Rightful Machines and Dilemmas

[discipline: philosophy]

Abstract

In this paper I set out a new Kantian approach to resolving dilemmas such as the trolley problem for semi-autonomous machine agents such as self-driving cars. First, I argue that efforts to build explicitly moral machine agents should focus on what Kant refers to as duties of right, or justice, rather than on duties of virtue, or ethics. Then, I show how the shift from ethics to a standard of justice illuminates the resolution of the conflict of obligation in what is known as the "trolley problem" for rightful machine agents. An action is rightful, Kant says, when it 'can coexist with the freedom of every other under a universal law;' hence Kant specifies duties of right by reference to their consistency within a system of equal rights of freedom. I interpret this demand for consistency in the system of legal norms as a normative requirement of justice. Hence when conflicts between strict legal obligations arise, we should not conceive them as cases where we must violate one or another obligation but, instead, as cases where we must *revise* or qualify obligations in order to meet the normative requirement of consistency.

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 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 paper, 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 universal law;" therefore, the rightfulness of any act is specified explicitly in terms of its consistency within a system of equal rights of freedom (DR: 6: 230).

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 a system of equal freedom for human beings in social interactions.

Finally, I will consider how a deontic logic suitable for governing explicitly rightful machines might meet the normative requirement of 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, and that a logic of belief revision may be preferred.

Rightful Machines

In the *Doctrine of Right*, 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 articulates the same idea in terms of 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 man by virtue of his humanity (DR: 6:237).

Hence while *freedom* is independence from being constrained by others, the *right of freedom* is that freedom limited by everyone else's equal rights of freedom under universal law. An unrestricted right of freedom would fail to secure everyone's rights of freedom; therefore Kant immediate 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 someone who infringes it. (DR: 6:231).

Self-defense and the defense of necessity

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. That is, when killing one's assailant in self-defense, there is no violation of the general obligation not to kill because the killing corrects the wrong the assailant is attempting to commit. Kant does suggest that one might nevertheless have an *ethical* duty not to kill one's assailant. This is possible because unlike legal duties, ethical duties are not specified by reference to their consistency within a system. 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" of action. Such conflicting grounds do not exist in a legal context, however, since legal obligations are completely specifiable in terms of their outward aspects.

In a case of the defense of "necessity," by contrast, in which one kills an innocent because that is the only way to preserve one's own life, one does act wrongfully, Kant argues. But 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, since even a punishment of death would not effectively deter the crime (6:235-6). Kant thus regards the defense of necessity, to the extent it is thought a *legal* defense, as premised on a confusion and so would reject any version of the "choice of evils" or general necessity defense sometimes raised in U.S. criminal law (see, e.g., MPC 3.02). Of course neither of these defenses could be raised by machine agents that are not real persons.

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 an ethical duty to keep one's promises, one has a legal duty to keep only those promises that meet the outward, public criteria that define 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 contractual obligation just the same. Similarly, I meet my legal obligations to avoid crimes such as theft and murder even if I avoid them solely because I fear punishment. The corresponding ethical duties, by contrast, require me to avoid such criminal acts because they are wrong.

The rightful enforceability and precise specifiability of duties of right have important implications for builders of explicitly normative machine agents. First, the precision necessary to specify duties of right should make such duties much easier to capture in governance systems. Second, rightful machines sidestep problems 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). But if, on the other hand, advanced machines are capable of autonomous ethical agency, then installing a coercive explicitly ethical governance system would violate the *machine's* right of freedom (Tonken). 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.

Finally, and perhaps most importantly, since ethical duties are not rightfully enforceable against those who violate them, explicitly ethical machine agents 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.

Solving the Trolley Problem

The original trolley "problem"

Consider the original ("Driver") version of the "trolley problem" (Foot 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 (90%) say they would turn the trolley (Mikhail).

Now contrast Driver with the following variation ("Fat Man") (Thomson 1976): 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 who leaning over the side of the 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 about (90%) say they would not, in a reverse mirror image of the intuitions in Driver (Mikhail).

The trolley "problem," 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). 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). 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 no matter what you do, it is better to turn the trolley and kill fewer people (Foot 5). By contrast, in Fat Man, you are faced with a conflict between a negative duty not to kill one (the fat man) and a *positive* duty to render aid to the five. In such cases, the negative duty is more important than the positive one, Foot claims (Foot 5). One therefore should kill the one to spare the five in Driver but avoid doing so in Fat Man.

