Ava Thomas Wright

8 Rightful Machines

Abstract: In this paper, I set out a new Kantian approach to resolving conflicts between moral obligations for highly autonomous machine agents. First, I argue that efforts to build explicitly moral autonomous machine agents should focus on what Kant refers to as *duties of right*, which are duties that everyone could accept, rather than on duties of virtue (or "ethics"), which are subject to dispute in particular cases. "Moral" machines must first be *rightful* machines, I argue. I then show how this shift in focus from ethics to a standard of public right resolves the conflicts in what is known as the "trolley problem" for autonomous machine agents. Finally, I consider how a deontic logic suitable for capturing duties of right might meet Kant's requirement that rightfully enforceable obligations be consistent in a system of equal freedom under universal law.

1 Introduction: (Im)moral machines

In a massive experiment conducted online (the "Moral Machine Experiment"), 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 over those of lower status, and the fit over the overweight, with some variations in preferences correlated with subjects' cultural backgrounds.[1] But while such results may be interesting, they seem to me at best irrelevant to the question of what a self-driving car faced with such a dilemma should do. Ethical preferences to spare more rather than fewer lives, or to spare humans over animals, are for the most part morally banal, while

1 These results are ceteris paribus preferences derived by aggregating individual decisions made by subjects across many different variations of the basic swerve-or-stay-on-course accident scenario (see Awad, et al. 2018, p. 60). They should not be understood to indicate absolute or overriding ethical preferences.

Ava Thomas Wright, California Polytechnic State University

https://doi.org/10.1515/9783110706611-008

ethical preferences to spare those of higher over lower status, or those of one gender or body type over another, are morally problematic. The latter preferences raise the strong moral intuition that choices guided by them would fail to respect the moral *equality* of persons. Self-driving cars programmed to enact such preferences would therefore be immoral machines.

According to Immanuel Kant, there are two kinds of moral duties: 1) *duties of right* ("legal" duties), which are duties that are rightfully enforceable by others, and 2) *duties of virtue* ("ethical" duties), which are not rightfully enforceable by others because their application in particular cases is subject to dispute.[2] Kant accordingly divides the *Metaphysics of Morals* into the *Doctrine of Right* and the *Doctrine of Virtue* (see TL, AA 06: 379). In this paper, I argue that efforts to build explicitly moral autonomous machine agents should focus on duties of right, rather than on duties of virtue, when resolving conflicts between obligations or rights. While dilemmas such as those in the (in)famous "trolley problem" – which inspired the experiment above – have received enormous attention in "machine ethics," there will likely never be a consensus as to their correct resolution.[3] What matters morally in such controversial cases 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 principles of public right and law. The point is not merely that conflict cases like those in the trolley problem likely will, as a practical matter, be resolved by the law rather than by ethical principles (Casey 2017). The point is, rather, that the resolution of such disputes between equals morally should be determined by principles of right and public law before controversial ethical principles may be applied. A "moral machine" must first be a *rightful machine*, I argue.

This paper has three main sections. In the first two sections, I set out some basic elements of Kant's theory of justice and then apply them to resolve the conflicts between duties in the trolley problem. An action is right, Kant says,

2 Following Kant, I will refer to duties of right as "legal," "rightful," or also "juridical," and reserve the term "ethical" to refer to duties of virtue (see MS, AA 06:219). I will use the term "moral" to refer broadly to any duty or power, whether legal or ethical or both. Kant occasionally appears to disregard his own distinction and use the term "ethics" to refer to morality generally, but I avoid this usage. For some critical discussion of Kant's distinctions, see O'Neill 2016, pp. 114–117.

3 The field of "machine ethics" is primarily concerned with building autonomous machine agents that can take moral considerations into account in their decision-making. Machine "ethics" therefore should not be understood as limited to what Kant would refer to as "ethics" (i.e., virtue).

when it "can coexist with the freedom of every other under universal law;" therefore, the rightfulness of an action is specified explicitly in terms of its consistency within a system of equal rights of freedom under universal law (RL, AA 06:230). I interpret this consistency not descriptively but normatively as a moral requirement that public right imposes upon any system of rightfully enforceable duties and rights. Without such consistency, the enforcement of either conflicting obligation in a disputed case would be arbitrary, and arbitrary enforcement is tantamount to coercion. Hence when dilemmas between duties of right 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 duties of right but, instead, as cases where a legitimate public authority must precisely *specify* our duties and rights in order to meet the 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 under universal law for human beings inevitably engaged in social interactions.

