Science and Engineering Ethics
https://doi.org/10.1007/s11948-020-00242-0

ORIGINAL RESEARCH/SCHOLARSHIP

Toward Implementing the ADC Model of Moral Judgment in Autonomous Vehicles

Veljko Dubljević[1]

Abstract
Autonomous vehicles (AVs)—and accidents they are involved in—attest to the urgent need to consider the ethics of artificial intelligence (AI). The question dominating the discussion so far has been whether we want AVs to behave in a 'selfish' or utilitarian manner. Rather than considering modeling self-driving cars on a single moral system like utilitarianism, one possible way to approach programming for AI would be to reflect recent work in neuroethics. The agent–deed–consequence (ADC) model (Dubljević and Racine in AJOB Neurosci 5(4):3–20, 2014a, Behav Brain Sci 37(5):487–488, 2014b) provides a promising descriptive and normative account while also lending itself well to implementation in AI. The ADC model explains moral judgments by breaking them down into positive or negative intuitive evaluations of the agent, deed, and consequence in any given situation. These intuitive evaluations combine to produce a positive or negative judgment of moral acceptability. For example, the overall judgment of moral acceptability in a situation in which someone committed a deed that is judged as negative (e.g., breaking a law) would be mitigated if the agent had good intentions and the action had a good consequence. This explains the considerable flexibility and stability of human moral judgment that has yet to be replicated in AI. This paper examines the advantages and disadvantages of implementing the ADC model and how the model could inform future work on ethics of AI in general.

Keywords Agent–deed–consequence (ADC) model · Autonomous vehicles (AVs) · Artificial intelligence (AI) · Artificial neural networks · Artificial morality · Neuroethics

✉ Veljko Dubljević
veljko_dubljevic@ncsu.edu

[1] Science Technology and Society Program, Department of Philosophy and Religious Studies, NC State University, 453 Withers Hall, 101 Lampe Dr, Raleigh, NC 27607, USA

Published online: 06 July 2020

Introduction

Autonomous vehicles (AVs) (Deng 2015)—and accidents they are involved in—attest to the urgent need to consider the ethics of artificial intelligence (AI). The problems that need to be addressed regarding their role in society include the extent to which they will be used and the ethical algorithms they will be programmed to follow (Wallach 2008). The most notable ethical question discussed so far has been whether we want them to behave in a 'selfish' (i.e., protect the owner and their property at all costs) or utilitarian manner (i.e., reduce the number of lives lost at all costs). Indeed, Shariff and colleagues argue that "[p]eople are torn between how they want autonomous vehicles to ethically behave; they morally believe the vehicles should operate under utilitarian principles, but prefer to buy vehicles that prioritize their own lives as passengers" (Shariff et al. 2017). Thus, society is faced with an ethical quandary, which could be resolved by several ways (Waldrop 2015). One is considering the impacts of varying levels of autonomy in vehicles on society (e.g., stopping short from fully autonomous vehicles). Another is addressing future social changes that seem necessary to incorporate AVs (e.g., separate lanes). A final option is considering alternative models of moral decision making for implementation in AIs. The latter approach is taken up in this article.

As noted above, work with AI in general, and AVs, in particular, was dominated by the use of utilitarianism. Utilitarianism is the 'go-to' ethical framework for implementing algorithmic responses for AVs. However, it is also the basis of empirical work (e.g., surveys) framing public opinion on the behavior of AVs in different morally significant situations (Bonnefon et al. 2016). As noted above, rather than focusing on a single moral system like utilitarianism, there is a need to offer alternative approaches that adequately capture the flexibility of human moral judgment. One possible way to approach implementing ethics for AVs would be to draw on recent work in neuroethics. This would have the advantage that AV technology would be appropriately grounded on research that probes human moral judgment and decision making, thus avoiding outlandish top-down or machine-learning outcomes[1] and promoting better integration into society.

