kant. But I will proceed as if Kant might have accepted some legal duty of philanthropy that might permit lying in some circumstances, in order to illustrate how answer set programming might model the conflict.

First, I determine what predicate actions generate the conflict. Here, it is telling a lie in an official proceeding in order to prevent someone from being harmed. One is either obligated to tell the truth, or permitted to lie, or obligated to lie, to prevent harm, in such a situation. These are the alternatives that Kant evaluates in SR. I first encode each rule in template form:

```
% r1: it is obligatory to tell the truth in testimony
ob(tell_truth(A)) :- testimony(A), not qual(r1(A)).
qual(r1(A)) :- act(A), not ob(tell_truth(A)).
ruling :- ob(tell_truth(A)).
```

```
% r2: it is permissible to lie out of philanthropy
pe(-tell_truth(A)) :- philanthropy(A), not qual(r2(A)).
qual(r2(A)) :- act(A), not pe(-tell_truth(A)).
ruling :- pe(-tell_truth(A)).

% r3: it is obligatory to lie out of philanthropy
ob(-tell_truth(A)) :- philanthropy(A), not qual(r3(A)).
qual(r3(A)) :- act(A), not ob(-tell_truth(A)).
ruling :- ob(-tell_truth(A)).
```

I then encode legal theories of testimony and philanthropy:

```
% legal elements of testimony
testimony(A) : intentional(A), tell(A), material(S), statement(S),
under_oath(P).
#show testimony/1.

% a general legal theory of "philanthropy" (note: not plausible at common law)
philanthropy(A) :- prevent_harm(A).
#show philanthropy/1.
```

I then describe the conflict situation:

```
% situation: someone lies in an official proceeding in order to avoid harm
intentional(response). tell(response). material(whereabouts).
statement(whereabouts). under_oath(me).

-tell_truth(response).
prevent_harm(response).
```

```
act(response).
```

Finally I add the preamble defining deontic conflict, deontic implication, and convenient equivalences:

```
% deontic conflict (contrary)
:- ob(A), ob(-A).

% deontic implication (subalternation)
pe(A) :- ob(A).  % obligation implies permission

% deontic equivalences
ob(A) :- -pe(-A).  -pe(-A) :- ob(A).
pe(A) :- -ob(-A).  -ob(-A) :- pe(A).
ob(-A) :- -pe(A).  -pe(A) :- ob(-A).
pe(-A) :- -ob(A).  -ob(A) :- pe(-A).

% directives to show only positive prescriptions in results
#show pe/1. #show ob/1.

% a ruling is required
problem :- not ruling, not problem.

% show rule qualifications
#show qual/1.
```

The resulting program generates the following answer sets:

```
Answer: 1

philanthropy(response) testimony(response) ob(tell_truth(response))
qual(r3(response)) pe(tell_truth(response)) qual(r2(response))

Answer: 2

philanthropy(response) testimony(response) qual(r1(response)) ob(-
tell_truth(response)) pe(-tell_truth(response))

Answer: 3

philanthropy(response) testimony(response) qual(r1(response))
qual(r3(response)) pe(-tell_truth(response))

SATISFIABLE
```

The total number of possible answer sets is the power set of the available qualifications, P(qual(r1()), qual(r2()), qual(r3())), or eight total sets, including the empty set: {(qual(r1), qual(r2), qual(r3)), (qual(r1), qual(r2)), (qual(r1), qual(r3)), (qual(r1)), (qual(r2), qual(r3)), (qual(r2)), (qual(r3)), ()}. Three combinations of qualifications are necessarily inconsistent. Qualifying only rule 2 or only rule 3 but not rule 1 is inconsistent because if it is obligatory to tell the truth (r1) (ob(tell_truth(response))), then it can be neither obligatory (r3) (ob(-tell_truth(A)) nor permissible (r2) (pe(-tell_truth(A)) to lie. Qualifying rules 1 and 2 but not 3 is inconsistent because of deontic subalternation: If lying is obligatory (r3), then it is permissible (r2); rule 2 (r2) therefore cannot be qualified unless rule 3 (r3) also is. Since we discard the empty set and eliminate the no-decision set in which all the rules are qualified via the "ruling" predicate, three answer sets remain. These three correspond to rulings that telling the truth is obligatory (Answer 1) in all testimony, or that lying is obligatory (Answer 2) or merely permissible (Answer 3) in cases where that lying is philanthropic.