Finally, in the third section, I consider how a deontic logic suitable for governing explicitly rightful machines might meet the normative requirement of consistency in the system of equal rights of freedom under universal law. I suggest that a non-monotonic deontic logic can meet the consistency requirement, though with certain reservations, and that a logic of belief revision may be preferred.

2 Rightful Machines

2.1 Kantian Right and the Innate Right of Freedom

In the *Doctrine of Right*, Kant defines the "Universal Principle of Right" 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. (RL, AA 06:230)

Kant thus defines the legal permissibility (rightfulness) of any action in terms of its systematic consistency with everyone's equal freedom under universal law. If the act is consistent with everyone's equal right of freedom, then 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, pp. 35–50).[4]

Kant reiterates justice as systematic freedom under universal law when defining the innate 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 or her] humanity. (RL, AA 06:237)

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

2.2 The Priority of Right

According to Kant, reason alone cannot specify a priori what our rights and duties, and powers and liabilities, with respect to each other are in particular cases (RL, AA 06:312). Since everyone is innately *equal*, each person has her "own [natural] right to do *what seems right and good to [her]* and not to be dependent on another's opinion about this," Kant says (RL, AA 06:312). No one individual or group has the innate moral authority to unilaterally define everyone's rights and duties with respect to others (i.e., legislate them), or to enforce them (i.e., execute them), or to resolve disputes (i.e., determine them) in particular cases. Intractable disputes over our rights and powers with respect to each other in particular cases are thus inevitable in a "state of nature" lacking public institutions to resolve them. While the state of nature is not necessarily a state of injustice, Kant says, "it would be a state *devoid of justice (status justitia vacuus),* in which when rights are in *dispute (ius controversum*), there would be no judge competent to render a verdict having rightful force" (RL, AA 06:312). Hence even if everyone were committed to acting perfectly ethically, according to Kant, rightful relations with others are impossible in a state of nature (RL, AA 06:312).

---

4 For example, if legal *duty* is taken as basic, then: person x has a *permission* to perform action P iff x has no duty not to P with respect to y; x has a (claim-) *right* that P iff person y has a duty to perform P for x; and x has what Hohfeld calls a "*no-right*" that P with respect to y iff y has no duty to not-P with respect to x.

What is needed, Kant says, 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. (RL, AA 06:311)

Kant refers to this system of public laws and institutions as "public right," and a society existing under such a system as one existing in a "rightful" or "civil" condition. The coercive enforcement of public law is rightful under such a system, Kant says, because

> when someone makes arrangements about another, it is always possible for him to do the other wrong; but he can never do wrong in what he decides upon with regard to himself (for *volenti non fit iniuria*). Therefore only the concurring and united will of all, insofar as each decides the same thing for all and all for each, and so only the general united will of the people, can be legislative. (RL, AA 06:313–14)

It is only by constituting a general or united will to authoritatively define, enforce, and adjudicate our rights and duties with respect to each other that we can avoid wronging one another in cases of dispute over our rights, Kant argues.

Hence determinations made in the system of public laws regarding what our rights or duties are take moral *priority* over individual ethical judgments in cases where those rights or duties are in dispute. To reject public authority and use one's own private judgment in such disputed cases is to act wrongfully, indeed, to do "wrong in the highest degree," Kant says (RL, AA 06:308n). Resolving such disputes in order to enable rightful relations with others is the very purpose of the system of public laws.

2.3 Duties of Rightful Machines

Duties of right concern only the public, outward aspects of one's actions and, according to Kant, are thus completely specifiable without reference to the agent's motive or "maxim" of the end of action (TL, AA 06:390). For example, while one has a 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 obligation just the same (see RL, AA 06:230). 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.[5]