The agent–deed–consequence (ADC) model (Dubljević and Racine 2014a, b) provides a promising descriptive and normative neuroethical account while also lending itself well for implementation in AVs. The ADC model explains moral judgments by breaking them down into positive or negative intuitive evaluations concerning the Agent, Deed, and Consequence, as well as by framing the evaluation as 'high-' or 'low-stakes' in any given situation (see Dubljević et al. 2018). These intuitive evaluations combine to produce a positive or negative judgment of moral acceptability. For example, the overall judgment of moral acceptability in a situation in which someone committed a deed that is judged as negative (e.g., breaking

[1] For instance, machine learning makes learned ethical rules opaque, thereby making transparency impossible. Additionally, disambiguating ethical from unethical discriminations or generalizations is no simple task, as the examples of racist chat-bots attest to. See Abel et al. (2016). See also Misselhorn (2018).

a rule) would be mitigated if the agent had good intentions and the deed had a good consequence. This explains the considerable flexibility and stability of human moral judgment, which has yet to be replicated in AVs (or any AI for that matter).

Further empirical work would be useful to determine the moral weight people assign to each ADC category in different situations. This data could be used when implementing ethics codes into AVs to ensure that it aligns with common human perceptions of morality in specific cases, but the way in which ethics is implemented will need to be complicated enough to adequately reflect the nuances present in human moral judgment. This paper examines the advantages and disadvantages of implementing the ADC model in AVs. It will first introduce the moral challenge that must be resolved if AVs are to share the road with humans. Then, it will explain how the ADC model of moral judgment can be used as a solution of the problem of an ethics code for use in AVs. Finally, it will review and refute several objections regarding computability, feasibility, relevance and weaknesses of the ADC model.

The Moral Challenge of AVs

Little progress has been made in determining what would amount to the single best moral framework. Moral frameworks that have been explored for implementation in AI so far are largely deontological. These frameworks include Asimov's Laws of Robotics (see Anderson 2008; Wallach 2008), Kantian moral theory (Powers 2006), Rossian *prima facie* duties (Anderson and Anderson 2007), and Rawlsian difference principle (Leben 2017). Virtue ethics has also been proposed as a framework for machine ethics (Wallach and Allen 2008, see also Tonkens 2012), and so has Aquinas's doctrine of double effect (Bonnemains et al. 2018). Additional approaches include profession-specific ethics codes (see Anderson et al. 2006; Dennis et al. 2016), general (deontic) logicist models (Bringsjord et al. 2006) and reinforcement learning models (Abel et al. 2016). However, in current discussions about the ethical codes for AVs specifically (as opposed to machine ethics in general), a utilitarian framework (Grau 2006) along with trolley-like dilemmas (Bonnefon et al. 2016) reworked into traffic incidents involving AVs is becoming dominant (Awad et al. 2018). Indeed, it seems that the future of AVs crucially depends on whether corporations will have their way in designing them for limiting financial or legal liabilities, or AVs will guard their passengers no matter the circumstances ('selfish AVs' in further text), or some sort of utilitarianism will be programmed into all AVs (see Fournier 2016). To anyone who thinks that utilitarianism does not adequately capture the intuitive moral sense of humans, this sounds like a grim prospect. At the same time, accidents involving AVs are, to some extent, a result of humans not following rules.[2] For instance, in the highly publicized fatal 2016 Tesla autopilot accident, the death of 40-year-old Joshua Brown was caused when a driver of a freight

[2] This problem is not limited to AVs. A recurring theme in machine ethics is that humans will break rules and this makes for implementing ethics in AI "very challenging" (Bringsjord et al. 2006, p. 12). See also Abel et al. (2016).

truck turned left at an intersection without giving priority to the AV (Singhvi and Russel 2016). This leads to a question on when AVs should be programmed to perform actions contrary to rules, such as swerving into the opposite lane to avoid an otherwise unavoidable crash.