Rules 2 and 3 regarding the (supposed) legal duty of philanthropy make no reference to lying testimony, however, and it seems reasonable that they might be qualified when the philanthropy requires a lie that constitutes perjury. (Certainly Kant would make these qualifications.) So I add those potential qualifications to rules 2 and 3 (in bold), while updating defeasible justifications accordingly. (I comment out the default qualifications in rules for now to remove clutter, although it is not necessary.)

```
% r2: it is permissible to lie out of philanthropy
pe(-tell_truth(A)) :- philanthropy(A), -tell_truth(A),
                          not qual(r2(A)), not qual(r21(A)).
%qual(r2(A)) :- act(A), not pe(-tell_truth(A)).
qual(r21(A)) :- perjury(A), not pe(-tell_truth(A)).  % except when it's perjury
ruling :- pe(-tell_truth(A)).

% r3: it is obligatory to lie out of philanthropy
ob(-tell_truth(A)) :- philanthropy(A), -tell_truth(A),
                          not qual(r3(A)), not qual(r31(A)).
%qual(r3(A)) :- act(A), not ob(-tell_truth(A)).
qual(r31(A)) :- perjury(A), not ob(-tell_truth(A)).  % except when it's perjury
ruling :- ob(-tell_truth(A)).
```

I also add a legal theory of perjury:

```
% legal elements of perjury
perjury(A) :- testimony(A), material(S), -tell_truth(A).
#show perjury/1.
```

Since rule 1 that one must tell the truth in testimony, on the other hand, might be subject to an explicit qualification for lies that are philanthropic, I add an explicit potential qualification to that rule as well:

```
% r1: it is obligatory to tell the truth in testimony
ob(tell_truth(A)) :- testimony(A), not qual(r1(A)), not qual(r11(A)).
%qual(r1(A)) :- act(A), not ob(tell_truth(A)).
qual(r11(A)) :- philanthropy(A), -tell_truth(A),
                    not ob(tell_truth(A)).    % except if a lie is philanthropic
ruling :- ob(tell_truth(A)).
```

These changes generate the following answer sets:

```
Answer: 1

philanthropy(response) perjury(response) testimony(response)
ob(tell_truth(response)) qual(r31(response)) pe(tell_truth(response))
qual(r21(response))

Answer: 2
```

```
philanthropy(response) perjury(response) testimony(response)

qual(r11(response)) ob(-tell_truth(response)) pe(-tell_truth(response))


Answer: 3


philanthropy(response) perjury(response) testimony(response)

qual(r11(response)) qual(r31(response)) pe(-tell_truth(response))


SATISFIABLE
```

Suppose a court chooses to qualify rules 2 and 3 in cases of perjury (so, Answer 1). Lying is neither permissible (rule 2) nor a fortiori obligatory (rule 3) for philanthropic reasons when the lie constitutes perjury (so qual(r21()) and qual(r31())). To do so, we simply make the qualifications on r2 and r3 for perjury strict (indefeasible) ones (and uncomment the default qualification to leave open the possibility of further qualifications of the rules in future conflicts):

```
% r2: it is permissible to lie out of philanthropy
pe(-tell_truth(A)) :- philanthropy(A), -tell_truth(A),
                      not qual(r2(A)), not qual(r21(A)).
qual(r2(A)) :- act(A), not pe(-tell_truth(A)).
qual(r21(A)) :- perjury(A).            % except when it's perjury
ruling :- pe(-tell_truth(A)).

% r3: it is obligatory to lie out of philanthropy
ob(-tell_truth(A)) :- philanthropy(A), -tell_truth(A),
                      not qual(r3(A)), not qual(r31(A)).
qual(r3(A)) :- act(A), not ob(-tell_truth(A)).
qual(r31(A)) :- perjury(A).           % except when it's perjury
ruling :- ob(-tell_truth(A)).
```

This resolves the conflict. According to these rules, it is generally obligatory to tell the truth with a possible exception for cases of philanthropy; however, one is neither permitted nor obligated to lie in cases of philanthropy when that lying constitutes perjury.

```
Answer: 1

philanthropy(response) perjury(response) testimony(response)
qual(r31(response)) ob(tell_truth(response)) qual(r3(response))
pe(tell_truth(response)) qual(r21(response)) qual(r2(response))

SATISFIABLE
```

Of course, according to Kant, since the duty of veracity is a duty of right, and indeed a constitutional duty of right, Kant likely would completely reject the idea that it could have any exceptions for philanthropy, which is a dubious *legal* duty anyway. Kant thus might have reached the same answer set as that above by rendering r1 strictly, like this:

```
% r1: it is obligatory to tell the truth in testimony
ob(tell_truth(A)) :- testimony(A).
ruling :- ob(tell_truth(A)).
```

This rule requiring truthfulness of testimony is subject to no possible defeat, and any other rules in conflict must be qualified to conform to its requirements.