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 conformity with those duties somewhat easier to achieve in a machine agent, since determining whether duties of right apply and what action they require should demand considerably less moral judgment in particular cases. It is much more difficult to determine what the duty of virtue to help others requires in particular cases than to determine what a positive legal duty to render assistance at the scene of an automobile accident requires (see, e.g., Minn Sec. 604A.01). Second, shifting the focus of machine ethics to conformity with duties of right sidesteps objections related to the machine agent's potential capacity for freedom. If a machine cannot act according to a principle that it freely chooses, then the machine cannot act ethically and can at best produce only a simulacrum of ethical action (Guarini 2012). On the other hand, if advanced machines of the future do become capable of genuine ethical agency (i.e., true Kantian "autonomy"), then installing a coercive, explicitly ethical control system would violate the *machine's* right of freedom (see Tonkens 2009). By contrast, duties of right require no particular subjective incentive for action; hence, mere conformity with the outward aspects of such duties is sufficient to act rightfully. And since duties of right are rightfully enforceable, a coercive control system might not violate even a truly "autonomous" machine's rights.

Finally, and perhaps most importantly, explicitly *ethical* machines that acted on preferences such as those collected in the Moral Machine Experiment might often violate rights of equality and freedom, and it is not difficult to imagine dystopias where such 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' equal rights of freedom and avoid paternalistic ethical meddling.

Self-driving cars and other machine agents programmed to act in accordance with popular ethical preferences would be immoral machines and seem to me to pose a threat to civil society. The goal of machine ethics should be *rightful machines*.

---

5 Kant also holds that one has a general ethical duty to obey legitimate law, which implies that all legal duties are therefore also indirectly ethical duties (see TL, AA 06:390–91). This indirect ethical duty to obey the law out of the incentive of duty is not my concern here, however, and the priority of public right does not depend upon it. For a perspicuous account of the relation between right and ethics in Kant's moral philosophy, see Guyer 2016.

3 Solving the Trolley Problem

3.1 The Original Trolley Problem: Driver versus Footbridge

Consider one ("Driver") variation of the "trolley problem" (Foot 1967, p. 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 (see, e.g., Mikhail 2007).

Now contrast Driver with the following variation ("Footbridge") (Thomson 1976, pp. 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 large man leaning over the footbridge railing. You could stop the trolley and save five people if you pushed the large 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 large 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 Footbridge, since most people are willing to kill one to spare five in the former but not in the latter case (Foot 1967, p. 3). Foot suggests that the answer is that "negative" duties such as to avoid injuring or killing others are qualitatively more important than "positive" duties such as to render aid to them (Foot 1967, pp. 5–6). In Driver, you are faced with an unavoidable conflict between negative duties not to kill five and not to kill one, Foot says, and since you must violate a negative duty not to kill someone no matter what you do, it is only rational to turn the trolley so as to inflict the least injury (Foot 1967, p. 5). By contrast, in cases like Footbridge, you are faced with a conflict between a negative duty not to kill one (the large man) and a positive duty to protect the five from harm, Foot says, and in such cases, the negative duty takes priority over the positive duty (Foot 1967, p. 6). One therefore should kill one to spare five in Driver but avoid doing so in Footbridge, according to Foot.

3.2 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 (Thomson 2008, p. 372). I argue that duties not to kill in the trolley problem take such normative priority not because they are negative duties but because they are *duties of right*, whereas conflicting positive duties to aid others in cases like Footbridge are *ethical* duties. Duties of right determined authoritatively in public law take normative priority over conflicting ethical reasons for action. 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 virtue.

Perhaps you are one of the 10% who think it might not be unethical for you to push the large man because that minimizes lives lost. But the large man's right to life in such a case of conflict 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 individual ethical judgment for it in the case, even if you disagree. The large man's right to his life includes at least the right not to be coerced to die in order to aid others. Indeed, this much of his right to life likely must be present in any legitimate system of equal freedom under public laws to which everyone could possibly consent (see ZeF, AA 08:349–50). Hence the large man's right to life in such a case has already been authoritatively determined in public law, and you therefore have a moral duty to respect it, whatever your ethical preference in the case may be. To do otherwise is to act lawlessly, Kant says, to commit wrong "in the highest degree" (RL, AA 06: 308n). This is the priority of right.