Work with utilitarian AVs is trying to answer this exact question, and yet many fault utilitarianism with justifying immoral actions (see also Fournier 2016). More importantly in this case, the blindness of utilitarian AVs to moral considerations other than consequences would at best inadvertently cause serious problems, and at worst be exploited by malicious actors for nefarious purposes. Consider the example of vehicles used as weapons. If five terrorists[3] use a truck to ram into pedestrians and run headlong into AVs, the utilitarian or selfish AVs sharing the road with humans would only exacerbate the problem. Namely, they would be programmed to either save more lives (for utilitarian AVs) or their passenger(s) and property (for selfish AVs) in each separate instance of collision, and terrorists could then trick AVs into doing their bidding by their sheer number and weight. Pedestrians (individuals and those in small groups) would likely take the brunt of the damage in each and every instance of such 'moral' decision-making of AVs. For instance, the choice of either colliding with a pedestrian or with the truck with five people onboard would predictably yield the decision to collide with a pedestrian (to minimize damage to the AV or to prevent more lives being lost, respectively), even though this is exactly what the terrorists are hoping to accomplish. This example is not idle nor far-fetched in the least. For instance, in 2016, a 19-tonne cargo truck was deliberately driven into crowds of people celebrating Bastille Day in Nice, France (see Hopkins et al. 2016). The attack lasted 5 min, spanned almost 2 km, and the final tally of innocent victims amassed to 86 dead (not including the perpetrator) and 458 injured. Such terror attacks will become much worse if committed when utilitarian or selfish AVs share the road.

Therefore, it is safe to assert that programming AVs with a functional equivalent of a moral theory that is abhorrent in certain situations is at best problematic. Indeed, as Awad and colleagues admit, "any attempt to devise [AI] ethics must be […] cognizant of public morality" (Awad et al. 2018, p. 59). Recent research on human moral decision making reveals that none of the major moral frameworks (such as utilitarianism) can by themselves explain the stability and flexibility of moral judgment, which needs to be somehow replicated if AVs are to share the road safely with humans. Yet, when the three dominant moral frameworks are combined together as in the ADC model of moral judgment, there is an outstanding level of stability and flexibility in both high and low stakes situations—a finding that is consistent between lay and expert moral judgments (see Dubljević et al. 2018). The purpose of this paper is to argue that the ADC model of moral judgment can be used as

[3] For purposes of this argument, the label 'terrorist' is used for any malicious actor that deliberately targets civilians, regardless of the ideology. So, 'ISIS' fighters, white supremacists and even individuals targeting others to protest their 'involuntary celibacy' all fall under the same term. Even though there is no space to argue for that here, utilitarianism fails to incorporate any kind of malicious intent, and would be likely exploited even more frequently in less tragic ways, say to commit acts of vandalism. I am grateful to Kevin Richardson for constructive comments that prompted me to make this clear.

a viable alternative for implementation in AVs as the basis for machine learning in an AI hybrid approach.[4] But what is the ADC model of moral judgment?

The ADC model gives an explanation of the stability and flexibility in moral judgment in terms of a balance of intuitions. Moral judgment relies on at least three different sets of moral intuitions: evaluations (positive vs. negative) of the character of a person (the agent-component, A), their deeds or actions (the deed-component, D), and the consequences of these actions (the consequences-component, C). Most untrained individuals do not have explicit knowledge of ethics, and yet their intuitive moral judgments correspond to certain moral precepts from ethical theories (Dubljević and Racine 2014a, b), such as virtue ethics, deontology, and consequentialism. Hence, in the example of breaking rules given above, the overall balance of moral intuition which relies on evaluations of agents, deeds, and consequences is integrated and might result in overriding rule-following in certain instances. Therefore, a wrong deed can be more acceptable if both the character of the person (the agent) and the consequences of the action are positive. The model provides formulas (see Dubljević and Racine 2017) for representing such situations (e.g., [A+], [D−] and [C+] = [MJ+] or positive moral judgment). Finally, the ADC model predicts that the intuitive plausibility of three major types of ethical theory (virtue ethics, deontology, and consequentialism) rests on the intuitive evaluations of A, D, and C respectively, that the explicit moral precepts from these theories can be sufficiently dissociated from each other and operationalized by using survey methodology. In fact, the ADC approach has a considerable advantage over other empirical approaches to moral judgment, as it explores both high stakes (i.e., involving life-threatening dilemmas) and low stakes situations (i.e., non-life threatening moral dilemmas). Indeed, the ADC model has been empirically confirmed in a recent study which operationalized intuitive aspects of all three moral theories with 152 professional philosophers and 1314 lay people with no training in ethics (Dubljević et al. 2018). This approach is an improvement to previous work, which has been limited to contrasting only two moral theories (see Christensen and Gomila 2012 for a review).