4. The indefeasibility of the duty of veracity in SR

The argument between Kant and Constant in SR is not about the best way to resolve conflicts between ethical obligations such as that between being truthful and protecting innocents from harm, however. Kant does not argue in SR that we should resolve conflicts by maintaining one or the other duty unconditionally, as opposed, on the other hand, to Constant's view that we should qualify one or the other duty by making a principled exception.

Kant's main goal in SR is, instead, to reframe the issue in order to clarify the question. Kant argues that a strict *legal* obligation not to lie in civil proceedings or contexts (e.g., when being interviewed by the police, or when testifying in court) must hold even when there are powerful competing *ethical reasons* to lie, such as that one's lie would protect an innocent from harm. The case Kant has in mind in SR is thus most similar to a case where committing perjury in testimony in a court of law might protect someone who is innocent from harm. This question is considerably more difficult to answer by consulting moral intuition.

What Kant points out is that if one does indeed have a strict legal obligation not to lie, then that obligation is already authoritatively specified in the system of equal freedom under universal laws. The question as to whether one might make exceptions to such a duty when there is a conflict is, therefore, a nonstarter for Kant. There cannot be conflicts between one's strict legal obligations and conflicting ethical duties, Kant insists; the very question is premised on a confusion. If there were such conflicts, then the priority of right would not exist; moreover, the prescriptive legal system would be inconsistent, which is normatively intolerable.

Recall that in DR Kant argued that in a case of self-defense, one acts legally rightfully when killing one's assailant, despite that there might be good ethical reasons not to defend oneself by killing another. Perhaps Kant imagines that an enlightened soul might indeed think it better to be killed oneself than to kill another human being under any circumstances. Kant's argument in SR follows a similar pattern: One acts legally rightfully when one is truthful in one's declarations, even despite that there might be strong ethical reasons to lie. Hence though Kant does not mention it in SR, it seems possible that one's ethical reasons to lie could ripen into an obligation that outweighs even the strict legal obligation to avoid perjury. Kant regards the duty of veracity in SR as considerably more important than almost any other duty one might have, however, as he thinks the rule against perjury is a foundation of the civil state, one as important as constitutional guarantees of freedom, equality and the rule of law. Perjury is thus a 'formal' wrong, Kant says, even when it is not a material one. So while the theoretical possibility is there, it seems doubtful that Kant would allow perjury under any circumstances.

Here again is the complete program:

```
% deontic conflict (contrary)
:- ob(A), ob(-A).

% deontic implication (subalternation)
pe(A) :- ob(A).  % obligation implies permission

% deontic equivalences
ob(A) :- -pe(-A).  -pe(-A) :- ob(A).
pe(A) :- -ob(-A).  -ob(-A) :- pe(A).
ob(-A) :- -pe(A).  -pe(A) :- ob(-A).
pe(-A) :- -ob(A).  -ob(A) :- pe(-A).

% directives to show only positive prescriptions in results
#show pe/1. #show ob/1.

%%%%%%%%%%%%%% legal theories
% a ruling is required
problem :- not ruling, not problem.

% r1: it is obligatory to tell the truth in testimony
ob(tell_truth(A)) :- testimony(A), not qual(r1(A)), not qual(r11(A)).
%qual(r1(A)) :- act(A), not ob(tell_truth(A)).
qual(r11(A)) :- philanthropy(A), not ob(tell_truth(A)).    % except when lying
is philanthropic
ruling :- ob(tell_truth(A)).

% r2: it is permissible to lie out of philanthropy
pe(-tell_truth(A)) :- philanthropy(A), not qual(r2(A)), not qual(r21(A)).
qual(r2(A)) :- act(A), not pe(-tell_truth(A)).
qual(r21(A)) :- perjury(A).              % except when it's perjury
ruling :- pe(-tell_truth(A)).

% r3: it is obligatory to lie out of philanthropy
ob(-tell_truth(A)) :- philanthropy(A), not qual(r3(A)), not qual(r31(A)).
qual(r3(A)) :- act(A), not ob(-tell_truth(A)).
qual(r31(A)) :- perjury(A).            % except when it's perjury
ruling :- ob(-tell_truth(A)).

% show rule qualifications
#show qual/1.

%%%%%%%%%%%%%% legal theories
% legal elements of perjury
perjury(A) :- testimony(A), material(S), -tell_truth(A).
#show perjury/1.

% legal elements of testimony
testimony(A) : intentional(A), tell(A), statement(S), under_oath(P).
#show testimony/1.

% a legal theory of "philanthropy"
philanthropy(A) :- prevents_harm(A).
#show philanthropy/1.

%%%%%%%%%%%%%%% conflict situation: someone lies in testimony in order to avoid
harm
intentional(response). tell(response).
material(whereabouts). statement(whereabouts). under_oath(me).
```

```
-tell_truth(response).

prevents_harm(response).

act(response).
```