In the Driver variation, by contrast, there is a conflict between a duty of right not to kill the one and duties of right not to kill each of the five. Some may object that by not turning the trolley, the driver avoids taking action and so avoids violating any legal duty of right not to kill the five. But this objection fails because as the driver of the trolley you are subject to a prior legal duty to drive the trolley safely, and failing to fulfill this duty therefore constitutes an action by omission. To see this prior legal duty more clearly, compare an analogous case where you are driving a car: if there are five people stranded in the lane ahead (let us assume, through no fault of their own), and you could safely change lanes to avoid killing them, then choosing to nevertheless maintain your lane and kill them would violate a prior legal duty to drive the car safely

(see Thomson, 2008, p. 369). There is, therefore, a conflict between (legal) duties of right in Driver. In cases of conflict between legal duties, the priority of right does not control, and this is what distinguishes Driver from Footbridge. Since the resolution of the conflict between legal duties in Driver is unclear, it seems only rational to minimize rights violations as a fallback ethical principle in the case.

Distinguishing right from ethics and observing the priority of right thus solves Foot's original trolley "problem." In Footbridge, one has a duty of right determined authoritatively in public law not to kill the large man that therefore takes priority over one's ethical duty to save the five from harm. In Driver, by contrast, there is a conflict between duties of right that the priority of right cannot resolve and so rational moral intuition falls back on minimizing harm. Prevailing intuitions to kill one to spare five in Driver but not to do so in Footbridge are thus both rational. This solves Foot's trolley problem.[6]

3.3 The Real Trolley "Problem:" 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 you should turn the trolley in Driver. But since principles of justice characteristically 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 may complain of being wronged in either case.

I propose the following approach to understanding the dilemma between duties of right in Driver. First, let us stipulate that the conflict is indeed a dilemma in which one is subject to contradictory strict legal obligations not to wrong another by intentionally killing her (i.e., 'OBa $\wedge$ OB~a', where 'OB' is obligation and 'a' is an action). That is, there is no other legally relevant factor,

6 Another trolley "problem" that has attracted some attention is the Bystander variation, which is like Driver except that instead of being the driver of the trolley, you are a bystander with access to a switch that can turn the trolley. This variation is a bad thought experiment because, unlike the Driver or Footbridge variations, the Bystander variation is subject to framing and ordering effects (see, e.g., Liao et al. 2012). These effects likely arise because intuitions about what one should do in Bystander will shift depending upon whether subjects take the control the bystander exercises over the trolley to be sufficient to make an analogy with the control the driver exercises in Driver, or not. Hence experimental results obtained by polling in the Bystander variation will be equivocal.

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. Now recall Kant's requirement that the prescriptive system of public laws specifying strict legal obligations must be consistent. What does this normative requirement of consistency imply in such a dilemma case?

The first implication is that *neither obligation in the dilemma can be rightfully enforced*. It is not possible to consent to be subject to the enforcement of contradictory narrow 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. Normative consistency is a constraining property of the system of enforceable public laws; hence a lack of consistency with other legal duties in the system cannot be the reason that a duty is not rightfully enforceable. A legal duty that contradicts another is simply inadmissible into the prescriptive system of legal duties, and the implication of a dilemma in the system is, rather, that the enforcement of either obligation is both rightful and wrongful, i.e., that its rightfulness *cannot be determined*.

The second implication of the normative consistency requirement is that public right requires that *the dilemma must be resolved* (i.e., either by legislative action or judicial verdict). It does not matter how it is resolved, so long as the procedural and substantive requirements of the universal principle of right are met when resolving it. What matters is that the conflict is resolved; and moreover, its resolution may vary by jurisdiction. 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 will completely bar 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.

From the point of view of justice, then, dilemmas like that in Driver are little different from other conflicts between obligations. The main difference appears to be that in the dilemma case we assume that there is no 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 must resolve other cases of conflict between moral equals. I do not mean to imply that civil institutions are authorized to resolve such conflicts irrationally or arbitrarily; rationality will still impose bounds upon acceptable resolutions and their public justifications. It is just that in the dilemma case there will be no decisive reason to resolve the conflict one way or the other.

4 Normative Consistency and Deontic Logic

4.1 Standard Deontic Logic and Non-Monotonic Reasoning Systems

One might think that the standard system of deontic logic would best reflect Kant's normative consistency requirement, since no-conflicts (i.e., '~(OBa & OB~a)') is a theorem of Standard Deontic Logic (SDL). But there seems 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 situation (Navarro/Rodriguez 2014, p. 179). The possibility of such conflicts seems a mundane fact about any actual 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 between obligations in such a case *descriptively* while maintaining some mechanism to resolve the conflict at the *prescriptive* level. The logic should not make it impossible to describe such conflicts, as SDL does. Efforts to strategically weaken the axioms or rules of inference of SDL in order to admit contradictions without generating a deontic explosion of inferences appear to merely quarantine rather than resolve contradictions, since the logic provides no mechanism for resolving the contradiction (see, e.g., Goble 2005). They therefore fail to meet the demand that contradictions be resolved at the level of prescriptive obligations.