[4] Hybrid approaches avoid the difficulties in both top-down (programming rigid rules) and bottom-up (relying solely on machine learning), and combine their strengths. As Wallach rightly notes “Engineers typically draw on both a top-down analysis and a bottom-up assembly of components in building complex automata. If the system fails to perform as designed, the control architecture is adjusted, software parameters are refined, and new components are added. In building a system from the bottom-up the learning can be that of the engineer or by the system itself, facilitated by built-in self-organizing mechanism, or as it explores its environment and the accommodation of new information.” (Wallach 2008, p. 468). Additionally, as Bringsjord and colleagues note, implementation must “begin by selecting an ethics code C intended to regulate the behavior of R [robots]. […] C would normally be expressed by philosophers essentially in English [or another natural language] […before] formalization of C in some computational logic L, whose well-formed formulas and proof theory are specified” (2006, p. 3).

Is the ADC Model Any Better at Providing Solutions for the Moral Challenge of AVs?

The purpose of this paper is not to defend the ADC model as a single unified moral theory, but only to show how it can be developed as an implementable natural language solution to complex socio-moral dilemmas facing AVs. So, it will not be argued here that the ADC model explains many of the problems that single factor moral theories struggle to resolve in terms of specific balances of ADC intuitions, similar to the rule-breaking example given above (e.g. that morally good people might do bad things or that benefitting the majority might come with horrible moral costs), as has been done by Dubljević and Racine in prior publications (e.g., Dubljević and Racine 2014a, b).

This paper has a fairly modest aims. First, it asserts that most prior research has been dominated by 'trolley-like' work, which has the disadvantage of reducing the dynamic decisional structure of morality into a simple binary choice. It assumes that we need to generate better dilemmas that could be applied in both human and AI decision making research and calibration. The most important claim is that the ADC model presents a viable alternative, and a major improvement to previous work.[5] But how does the ADC model fare in a vehicle terror attack situation as discussed above?

For the sake of the argument, let's assume that a functional equivalent of morality in terms of both assessing agents, deeds, and consequences and distinguishing between high and low stakes situations is computable—and, in fact, programmed—into AVs. At the beginning of a 2 km stretch of road shared by pedestrians, regular vehicles and AVs, 5 terrorists in a truck start ramming pedestrians and running headlong into AVs. What would happen?

Well, in the first couple of seconds, the AVs would behave very much like the way in which humans behaved in the real-world example or in the situation with utilitarian AVs: they would try to avoid any casualties. Crucially, however, they would register and transmit the information that this particular truck is consistently engaging in 'high-stakes' unlawful behavior. After the truck has been tagged as 'negative' in the system, AVs with no passengers on board (e.g., AV freight trucks) would no longer swerve to avoid the truck holding the 5 terrorists if that meant breaking laws (i.e., the algorithm would calculate that both [A-] and [D-] are not acceptable). Additionally, AVs with passengers would not count the five terrorists in any trade-off of human lives as having weight over individual pedestrians. Finally, fully autonomous

[5] Unlike trolley problems, which are simple binary choices, vignettes and situations designed for the ADC approach have eight distinct versions and can capture weights that people assign to these factors in dramatic or mundane situations (see Dubljević et al. 2018). At this point, one can be neutral on the computer engineering question of implementation via classical symbol system or connectionist/neural network system or even a compatibilist connectionist-simulated-on-classical system. The main concern is only that the AVs should be able to encode the ADC model of moral judgment. The problem is still at the level of human agreement on a specific code to be implemented. As Bringsjord and colleagues rightly note "if humans cannot formulate an ethical code […] [a] logic-based approach is impotent" (2006, p. 13). I am grateful to Ron Endicott for constructive comments that prompted me to make this explicit.

AVs with no passengers might actually prevent the truck in question from traveling further by creating a road-block. The low-stakes negative consequence of disrupting traffic would be overridden by the high-stakes imperatives to detain malicious agents and prevent further loss of life. In fact, AVs would behave more like humans, similar to a motorcyclist who tried to board the freight truck in the Nice attack but had to jump off once the terrorist tried to shoot him (see Hopkins et al. 2016). Such AVs would also be much more efficient though, because—unlike humans—they wouldn't panic or fear for their own safety when intervening in malicious acts.