The priority of right

Foot's analysis is roughly correct but incomplete. To complete the analysis Foot needs to provide some account of why "negative" duties to avoid acts such as killing others take normative priority over "positive" duties to perform acts such as aiding others (Thomson 2008). I argue that duties not to kill in the trolley problem take priority not because they are negative duties but because they are strict *duties of right*, whereas conflicting positive duties in cases like Fat Man are *ethical* duties. According to Kant, duties of right take priority over ethical duties, in the special sense that legal but not ethical obligations are rightfully enforceable. 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 that are not also legally enforceable are positive. But the relevant distinction here is between duties or right and those of ethics.

Now, while it is characteristic of Kantian deontology that one's duties constrain the goals one may permissibly pursue, this priority of duties over goals is not what distinguishes Driver from Fat Man. Both negative and positive ethical duties 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, while the latter are usually positive. But it seems clear that wide duties might sometimes ripen into constraining ethical obligations that should take priority over reasons for action that strict ethical duties might generate. 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 had made to meet someone for lunch.

It is the *priority of right* over ethics, not the deontological priority of duties over goals, that I argue explains why killing is worse than letting die in Fat Man. While Kant does not clearly explain why duties of right should take

priority over ethical duties, this is likely because he did not think that conflicts between duties were even possible (Timmerman). All duties of right are perfect duties of strict obligation, whereas ethical duties are often imperfect and of wide obligation. Hence if legal obligations always apply in every situation, and Kant thinks that conflicts between obligations are impossible, then the question of the priority of a strict duty of right over a wide ethical duty for Kant cannot even arise. But if, as many believe (and as Driver and Fat Man seem to me to illustrate), conflicts between duties are indeed possible, then the priority of duties of right over ethical duties requires explanation. I suggest the following understanding of this priority.

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. Hence one aspect of the priority of right amounts to just the recognition that if others are free, then one cannot coercively impose one's own ethical preferences upon them. One can only force others to conform to those (legal) duties necessary to secure an equal system of freedom for everyone (i.e. justice). In simplest terms it is wrong of me to force you to do what I think you ought to do unless doing so is required for a civil society.

I thus cannot push the fat man off the footbridge in order to save the most lives, since even if the fat man had an ethical duty to jump and sacrifice his life to save the five, I could not force him to act ethically. Perhaps I could push him if I first obtained his consent, or if I could ascertain his will. But he fat man has a right of freedom to determine for himself what should be done, limited only by the equal rights of freedom of everyone else in a system.

Much more could be said concerning the priority of right over ethics, but my aim here is only to give a sense of why Kant endorses the idea. There is no question that he does, as does almost every modern philosopher in some form. For example, the utilitarian John Stuart Mill's principle of justice, the "Harm Principle," states a version of the priority of right, and John Rawls defends the priority of right explicitly and at length (Rawls).

The real trolley "problem"

Distinguishing right from ethics and observing the priority of right thus solves the original trolley "problem." One has a duty of right not to kill the fat man that takes priority over one's ethical duty to render aid to the five, whereas in Driver, there is a conflict between duties of right not to kill the one and not to kill each of the five. But the more interesting and important trolley "problem" for my purposes is the conflict between duties of right in Driver. 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 one should turn the trolley in Driver. 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 clear why justice should allow 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.

Kant claims, moreover, that conflicts between one's strict legal obligations such as in Driver cannot occur. One cannot possibly have a duty to perform an action that one is simultaneously obligated to avoid, Kant argues; hence "...a collision of duties and obligations is not even conceivable (*obligationes non colliduntur*)" (MM: 7:224). What is known as the standard system of deontic logic (SDL) reflects Kant's view, since it is a theorem of SDL that "Op --> ~O~p" (i.e., "~(Op & O~p)"), where we take "O" as a monadic operator for an obligation one has and "p" as an action one performs. That is, if one has an obligation to perform an act, then one cannot at the same time have an obligation not to perform that act. Admitting such a conflict in SDL would imply that one has an obligation to perform and not to perform the very same act (O(p&~p)), which is not even conceivable. Kant's view and SDL thus appear to imply that there is no conflict between one's obligations in the trolley problem, and that one's obligation is clear.