CHAPTER EIGHT

ANSWER SET PROGRAMMING THE TROLLEY PROBLEM

1. The trolley problem and the Doctrine of Double Effect

Much of the discussion in the AI and cognitive science community about the trolley problem concerns the so-called Doctrine of Double Effect (DDF) or Triple Effect (DTE). The DDE is a controversial ethical principle, however, and therefore largely irrelevant to the trolley problem, at least for rightful machines. The DDE is supposed to distinguish Driver from Fat Man in the following way: According to the DDE, turning the trolley is permissible since one's intent with regard to the one person on the side track is not to *cause her death* in order to save the five on the main track, but instead, to *turn the trolley* in order to save the five, where the one's death is merely a foreseen (double) effect of turning the trolley. Whereas in the Fat Man variation, one's specific intent is to act to cause the fat man's death in order to save five.

Hence the permissibility of an action under the DDE depends on what Kant refers to as the maxim governing one's action—that is 'I will turn the trolley in order to save five' as opposed to 'I will kill one person in order to save five'—and, even more specifically, the maxim of the *end* of one's action (i.e., ' I will save five'). But maxims are not rightfully enforceable, and indeed cannot be enforced, according to Kant. Another might make me act in some way that serves her end, but short of brainwashing me or mind control, I cannot be forced to make someone else's *end* my own end. No one can make me have a specific intent or adopt a particular maxim when I act; moreover, the priority of right implies that doing so would be wrongful, anyway. I can only be rightfully forced to comply with a duty when that force is a product of my own will united with everyone else's in legitimate public authority.

Suppose I turn the trolley with the same intent I might have in the Fat Man case, that is, I specifically intend to sacrifice the one on the side track in order to save the five. Perhaps I would have been willing to push the fat man, too, for the same reason. Then the DDE would not apply, but my *legal* obligations in the case are no different. Or suppose I turn the trolley solely because I believe I will be punished if I do not turn it (rather than to save the five), or perhaps just because I think turning trolleys is fun. While these varying maxims of my action certainly affect whether my action is ethical or not, they do not affect my *legal* liability or culpability in the case. It does not matter why I turn the trolley from a legal point of view. This shows that the DDE has no application for rightful machines.

2. A Flawed Logic Programming Approach to Bystander

Recall that in the Bystander variation of the trolley problem, you are a bystander who must choose between pulling a switch to turn the trolley, so killing one, or not pulling the switch, so killing five. Pereira and Saptawijaya (2011) model Bystander in logic programming as follows. First, they identify two possible actions the bystander might take, either "watching" the trolley continue on the main track and kill five people, or "throwing_switch" to turn the trolley and kill one person:

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

The expect/1 and exclusive/2 predicates are not given in the paper, but expect/1 appears to be a wrapper to process possible choices, and they clarify that exclusive/2 requires that one or the other choice but not both appear in answer sets (Pereira and Saptawijaya 2011: 106). The exclusive/2 likely asserts something like the following behind the scenes:

```
watching :- not throwing_switch.
throwing_switch :- not watching.
```

This would generate the desired answer sets:

```
Answer: 1

throwing_switch

Answer: 2
```

```
watching


SATISFIABLE
```

Pereira and Saptawijaya then encode the first choice, "throwing_switch" to turn the trolley to the sidetrack, so killing one, as follows (the "<-" symbol in their system is similar to ":-" in our encoding):

```
redirect_train <- consider(throwing_switch).
kill(1) <- human(X), side_track(X), redirect_train.
end(save_men, ni_kill(N)) <- redirect_train, kill(N).
```

The end/2 predicate signifies the outcome of the decision and will figure in a post-processor that selects answer sets that minimize the total number of deaths, which are collected by a die/1 function. The first two clauses here say that 1) if you throw the switch, then that redirects the train; and that 2) if you redirect the train and there is a human on the side track, then one person is killed ("kill(1)"). The kill/1 predicate does not imply that you intentionally kill anyone, however; it is used merely as a synonym to connect with the die/1 predicate, which is used elsewhere. (The consider/1 predicate likely asserts an appropriate rule in the background, and we can ignore it here.)