AVs based on ADC would also outperform selfish or utilitarian AVs. Selfish AVs would avoid danger to their passengers or damage to property, thus playing into the plans of the terrorists. It will be remembered that a 'selfish AV' has the choice of either colliding head-on with the truck with terrorists or killing a pedestrian which is running away on a side-walk. 'Preventing or minimizing damage,' along with saving the passenger would predictably yield a choice of ramming a pedestrian (e.g., [1 death + $100 damage vs. 1 death + $20,000 damage]). Similarly, utilitarian AVs (whether they carry passengers or not) would calculate every instance of potential collision with the truck with 5 terrorists as 'immoral' and killing one pedestrian to save those 5 terrorists as preferable [e.g., repeated instances of binary decision making (1 death vs. 5 deaths)], as programming which recognizes malicious intent is not entertained even in principle. Both selfish and utilitarian AVs would actually make things much worse since terrorists might plan their attack by knowing the limitations (and rigidity) of AI decision making software implemented in AVs.[6]

Are There Any Issues with Programming ADC into AVs?

There are several objections that need to be addressed at this point: objections questioning the computability, feasibility, relevance and weaknesses of the ADC framework for implementation in AI in general and AVs in particular. First of all, a fundamental objection would be that moral understanding can't be modeled computationally (see Misselhorn 2018). Indeed, how does one program duties and intentions?

In short: one doesn't. Instead of programming duties and intentions, AI and AVs can be equipped with functional equivalents or even 'limited functional simulacra' of morality.[7] Deontological aspects have already been implemented, since AVs have a functional equivalent of 'observing the law' (D-component in the ADC algorithm). Facial recognition technology can be used to implement avoidance algorithms that explicitly limit the number of humans that would be hurt in collision situations (C

[6] It is perhaps possible that a more complex version of utilitarian-inspired decision making algorithms would fare better in this regard, but to my knowledge, no current work on utilitarian AVs is entertaining malicious intent, or the difference between low and high stakes situations, as serious issues for implementation. I am grateful to Bill Bauer for constructive comments that prompted me to make this clear.

[7] I'm grateful to Michael Pendlebury and other audience members at the "Work in progress in philosophy" session at NC State University, on Oct 26th, 2018, for helpful and constructive comments that prompted this distinction.

component in the ADC algorithm). Finally, tagging of malicious actors (as in the example above) can be achieved by using the 'Identification Friend or Foe' (IFF) technology, which has been developed to distinguish between friendly and enemy individuals, vehicles or aircrafts in military engagements (Bowden 1985) by using transponders and broad characterization categories of 'friend,' 'enemy,' 'neutral,' or 'unknown.' It is possible to use this technology in civilian settings to designate all apparently law-abiding human participants in traffic as 'friend' (A+), those humans (or vehicles) that are harmlessly breaking the law as 'unknown' or 'neutral' (A?) and those engaging in considerable harm (such as ramming pedestrians, but also 'hit-and-run') as 'enemy' (A-) who might need to be harmlessly detained if possible.[8] An additional designation of 'malfunctioning' could be added into the system to facilitate functional equivalents of moral decisions in the hybrid environment where AVs share the road with humans. This would still leave issues with the moral status of animals involved in traffic accidents and collisions (see e.g., Luetge 2017), but for now, it is safe to assert that the ADC model performs better than competing frameworks considered for implementation in AVs, namely selfish and utilitarian AVs.