I argue that the best way to render Kant's claims about the systematic consistency of one's strict juridical duties is to think of it as a *normative requirement* of justice, rather than a necessary truth about any system of norms we might call legal. Whether conflicts of legal obligation are possible or not, it would be wrongful to enforce contradictory legal obligations, as then force would be applied arbitrarily, since one cannot possibly consent to such force. But note that since ethical obligations that are not also legal obligations are not rightfully enforceable, this normative requirement would not apply if conflicts between ethical obligations were to occur. Hence the analysis of legal as opposed to ethical conflicts must be quite different.

I can now offer an approach to the solution of the trolley problem dilemma in Driver. First, I argue that the conflicting obligations at issue are strict legal obligations (not to wrong another by intentionally killing her, even to save many others), although there also appears to be considerable conflict between one's ethical grounds of action, as well. I further stipulate that the problem is indeed a dilemma in which we are subject to contradictory juridical obligations (Op & O~p). 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, or some controlling positive law, that would eliminate one of the obligations.

I then appeal to the normative requirement that strict legal obligations must be made consistent in the prescriptive system of legal norms. 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 note that this requirement of consistency in the system of legal norms is a second-order principle of justice, not a property of the system. Enforcement of either obligation if taken by itself is rightful in principle at the level of the prescriptive system of legal norms. At this prescriptive level, consistency is a constraining property of the system; hence a lack of consistency with other legal norms in the system cannot be the reason that a norm is not rightfully enforceable. Contradictory norms are simply simply inadmissible, 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. At level of of the descriptive system, however, obligations conflicting in a dilemma are unenforceable and must be excised or revised.

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, since either legal obligation would be rightfully enforceable in principle in the absence of the inconsistency. What matters is that it is resolved; and moreover, the resolution may vary by jurisdiction, so long as there is due process. And in fact this is precisely how (U.S.) law handles many such cases: in some states, contributory negligence completely bars recovery by injured plaintiffs in accidents, while in other states, fault might play no or a very limited role. Yet in each state, the law is rightfully enforceable.

Suppose five people are attempting to cross an interstate highway (which is generally illegal), and the 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, the car will be programmed to swerve and kill the motorcyclist, because in such a jurisdiction, liability for the deaths will be assigned strictly without regard to fault, and one death is less costly to compensate than five from the point of view of the manufacturer's liability. 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 manufacturer 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). Yet no one argues that the application of either state's rule is illegitimate. They are both rightfully enforceable within their respective jurisdictions. Note that principles of ethics are likely to play no role in determining the behavior of self-driving cars.

In the absence of any controlling positive law or regulation, or governing judicial precedent, however, neither legal obligation can be justly enforced in the trolley problem, and the practical result, speaking strictly legally, is that one may resolve the conflict however one wishes.

Now, if the problem is framed as one where we are forced to choose between ethical duties to avoid harming one as opposed 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, and even Kantians would disagree on its proper ethical resolution. By contrast, whatever resolution a public authority makes of the trolley problem is rightfully enforceable and so decides the issue.

Normative Consistency and Deontic Logics

One might think that the standard system of deontic logic would best reflect the normative consistency requirement, since no-conflicts (~(Op & O~p)) is a theorem in SDL. But there seems to me no reason to think that even a rational public authority might not inadvertently create legal obligations that contradict in situations that authority did not foresee. For example, suppose a municipal authority passes a traffic law that requires stopping at stop signs and another that forbids stopping in front of military bases. 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 situa-

tion (Navarro and Rodriguez, 179). The possibility of such conflicts seems a mundane descriptive fact about our 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. Formal systems should be able to represent the conflict of obligations in such a case descriptively while maintaining some mechanism to meet the normative demand for consistency. The logic should not deny the descriptive possibility of such conflicts, as SDL does.

At the other extreme are deontic logics that accept classical contradictions between norms and attempt to draw reasonable inferences despite them. Semi-classical logics and some paraconsistent logics abandon classical semantics with its two truth values (true, false) to replace it with a semantics of many values (e.g., null, just true, just false, and both true and false). Such systems are often regarded as too weak to be useful, but the problem with them in the present context is that their very purpose is to tolerate contradictions. Such logics accept inconsistencies not only descriptively but also normatively. Efforts to strategically weaken axioms or rules of inference of the standard system in order to avoid the deontic explosion offend the normative demand for consistency in the same way (Goble). 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 resolved.