What the third clause says is that if you redirect the train and as a result kill some number of people N, then *you do not intentionally kill* those N people, which is what the somewhat cryptic "ni_kill(N)" predicate means in their system. You also save the five men on the main track ("save_men"). Pereira and Saptawijaya thus encode the choice to turn the trolley in Bystander as one where the "end" (outcome) of the choice is that men are saved ("save_men") without intentionally killing some number of people ("ni_kill(N)").

Whereas they model the alternative choice to *not* turn the trolley, so killing five men, as "watching" the train continue on the main track, in the following way:

```
train_straight <- consider(watching).
end(die(5)) <- train_straight.
```

Here, they encode the decision as one where the "end" (again, outcome) is simply that five people die. (The "kill" predicate from before is translated into a "die" predicate later in post-processing.) This outcome does not include the predicate they designate for intentional killings ("i_kill/1"), but also does not save anyone.

Pereira and Saptawijaya do not attempt to justify the normative choices made in these encodings, except with the brief remark that "...merely watching the trolley go straight is an omission of action that just has negative consequence, whereas throwing the switch is an action that is performed to achieve a goal and additionally has negative consequence" (Pereira and Saptawijaya 2011: 105).

They then model the Principle of Double Effect with 1) an integrity constraint that "intentional killings" are never allowed, and 2) by defining outcomes ("end") that contain the predicate ("i_kill(Y)") as ones where there was an intentional killing, even if that killing also saves people ("save_men").

```
falsum <- intentional_killing.
intentional_killing <- end(save_men, i_kill(Y)).
```

The predictable result is that either choice in Bystander is permissible, since as they encoded the alternative decisions, there is no intentional killing either way. Their encoding thus produces two possible answer sets for actions in Bystander (Pereira and Saptawijaya: 109):

```
[throwing_switch], [watching]
```

They then employ a post-processing function to select the answer set in which the fewest people die:

```
[throwing_switch]
```

They do not justify this normative decision, either, although perhaps it requires little in this context.

Pereira and Saptawijaya's approach is normatively flawed in two ways. First, they make no distinction between legal and ethical duties, and no attention is paid to the priority of right. The DDE is not a legal principle but, instead, a controversial ethical one inappropriate for governing rightful machines. Many argue it is unjustifiable as an ethical principle, though I will not evaluate the principle here. Pereira and Saptawijaya apply the DDE and the related Doctrine of Triple Effect (DTE) as if these principles were authoritative, however:

> By appropriate moral decisions, we mean the ones that conform with those the majority of people make, in adhering to the principle of double effect (105).

As support for this claim, they appeal to an experiment that gathered intuitions in various trolley problem scenarios and assert that the DDE explains these intuitions (Pereira and Saptawijaya 2011: 105; see Mikhail 2007). But the cited experiment provides ambiguous support for the DDE at best. In the three trolley problem variations that are supposed to illustrate a proper application of the DDE, the actual percentages of those who judged actions permissible in the experiment were 37%, 48%, and 62%. Unlike Driver and Fat Man variations, intuitions are not as clear in these cases, yet Pereira and Saptawijaya do not report exact percentages, and instead display only a table with "permissible/impermissible" binary judgments beside each trolley problem variation (Pereira and Saptawijaya 2011: 104).

One prominent such variation referred to as "Loop track" is a case like Bystander except that the track loops back around such that it would kill the five if it did not strike the one. According to the DDE, one should not turn the trolley in Loop track, since one would choose to act to *cause the death* of the one in order to stop the trolley from looping back to kill the five. One's maxim in Loop is "I will act to cause the trolley to kill the one in order to save the five," which violates the DDE. Yet 48% of subjects nevertheless thought turning the trolley was *permissible* in Loop, or almost exactly half. Pereira and Saptawijaya do not report this percentage but simply report that the majority (i.e. 52%) thought that turning the trolley is "impermissible," in accordance with the DDE. To be fair, the original experiment also draws tendentious conclusions from sparse data concerning the trolley problem variations, though at least the original paper warns that its conclusions regarding the trolley problem are a "proposal, but both the data and hypothesis presented are preliminary" (Mkhail 2007:149). The main thesis of that paper is that there may be a universal moral grammar underlying human moral intuitions, and the trolley problem variation experiments upon which Pereira and Saptawijaya rely are a minor part of the paper set out in a sidebar. (Kant would agree on the universality of the supreme principle of morality, and that every rational being is aware of it, though I think Kant would reject the DDE.) Another problem is that when a case like Fat Man is presented before Loop then a majority choose not to turn the trolley (56%), whereas if a case like Bystander or Driver is presented before Loop, then most do choose to turn the trolley (68%) (Liao, et al., 2007: 666). Once again, intuitions are simply unclear in these cases, and "modeling" them as if they had clear solutions by appeal to ethical principles like the DDE is misleading.