A second objection concerns feasibility—how feasible is it to have international standardization based on a specific model of moral judgment? Indeed, a significant problem emphasized by many authors is that people can't agree on a single moral theory in any given society, let alone in an international context. However, disagreement in the international context is not as big of a hurdle as it may seem at first glance. For instance, there is no international agreement on whether vehicles should drive on the left side (as in the United Kingdom and some of its former colonies) or the right side (as in the rest of the world), and this does not prevent motor vehicles from being driven nationally or internationally. On the point of reasonable disagreement on morality within a society, the ADC model actually helps to overcome such disagreement because it is not based on any single moral theory but, in fact, systematizes the intuitive bases for three major moral theories. The 'fact of reasonable pluralism' regarding morality may not need to be resolved at all since the disagreement concerns the 'ultimate' normative basis for moral theory, whereas there is ample agreement on many practical moral matters. This applies to many practical moral issues in many countries, ranging from principles of biomedical ethics as developed in the United States (Beauchamp and Childress 2013) to the ethics code for automated driving in Germany (Luetge 2017). On the latter point, Luetge notes that "while there was considerable disagreement in the discussions, ultimately, in most questions, a consensus in practical matters could be reached" (Luetge 2017, p. 557).

The third objection concerns practical relevance. Namely, one could argue that it is inefficient to instantiate a whole IFF technology with 'transponder tagging' based on the possibility of infrequent events where the technology might be useful. Indeed, there were no more ramming incidents in France after the Nice attack, and other similar terrorist activities involving vehicles around the world were mostly single

[8] Indeed, 'hit-and-run' incidents are the most likely moral situation that AVs will encounter, but the difference in 'high-' and 'low-stakes' moral situations is a crucial addition. More on that below.

instance occurrences. As Shariff and colleagues note, “overreactions risk slowing or stalling the adoption of autonomous vehicles” (Shariff et al. 2017). Indeed, if the implementation of ADC in AVs required additional technical solutions that were costly and rarely useful, this would militate against the adoption of this framework. However, the ADC model is far from being useless in day-to-day moral decisions, since it explicitly addresses common ‘high-stakes’ problems, such as hit-and-run incidents, which would be automatically reported to the police, as well as low-stakes moral judgments.

Consider the example of a stalled freight vehicle on a country road with a no-passing line in the middle. Human drivers immediately assess whether the situation is high- or low-stakes with the use of intuition, but the behavioral repertoire of AVs is not so adept at problem-solving. Indeed, if we take seriously the lessons learned from Asimov’s stories, rigidity in machine principles leads to many practical problems (see Wallach 2008). In this particular case, utilitarian AVs would simply wait, since no lives are in danger. Indeed, several AVs would endlessly wait behind the stalled vehicle, causing completely avoidable traffic jams. The ADC framework, though, as discussed above, offers a simple solution: if the IFF system is in place, the transponder of the stalled vehicle will transmit the ‘malfunctioning’ signal.[9] In the least beneficial scenario, this would allow AVs to infringe rules proscribing crossing a no-passing line if no immediate harm would result. In the most beneficial scenario, any human passengers that were traveling with that stalled vehicle would be tagged as ‘friendly’ and AVs with no passengers might provide them with transportation to the next settlement or safe environment.[10]

In any case, these and other scenarios need to be tested in a simulated environment before AVs start sharing the road with humans. Indeed, the situations discussed here and methods tested with the aid of the ADC model will contribute to more valid assessments of moral evaluation for AI in general and AVs in particular. This is not to say that the implementation of the ADC model will avoid all irresolvable or tragic dilemmas (see Hursthouse 1999). However, at the very least, the ADC model will provide a platform for assessment of what went wrong. As Awad and colleagues note: “[P]eople’s willingness to buy autonomous vehicles and tolerate them on the roads will depend on the palatability of the ethical rules that are adopted” (Awad et al. 2018, p. 61).

This leads to the final objection: it could be argued that the weakness of the ADC model of moral judgment is that it merely addresses the issue how people make moral judgments, and not how they *ought to* make moral judgments. In other words, it could be argued that the ADC model is adequate descriptively, but not normatively. This is a recurring challenge that is frequently leveled against empirically informed approaches and neuroethics is no exception. It is rooted in Hume’s work and later anti-naturalist positions (e.g., Moore). A classical

[9] The assumption here is that implementation of the transponder system would be mandatory at vehicle registration for both AVs and regular vehicles.