Non-monotonic reasoning systems with a classical base appear to meet this minimum requirement, though perhaps not as explicitly as they might. NMRs are able to admit contradictions because they reject monotonicity, that is, "if $A|\sim p$ and $A \subseteq B$ then $B|\sim p$". What this means is that some inferences might no longer be drawn when new premises are introduced; for example, one might introduce a new fact that directly contradicts some fact upon which an inference depended, so defeating that inference. They therefore avoid deontic explosion of inferences from the contradiction. NMRs on a classical base meet the normative consistency requirement because, semantically, they require an explicit preference or choice relation between possible worlds that are (classically) maximally consistent, in order to continue to draw defeasible inferences. Literal contradictions of facts in a NMR will generate the same deontic explosion of inference that occurs when contradictions are admitted into the standard system of deontic logic. The answer sets semantics, for example, reflects this, returning an answer set containing all literals when factual contradictions are unavoidable (Gelfond).

Carlos Alchourron rejects defeasible deontic logics because he argues such systems obscure the distinction between descriptive and prescriptive activity in the law (Maranhao). 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 would reject positivism. Law that conforms to the UPR (and perhaps other implicit requirements, as well) and is enacted in accordance with civil institutions, etc., is valid law for Kant because of its form, and to some degree because of its substantive content (not violating equality, freedom) and procedural history (arising as out of just institutions). All of these requirements flow from the distinctively legal principle of justice, the UPR, and ultimately, the categorical imperative.

Yet Kant would also recognize that a number of diverse consistent bodies of positive law are possible and are all 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 law toward the strongest and most coherent system realizing equal freedom. Logics of belief revision such as 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). 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 goal of system 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 revision operator (Aqvist), a logic such as AGM may better reflect Kant's normatively consistent system of equal freedom under universal laws constructed by a civil community.

Conclusion

I have argued that efforts to build explicitly normative machine agents should focus on duties of right rather than ethics. Rightful machines by definition avoid acting in ways that paternalistically interfere with rights of freedom,

whereas ethical machines may not. Shifting from ethics to a standard of right, moreover, provides a new approach to the law and logic of deontic dilemmas such as the trolley problem for semi-autonomous machine agents like self-driving cars.

References

Alchourrón, C. (1991). Conflicts of norms and the revision of normative systems. *Law and Philosophy* 10:413-425,.

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

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

Anderson, M and Anderson, S. L.. (2011). *Machine Ethics* (1st ed.). Cambridge University Press, New York, NY.

Anderson, M., & Anderson, S. L. (2006). Machine ethics. *IEEE Intelligent Systems*, 21(4),.

Aqvist, 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 (*AAAI)..

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

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

Ulrich Furbach, Claudia Schon, and Frider Stolzenburg. Automated Reasoning in Deontic Logic. arXiv: 1411.4823v1. 2014.

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

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

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

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

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

I. Kant in *The Cambridge Edition of the Works of Immanuel Kant*, trans. Mary Gregor, ed. Paul Guyer and Allen Wood (Cambridge: Cambridge University Press, 1992). All references to Kant's work are from the Cambridge Edition unless otherwise noted; citations are made according to Academy pagination.

I.. Kant, "The Doctrine of Right" (DR)

I.. Kant, "The Doctrine of Virtue" (DV)

I. Kant, "Groundwork of the Metaphysics of Morals" (GM)

I. Kant, "The Metaphysics of Morals" (MM)

I. Kant, 'On the Common Saying: 'That May Be Correct in Theory but It Is of No Use in Practice' (T).

I. Kant. "On a Supposed Right to Lie From Philanthropy" (SR)

I. Kant, 'Toward Perpetual Peace' (PP)

S. Matthew Liao, Alex Wiegmann, Joshua Alexander, and Gerard Vong. Putting the trolley in order: Experimental philosophy and the loop case. Philosophical Psychology Vol. 25, No. 5, October 2012, 661–671

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

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

\Model Penal Code (MPC).

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

Navarro, P. and Rodriguez, J. *Deontic Loigc and Legal Systems*. Cambridge 2014.

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

Lu´ıs Moniz Pereira and Ari Saptawijaya. Modelling Morality with Prospective Logic. J. Neves, M. Santos, and J. Machado (Eds.): EPIA 2007, LNAI 4874, pp. 99–111, 2007. Springer-Verlag Berlin Heidelberg 2007

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

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

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

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

Timmermann, J. (2013) Kantian Dilemmas? Moral Conflict in Kant's Ethical Theory. AGPh ; 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. Wiley Periodicals, Inc. *Philosophy & Public Affairs* 36, no. 4

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

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