At the other extreme from SDL are deontic logics that accept contradictions between norms and then attempt to draw reasonable inferences despite them. Semi-classical logics and some paraconsistent logics abandon classical semantics with its two truth values (true, false) and 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 very useful, but the problem with them in the present context is that their very purpose is to tolerate contradictions. Such logics thus appear to accept contradictions not only descriptively but *prescriptively* as well. What the normative demand for consistency requires, however, is a deontic logical system that admits the presence of contradictions

descriptively but whose semantics insists that they be resolved at the level of prescriptive obligations.

Non-monotonic reasoning systems (NMRs) with a classical base can describe contradictions while meeting the normative consistency requirement at the prescriptive level, though perhaps not as explicitly as might be desired. NMRs are able to admit contradictions descriptively because they reject monotonicity (i.e., "if K' $\vdash$ p and K' $\subseteq$ K, then K $\vdash$ p"). What monotonicity 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 depends, so defeating that inference. NMRs therefore can describe contradictions while avoiding the deontic explosion of inferences from a contradiction that plagues SDL. NMRs with a classical (rather than paraconsistent) base meet the normative consistency requirement at the prescriptive level because, semantically, they require an explicit preference or choice relation between possible worlds that are maximally consistent in order to continue to draw defeasible inferences. Each possible world of obligations is thus one that meets the normative consistency requirement at the prescriptive level. NMRs also seem promising for purposes of programming autonomous machine agents because they have known efficient implementations such as answer set programming (Gelfond 2008).

4.2 Logics of Belief Revision

Carlos Alchourrón rejects non-monotonic deontic legal logics, however, on the grounds that such systems obscure the distinction between descriptive and prescriptive activity in the law (Maranhao 2006). Alchourrón is a legal positivist who looks outside any formal property of positive law for sources of that law's moral authority. By contrast, Kant understood there to be a necessary connection between law and the moral obligation to obey it. For Kant, a public law that conforms to the Universal Principle of Right will be morally obligatory because of the law's formal structure (universality, consistency, etc.) as well as, to some degree, its substantive content (respect for the constitutional rights of equality, freedom, etc., that the UPR generates for social human beings).

Yet for Kant a number of diverse but internally consistent bodies of legitimate positive public law are possible. Hence like Alchourrón Kant may have some reason to prefer a deontic legal logic that shows the explicit evolution of such a body of law toward the strongest and most coherent system realizing equal freedom under universal law. Logics of belief revision such as Alchourrón's

"AGM" (named after Alchourrón, Gardenfors, Makinson 1985) may thus provide the best approach to implementing Kant's normative requirement of consistency. AGM has 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 (Alchourrón, Gardenfors, Makinson 1985). Rules are not described as defeasible defaults, although they may still achieve appropriately defeasible inferences by Alchourrón's use of a revision operator on the antecedents of conditional obligations (Alchourrón 1991). The ultimate goal of a 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 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 Alchourrón's "f" revision operator (Aqvist 2008), a logic such as AGM may better reflect the normatively consistent system of equal freedom under universal laws constructed by a civil community.

It is important to note that while a deontic logic like AGM may be necessary to capture and reason about duties of right, conformity with those duties might be engineered in a machine agent in a number of different ways (e.g., by symbolic or by statistical, machine-learning techniques, or by some hybrid). The problem of what the *right-making properties* of action are is not the same as the engineering problem of *how to implement right action* in accordance with those properties (see Keeling 2020).

5 Conclusion

I have argued that efforts to build explicitly moral machine agents should focus on public right rather than ethics. *Rightful machines* that respect the priority of right will avoid acting in ways that paternalistically interfere with equal rights of freedom, whereas "ethical" machines that act on popular ethical preferences such as those collected in the Moral Machine Experiment may not. I then showed how shifting the focus from ethics to a standard of public right provides a new approach to resolving deontic conflicts such as those in the trolley problem for autonomous machine agents. Finally, I argued that this shift has important implications for how a deontic logic should handle conflicts between duties or rights.