[10] This might need to be qualified with an override preventing those human passengers from taking control of the AV, so as to thwart any malicious or nefarious plans of exploiting the system.

philosophical response to Hume's challenge is the so-called Kant's dictum: 'ought implies can'. Now, there is still some philosophical debate over whether Kant in fact argues for the strong or weak version of the dictum, what the proper application of it entails and whether Kant's dictum proves that Hume's challenge is completely false or not (see Spielthenner 2017). However, regardless of such controversies, Kant's dictum is rarely rejected altogether and is, in fact, the basis of moderate naturalistic approaches, including Dewey's pragmatic naturalism (see Dewey 1929). Thus, well-conducted empirical research that provides certain facts about the human condition is admissible in ethical reasoning and may have legitimate normative implications. However, critics may be unimpressed by falling back to Kant and argue that the ADC model has normative weaknesses that make it ineligible for implementation in AVs.

However, it may be prudent first to establish what the real problem is. Historically, the charge of using illicit reasoning from *is* to *ought* has been very useful in debunking claims or attempts to misuse some empirical data to reach morally repugnant normative conclusions. A good example is racism, which claimed that certain crucial moral properties are genetically determined since allegedly some races are less intelligent or capable of moral judgment. Another good example is sexism, for instance, in the claim that is simply a fact that women are paid less than men, and this is as it should be.

The real worry then is about jumping to conclusions, a sort of cautionary tale, since further work needs to be added before normative implications of any single fact can be imputed. Thus, it is evident in cases where a mere social convention, say that women are paid less than men, is reified and made into a sort of natural fact, which would somehow or the other justify the normative conclusion that women should be paid less than men. Thus, the is-ought gap is a useful philosophical tool for assessing whether any given empirically informed approach is normatively unsound. On the other hand, if this was allowed to be a 'carte blanche', a general 'debunking' of any empirically informed discussion, this is yet another example of over-reaching. This produces not only counter-intuitive but also bizarre conclusions. Let's say that someone claims that to promote real gender equality, men have a moral obligation to give birth. This (widely implausible) example is immediately precluded by Kant's dictum: ought implies can, and since men cannot give birth (a natural fact) no such obligation can hold any normative force. The research showing that people do in fact make a moral decision in line with the ADC model of moral judgment, and that it incorporates normative intuitions underpinning the three dominant moral theories (virtue ethics, deontology, and consequentialism, respectively) gives credence to a prima facie assumption that the ADC model is normatively sound. It is then up to the critic of the ADC model to provide evidence of normative inadequacy. Therefore, it is safe to assert now that the ADC model is an improvement to existing alternatives and that this will motivate more research that goes beyond the 'trolley problems' and takes seriously not only the need to discuss low-stakes moral decisions but also the intuitive sources of human moral judgment.

Conclusion

The ADC model provides a viable alternative to utilitarian or selfish AVs. Unlike utilitarian and selfish AVs, which could be exploited for nefarious purposes by individuals and groups with malicious intent, the AVs with the ADC model-based ethical code would in principle be functionally equipped with recognition of morally bad agents in the real world. The example of AVs is chosen as the first practical application of the ADC model for AI, but others are possible as well. Namely, if there is agreement that the ADC model is a better alternative for implementing an ethics code into AI than currently available single-focus approaches, then we could reasonably expect more applications. However, more work needs to be done to ascertain the applicability of such ADC model inspired systems, both in terms of surveying public attitudes and in terms of simulating the programming of ADC into AI before actual people are impacted. Such work will enrich the research on AI ethics.

References

Abel, D., MacGlashan, J., & Littman, M. L. (2016). Reinforcement learning as a framework for ethical decision making. In B. Bonet et al. (Eds.), *AAAI workshop: AI, ethics, and society. AAAI workshops*. AAAI Press.

Anderson, S. (2008). Asimov's 'three laws of robotics' and machine metaethics. *AI & SOCIETY, 22*(4), 477–493.

Anderson, M., & Anderson, S. (2007). The status of machine ethics: A report from the AAAI symposium. *Minds and Machines, 17*(1), 1–10.

Anderson, M., Anderson, S., & Armen, C. (2006). MedEthEx: A prototype medical ethics advisor. In *Proceedings of the national conference on artificial intelligence* (p. 1759). MIT Press.

Awad, E., Dsouza, S., Kim, R., Schulz, R., Henrich, J., Shariff, A., et al. (2018). *The Moral Machine Experiment*. https://doi.org/10.1038/s41586-018-0637-6.