Hence the first problem with Pereira and Saptawijaya's approach is that they model the trolley problem by relying on controversial ethical principles, rather than on any legal principle subject to a standard of justice. The second normative problem with their approach is that their encodings are tailored to generate the results they want in the variations. There is no moral reasoning or decision-making going on in their models, such as a decision as to whether the principle of DDE should be applied to a case. Pereira and Saptawijaya themselves decide whether and how the DDE applies to decide the cases by how they choose to encode them.

3. Answer set programming the trolley problem I: Fat Man versus Driver

Here are programs modeling Fat Man and Driver:

```
% deontic conflict (contrary)
:- ob(A), ob(-A).

% deontic implication (subalternation)
pe(A) :- ob(A).  % obligation implies permission
#show pe/1. #show ob/1.

% deontic equivalences
ob(A) :- -pe(-A).  -pe(-A) :- ob(A).
pe(A) :- -ob(-A).  -ob(-A) :- pe(A).
ob(-A) :- -pe(A).  -pe(A) :- ob(-A).
pe(-A) :- -ob(A).  -ob(A) :- pe(-A).

%%%%%%%%%%%%%% rules
% a ruling is required
problem :- not ruling, not problem.

% r1: it is obligatory not to kill when the killing is a murder
```

```
ob(-kill(A)) :- murder(A), not qual(r1(A)).
qual(r1(A)) :- act(A), not ob(-kill(A)).
ruling :- ob(-kill(A)).

% r2: it is permissible to kill out of necessity
pe(kill(A)) :- necessity(A), not qual(r2(A)).
qual(r2(A)) :- act(A), not pe(kill(A)).
ruling :- pe(kill(A)).

#show qual/1.

%%%%%%%%%%%%%% legal theories
% legal elements of murder
murder(A) :- malice(I), kill(A), person(P).
#show murder/1.

% legal elements of necessity: five prongs
necessity(A) :- lesser_evil(A, A1), imminent(A1), causal_nexus(A, A1),
no_alternative(A, A1), no_fault(P).
% any act is a lesser evil than an act that is a murder (Dudley)
lesser_evil(A, A1) :- act(A), murder(A1), A != A1.

#show lesser_evil/2.
#show necessity/1.

% action by omission
act(A) :- prior_duty(D), inaction(A).

%%%%%%%%%%%%%%% conflict situation: Fat Man
kill(push).  person(fatman).  malice(knowing).
```

```
imminent(let_five_die). causal_nexus(push, let_five_die).  no_alternative(push,
let_five_die).  no_fault(me).

act(push).  inaction(let_five_die).

%%%%%%%%%%%%%%% conflict situation: Driver
kill(turn).  person(one).  malice(knowing).

imminent(maintain). causal_nexus(turn, maintain). no_alternative(turn,
maintain). no_fault(me).

kill(maintain). person(five). malice(knowing).

act(turn).  inaction(maintain).
```

Here are the answer sets for Fat Man, reflecting the unavailability of the necessity defense in the case and, moreover, the failure of the action by omission theory:

```
Answer: 1

murder(push) qual(r2(push)) ob(-kill(push)) pe(-kill(push))

SATISFIABLE
```

Here are answer sets for Driver, reflecting the presence of a dilemma:

```
Answer: 1
```

```
murder(turn) murder(maintain) necessity(turn) lesser_evil(turn,maintain) ob(-
kill(maintain)) pe(-kill(maintain)) ob(-kill(turn)) pe(-kill(turn))
qual(r2(turn))

Answer: 2

murder(turn) murder(maintain) necessity(turn) lesser_evil(turn,maintain) ob(-
kill(maintain)) pe(-kill(maintain)) qual(r1(turn)) pe(kill(turn))

Answer: 3

murder(turn) murder(maintain) necessity(turn) lesser_evil(turn,maintain) ob(-
kill(maintain)) pe(-kill(maintain)) qual(r1(turn)) qual(r2(turn))

SATISFIABLE
```

Unlike Fat Man, Driver is unresolved in public law. The necessity defense has traditionally been barred in cases of homicide; however, the law is not settled, as some states have restructured the necessity defense as a defense to murder (see Cohan 2006). Wisconsin state law, for example, reduces a murder charge to manslaughter when a necessity defense is successful (Wis. Stat. Ann.: Sec. 939.47), and the Model Penal Code's commentary argues that necessity generally should be available as a defense to homicide, although this is a minority view (MPC 8.302).