References

All references to Kant's works follow the German original, *Gesammelte Schriften*, ed. by the Königliche Preussische Akademie der Wissenschaften, 29 Volumes. De Gruyter et al. 1902. English citations follow *The Cambridge Edition of the Works of Immanuel Kant.* Cambridge: Cambridge University Press (1992).

GMS Grundlegung zur Metaphysik der Sitten/*Groundwork of the Metaphysics of Morals*
MS Metaphysik der Sitten/*Metaphysics of Morals*
RL Rechtslehre/*The Doctrine of Right*
TL Tugendlehre/*The Doctrine of Virtue*
ZeW Zum ewigen Frieden/*Toward Perceptual Peace*

Alchourrón, Carlos E. (1991): "Conflicts of Norms and the Revision of Normative Systems". In: *Law and Philosophy* 10, pp. 413–425.
Alchourrón, Carlos E./Gärdenfors, Peter/Makinson, David (1985): "On the logic of theory change". In: *Journal of Symbolic Logic* 50. Nr. 2, pp. 510–530.
Åqvist, Lennart (2008): "Alchourron and Bulygin on deontic logic and the logic of norm-propositions, axiomatization, and representability results". In: *Logique & Analyse* 203. pp. 225–261.
Awad, Edmond/Dsouza, Sohan/Kim, Richard/Schulz, Jonathan/Heinrich, Jospeh/Shariff, Azim Bonnefon, Jean-Francois/ Rahwan, Iyad (2018): "The Moral Machine experiment". In: *Nature* 563. pp. 59–64.
Casey, Bryan (2017): "Amoral Machines, Or: How Roboticists Can Learn to Stop Worrying and Love the Law". In: 111 *Nw. U. L. Rev*. 1347.
Foot, Phillippa (1967): "The problem of abortion and the doctrine of double effect". In: *Oxford Review* 5. pp. 5–15.
Gelfond, Michael (2008): "Chapter 7: Answer Sets". In: *Foundations of Artificial Intelligence* 3. pp. 285–316.
Goble, Lou (2005): "A logic for deontic dilemmas". In: *Journal of Applied Logic* 3. pp. 461–483.
Guarini, Marcello (2012): "Conative Dimensions of Machine Ethics: A Defense of Duty". In: *IEEE Transactions on Affective Computing* 3. Nr. 4, pp. 434–442.
Guyer, Paul (2016): "The Twofold Morality of Recht: Once More Unto the Breach". In: *Kant-Studien* 107. Nr. 1, pp. 34–63.
Hohfeld, Weasley (1919): *Fundamental Legal Conceptions as Applied in Judicial Reasoning*. New Haven, CT: Yale University Press.
Keeling, Geoff (2020): "Why Trolley Problems Matter for the Ethics of Automated Vehicles". In: *Science and Engineering Ethics* 26. Nr. 1, pp. 293–307.
Maranhao, Juliano S. A. (2006): "Why was Alcourron afraid of snakes?". In: *Analisis Filosofico XXVI*. Nr. 1, pp. 162–92.
Liao, S. Matthew/Wiegmann, Alex/Alexander, Joshua/Vong, Gerard (2012): "Putting the trolley in order: Experimental philosophy and the loop case". In: *Philosophical Psychology* 25. Nr. 5, pp. 661–671.
Mikhail, John (2007): "Universal moral grammar: Theory, evidence, and the future". In: *Trends in Cognitive Sciences* 11. Nr. 4, pp. 143–152.

Navarro, P. E., & Rodríguez, J. L. (2014): *Deontic Logic and Legal Systems*. Cambridge University Press.
O'Neill, Onora (2016): "Enactable and Enforceable: Kant's Criteria for Right and Virtue". In: *Kant-Studien* 107. Nr. 1, pp. 111–124.
Thomson, Judith (1976): "Killing, Letting Die, and the Trolley Problem". In: *The Monist* 59. Nr. 2, pp. 204–17.
Thomson, Judith (2008): "Turning the Trolley". In: *Philosophy & Public Affairs* 36. Nr. 4, pp. 359–374.
Tonkens, Ryan (2009): "A Challenge for Machine Ethics". In: *Minds & Machines* 19. Nr. 3, pp. 421–438.