Beauchamp, T. L., & Childress, J. F. (2013). *Principles of biomedical ethics* (7th ed.). New York: Oxford University Press.

Bonnefon, J., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*. https://doi.org/10.1126/science.aaf2654.

Bonnemains, V., Saurel, C., & Tessier, C. (2018). Embedded ethics: Some technical and ethical challenges. *Ethics and Information Technology, 20,* 41–58.

Bringsjord, S., Arkoudas, K., & Bello, P. (2006). Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intelligent Systems, 21*(4), 38–44.

Christensen, J. F., & Gomila, A. (2012). Moral dilemmas in cognitive neuroscience of moral decision-making: A principled review. *Neuroscience and Biobehavioral Reviews, 36,* 1249–1264.

Deng, B. (2015). Machine ethics: The robot's dilemma. *Nature, 523,* 24–26.

Dennis, L., Fisher, M., Slavkovik, M., & Webster, M. (2016). Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems, 77,* 1–14.

Dewey, J. (1929). *The quest for certainty: A study of the relation of knowledge and action*. New York: Milton, Balch & Company.

Dubljević, V., & Racine, E. (2014a). The ADC of moral judgment: Opening the black box of moral intuitions with heuristics about agents, deeds and consequences. *AJOB Neuroscience, 5*(4), 3–20.

Dubljević, V., & Racine, E. (2014b). A single cognitive heuristic process meets the complexity of domain-specific moral heuristics. *Behavioral and Brain Sciences, 37*(5), 487–488.

Dubljević, V., & Racine, E. (2017). Moral enhancement meets normative and empirical reality: assessing the practical feasibility of moral enhancement neurotechnology. *Bioethics, 31*(5), 338–348.

Dubljević, V., Sattler, S., & Racine, E. (2018). Deciphering moral intuition: How agents, deeds and consequences influence moral judgment. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0204631.

Fournier, T. (2016). Will my next car be libertarian or utilitarian? Who will decide? *IEEE Technology & Society, 35*(2), 40–45. https://doi.org/10.1109/mts.2016.2554441.

Grau, C. (2006). There is no 'I' in 'Robot': Robots and utilitarianism. *IEEE Intelligent Systems, 21*(4), 52–55.

Hopkins, N., Chrisafis, A., & Fischer, S. (2016). *Bastille day attack: 'Hysterical crowds were running from death*. The Guardian. https://www.theguardian.com/world/2016/jul/15/nice-truck-attack-victims-survivors-bastille-day-crowds. Accessed November 15, 2018.

Hursthouse, R. (1999). *On virtue ethics*. Oxford: Oxford University Press.

Leben, D. (2017). A Rawlsian algorithm for autonomous vehicles. *Ethics and Information Technology, 19,* 107–115.

Lord Bowden of Chesterfield. (1985). The story of IFF (identification friend or foe). *IEE Proceedings, 132*(6 pt. A), 435–437.

Luetge, C. (2017). The German ethics code for automated and connected driving. *Philosophy & Technology, 30,* 547–558.

Misselhorn, C. (2018). Artificial morality: Concepts, issues and challenges. *Society, 55,* 161–169.

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

Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2017). Psychological roadblocks to the adoption of self-driving vehicles. *Nature Human Behavior*. https://doi.org/10.1038/s41562-017-0202-6.

Singhvi, A., & Russel, K. (2016). *Inside the self-driving tesla fatal accident*. The New York Times. https://www.nytimes.com/interactive/2016/07/01/business/inside-tesla-accident.html Accessed November, 2018.

Spielthenner, G. (2017). The is-ought problem in practical ethics. *HEC Forum, 29*(4), 277–292.

Tonkens, R. (2012). Out of character: On creation of virtuous machines. *Ethics and Information Technology, 14,* 137–149.

Waldrop, M. M. (2015). Autonomous vehicles: No drivers required. *Nature, 518,* 20–23.

Wallach, W. (2008). Implementing moral decision making faculties in computers and robots. *AI & SOCIETY, 22,* 463–475.

Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong*. Oxford: Oxford University Press.