In a majority of U.S. state jurisdictions, however, necessity likely will not justify the driver who turns the trolley. But at the same time, on a theory of omission of the prior duty, the driver could also be subject to a murder charge if she maintains her lane and kills five people, given the variance in moral intuitions. Hence there is a need for either a new defense of dilemma or some innovation in the law of necessity to handle true dilemma cases as I have stipulated Driver to be. Whatever legal device is employed, however, the dilemma must be resolved in public law.

CHAPTER NINE

CONCLUSION

Supreme utilitarian and deontological normative principles operate in different ways when determining whether actions are rightful as opposed to ethical, yet "machine ethics" thus far appears to have largely neglected this crucial distinction. Systems have explored how to specify and automate John Stuart Mill's Principle of Utility or Immanuel Kant's Categorical Imperative, or model other ethical principles such as the Doctrine of Double Effect, but I argue that these principles are simply the wrong normative standards to apply, or at best, incomplete. The correct normative standards for semi-autonomous machine agents are principles of *justice* such as Mill's Harm Principle, or Kant's Universal Principle of Right, as well as the positive law of a legitimate state.

Principles of justice scope supreme moral principles in order to structure the public space of freedom in public law for human beings in social interactions who otherwise cannot avoid wrong one another. Public law sets out strict, fully specifiable duties of right, and such duties take normative priority over conflicting ethical reasons for action. Duties of virtue or ethics, by contrast, are neither normatively authoritative in cases of conflict, nor rightfully enforceable against agents that violate them, and thus the purpose of automating such ethical duties is unclear. I believe that machine ethics has as a result rendered itself largely irrelevant to the actual engineering and regulatory problems that semi-autonomous machine agents pose for civil society. My primary aim in this thesis has therefore been to promote a shift in the focus of machine ethics toward creating *rightful machines.*

The answer sets programming approach I set out for defeasible deontic reasoning in the more technical chapters of the thesis was intended to serve this main aim. But the approach seems regimented and flexible enough to encode a large number of legal rules that may come into conflict when framed by the facts of some case. Enumerating the credulous extensions (answer sets) of a logic program consisting of such rules and facts rather than resolving conflicts by more or less arbitrary logical rules of priority seems to me an appropriate response to the normative demands of justice, and, moreover, an approach that might function modularly with other systems. I envision a more complete governance system for rightful machines that may include various control systems to select an answer set when necessary or appropriate, or to fill legal gaps, in varying contexts, as well as a semantic legal knowledge base. While technical aspects of the approach seems promising, however, I also think it might be worthwhile to explore other logics with the normative demands of justice in mind, particularly a revision logic such as AGM.

REFERENCES

Alchourrón, C. (1991) Conflicts of Norms and the Revision of Normative Systems. *Law and Philosophy* 10: 413-425.

Alchourrón, C. (1969) Logic of Norms and Logic of Normative Propositions. *Logique et Analyse* 12.

Alchourrón, C., Gärdenfors, P. and Makinson, D. (1985), On the logic of theory change, *Journal of Symbolic Logic*, 50(2): 510-530.

American Legal Institute (1985). *The Model Penal Code*. [MPC]

Anderson, M., & Anderson, S. L.(Eds.) (2011) *Machine ethics*. New York: Cambridge University Press.

Åqvist, L. (2008) Alchourron and Bulygin on deontic logic and the logic of norm-propositions, axiomatization, and representability results. *Logique & Analyse* 203: 225-261.

Asaro, P. (2015) The Liability Problem for Autonomous Artificial Agents. *Association for the Advancement of Artificial Intelligence*.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J. F., et al. (2018). The Moral Machine Experiment. *Nature*.

Casey, B. (2017) Amoral machines, or: How roboticists can learn to stop worrying and love the law. *Northwestern University Law Review* 111 NW. U. L. REV. 231.

Cohan, John Alan (2006) *Chapman Law Review*, Vol. 10, Issue 1, Fall pp. 119-186

Foot, P. (1967) The problem of abortion and the doctrine of double effect. *Oxford Review* 5, 5–15.

Ganascia, J. (2007) Modelling ethical rules of lying with Answer Set Programming. *Ethics and Information Technology* 9:39–47.

Gelfond, M., and Lifschitz, V. (1988) The stable model semantics for logic programming. In: Kowalski, R., Bowen, K.A. (eds.) *5th Intl. Logic Programming Conf.*, MIT Press, Cambridge.

Gelfond, M. (2008) Chapter 7: Answer Sets. *Foundations of Artificial Intelligence*, 3, 285–316.

Goble, L. (2005) A logic for deontic dilemmas. *Journal of Applied Logic* 3 461–483.

Girle, R. (2017) *Modal Logics and Philosophy*, 2d ed. Montreal: MQUP.

Guarini, M. (2012) Conative Dimensions of Machine Ethics: A Defense of Duty. *IEEE Transactions on Affective Computing*, vol 3, no. 4.

Hart, H. (1973) Rawls on Liberty and Its Priority. *The University of Chicago Law Review* 40: 534-55.

Hohfeld, W. (1919) *Fundamental Legal Conceptions as Applied in Judicial Reasoning*, ed. Walter Wheeler Cook. New Haven, CT: Yale University Press.

Horty, J. (2001) *Agency and Deontic Logic*. Oxford University Press.

Kant, I. (1992) In P. Guyer and A. Wood (Eds.), *The Cambridge Edition of the Works of Immanuel Kant.* Cambridge: Cambridge University Press. All references to Kant's work are from the Cambridge edition unless otherwise noted. Citations are according to standard Academy pagination.

- *The Doctrine of Right*, Part One of *The Metaphysics of Morals,* trans. M. Gregor [DR]
- *The Doctrine of Virtue*, Part Two of *The Metaphysics of Morals,* trans. M. Gregor [DV]
- *Groundwork of the Metaphysics of Morals*, trans. M. Gregor [GM]
- *On the Common Saying: 'That May Be Correct in Theory but It Is of No Use in Practice',* trans. M. Gregor. [T]
- *On a Supposed Right to Lie out of Philanthropy.* trans M. Gregor [SR]
- *Toward Perpetual Peace*, trans M. Gregor [PP]

Liao, S.,Wiegmann, A., Alexander, J., and G. Vong. (2012) Putting the trolley in order: Experimental philosophy and the loop case. *Philosophical Psychology* Vol. 25, No. 5, October 2012, 661–671.

Maier, F. and Nute, D. (2010) Well-founded semantics for defeasible logic *Synthese* 176:243-274.

Maranhao, J. (2006) Why was Alcourron afraid of snakes? *Analisis Filosofico, XXVI* Nº 1 - ISSN 0326-1301 (mayo 2006) 62-92.

Mikhail, J.: (2007) Universal moral grammar: Theory, evidence, and the future. *Trends in Cognitive Sciences* 11(4), 143–152.

Mill, J.S. (1977) *On Liberty* (1859), n *Essays on Politics and Society.* In J.M. Robson, ed., *The Collected Works of John Stuart Mill*, vol. 18.

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems,* 21(4), 18–21.

Nute, D. (1993). Defeasible Prolog. *Technical report: American Association for Artificial Intelligence FS*, 105.

O'Neill, O. (2011) *Constructing Authorities.* Cambridge: Cambridge University Press.

Pereira, L. and Saptawijaya, A. (2011) Modeling Morality with Prospective Logic. In Anderson, M., & Anderson, S. L. (eds.) *Machine ethics*. New York: Cambridge University Press.

Petrinovich, L., & O'Neill, P. (1996). Influence of wording and framing effects on moral intuitions. *Ethology and Sociobiology*, 17, 145–171.

Potassco, the Potsdam Answer Set Solving Collection. University of Potsdam.

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

Rawls, J. (2001) *Justice as Fairness: A Restatement* (Cambridge, MA: Harvard University Press.

Rawls, J. (1993) *Political Liberalism.* New York: Columbia University Press.

Reiter. R. (1980) A Logic for Default Reasoning. *Artificial Intelligence*, 13: 81–132.

Timmermann, J. (2013) Kantian Dilemmas? Moral Conflict in Kant's Ethical Theory. *DeGruyter* 95(1): 36–64.

Thomson, J. (1976) Killing, Letting Die, and the Trolley Problem. *The Monist* 59:204–17.

Thomson, J. (1985) The Trolley Problem. *The Yale Law Journal* 94:1395–415.

Thomson, J. (2008) Turning the Trolley. *Philosophy & Public Affairs* 36, no. 4

Tonkens. R. (2009) A Challenge for Machine Ethics. *Minds & Machines* 19:421–438

Wood, A. (2011) Kant and the Right to Lie. *Eidos* 15:96-117.