Algorithmic Suitability of Ethical Frameworks

A Literature Review of Computational Ethics in Autonomous Vehicles

Alex Crist

Abstract

Three different ethical theories are examined for their application in computational ethics in relation to the trolley problem posed by autonomous vehicles. A deontological ethical approach, a Rawlsian approach, and a utilitarian approach that utilizes prima facie duties are all judged based on their consistency, completeness, and practicality, and agreement with human intuition. The Kantian approach, the deontological framework of choice, struggles to create a specific plan that an algorithm could follow while also failing to prove that an action could be produced in all scenarios. Practicality is gained through the Rawlsian approach, but agreement with human intuition is not universal given specific edge-case scenarios. Finally, Anderson's prima facie duty based utilitarian algorithm satisfies all the same needs as the Rawlsian approach while also producing decisions more inline with human intuition due to the algorithms reliance on human trainers and complex relationship modelling through inductive-logic programming.

Introduction

Humans are terrible drivers; in the United States, human mistakes account for 93% of all vehicle accidents (Sütfied 2017). By comparison, autonomous vehicles never fatigue, never lose focus, and make all decisions with intention, which prevents a large majority of their would-be collisions. Despite their precision, self-driving cars are still put in positions that will inevitably involve crashes and in rare cases, a computer will have to decide who gets hurt and who does not. This contemporary issue can be simplified to resemble the classic ethical trolley-problem in which a moral agent must make a choice between allowing a trolley to impact five people or to instead divert its path to impact a single person. The key difference between the classic trolley problem and the self-driving trolley problem is that autonomous vehicles may be faced with a multitude of decisions rather than two, but the basis of the ethical dilemma remains the same. A solution will require a the synthesis of algorithms that can crunch ethical calculations.

While any single ethical framework is likely to be imperfect in making ethical decisions, some are more suitable than others for producing a solution. In this paper, the existing literature on computational ethics using different ethical theories is reviewed and their application in the self-driving trolley problem is analyzed. To precisely evaluate each theory, Anderson's four principles of good action-based ethical theories are used. The principles dictate that an ethical theory should be consistent in its decisions, complete in its ability to produce an action in all scenarios, practical to implement, and in agreement with human intuition (2006).

Deontological ethical algorithms

Among the most popular ethical theories are deontological ethics, which judge whether actions are right or wrong based on sets of rules. One such rule-based framework is Immanuel Kant's system of ethics in which he states, "Act only according to the maxim whereby you can at the same time will that it should become a universal law" (1981). From any moral action, a rule can be extracted that should universally be applied to all people regardless of anything else. To see this in practice, from the specific moral action, "I will not cheat on my test", the universal moral rule that "cheating is wrong" could be extracted.

In Thomas Powers's paper, "Prospects for a Kantian Machine", he makes the argument that a deontological ethical framework is most suitable for an algorithmic implementation because these frameworks are based on rules and algorithms can act on rules (2006). Powers explores three different directions for how Kant's universal rules could be used in a computer system, all of which rely on the idea that the algorithm will be in possession of a set of universal rules that it can then use to evaluate moral actions. Powers claims that "the formation of normative claims" is the human step that allows people to connect the factual inputs of their lives to their moral actions and subsequently will be the key algorithmic step in connecting machine ethics to ethically calculated actions (2006).

While Powers expertly lays out many subtle facets of Kantianism, his argument that a deontological ethical framework is ideal for computational ethics is overshadowed by a lack of clear connection to exactly how such an algorithm would actually be implemented. The details are left vague, and unlike the next two ethical algorithms, no specific guidelines are offered for how the discussed algorithm could be created. The missing keystone to this argument lies in the lack of access to a set of universal rules that moral agents should follow. Powers's only response to this is, "the machine might itself construct a theory of ethics by applying the universalization step to individual maxims and then mapping them onto traditional deontic categories- namely, forbidden, permissible, and obligatory" (2006). However, even if such a set of universal morals existed, it would be mathematically impossible to prove that the set of rules is able to handle all inputs according to Rice's theorem which proves that any nontrivial property of an algorithm is undecidable (2015).

In the frame of the trolley problem, it is unclear what universal rules might be created that would guide what actions should be taken in an accident. While Kantianism is a legitimate ethical theory that could meet Anderson's criteria of consistency and agreement with human intuition, Rice's theorem shows that Powers's algorithm will always be impossible to prove for completeness in producing an action in all situations. This incompleteness in combination with the algorithm's difficulty of implementation makes it entirely impractical for usage in computational ethics.

Rawlsian ethical algorithms

A far more practical ethical framework for application in the self-driving trolley problem comes in an adaptation of philosopher Thomas Rawls's beliefs about justice in society. To summarize Rawls's beliefs, he presents that society is unjust due to the fact that if given member of the society were to create the society before being born into it, they would choose to make it more equitable due to the risk of being born into an unfavorable position. Rawls's labels his idea of not knowing where one will be born as "the veil of ignorance". Finally, Rawls argues that from behind the veil of ignorance, one will design a society such that the least well-off person is as well-off as possible (1972).

Derek Leben applies this Rawlsian idea of maximizing the well-being of the worst-off person in a situation to computational ethics in his paper, "A Rawlsian algorithm for autonomous vehicles" (2017). Leben makes the assumption that in a self-driving trolley situation, the vehicle's computer will be able to calculate the probability for survival for each human affected by the crash on a scale from zero to one, where zero represents death, one represents no-injury, and intermediate values represent varying degrees of injuries. Given multiple sets of these survival-probability values which each represent outcomes of different crash possibilities, Leben's algorithm's primary priority is to choose a situation where the lowest value the chosen outcome's survival-probability set is as high as possible.

In an accident, the autonomous vehicle might calculate three possibilities, corresponding to driving straight, swerving left, or swerving right with {0.3, 0.8, 1}, {0.1, 1, 1}, and {0.4, 0.4, 1} representing the life-probability values of the three involved humans respectively. Using Leben's strategy, the computer would swerve right, choosing the life-probability set of {0.4, 0.4, 1} because the set's minimum value, 0.4, is the maximum of the three sets' minimum values.

Leben's algorithm has a key improvement over Powers's. The implementation of his algorithm is specific making it a highly practical option. Additionally, the framework satisfies two more of Anderson's conditions for good ethical theories: consistency and completeness. One potential issue, however, is whether the result of the algorithm always agrees with human intuition. Leben himself notes, "Technically the Maximin principle prefers an infinite number of severe injuries to the death of a single person" (2017). This result defies human intuition that the maiming of a million lives is inferior to the death of a single person. Consequently, this algorithm is still imperfect and in certain scenarios, a different approach could be more appropriate.

Utilitarian ethical algorithms

In Anderson's foundational paper, "An approach to Computing Ethics", a third ethical theory is applied in computer science: utilitarianism (2006). Hedonistic Act Utilitarianism, the specific variety Anderson and his colleagues are concerned with, evaluates actions simply by summing how many units of pleasure and displeasure they produce. Positive sums correlate to moral actions while negative sums correlate to immoral actions. Anderson feels that a single pleasure value is too simplistic as a means of evaluating actions and therefore argues that multiple *prima facie duties* should be used to break the problem into multiple values from which an algorithm produce more complex analysis. As Anderson explains, "A prima facie duty is an obligation that we should try to satisfy but that can be overridden on occasion by another, stronger duty" (2006). For the purpose of general ethics, he proposes W.D. Ross's seven-principles: fidelity, reparation, gratitude, justice, beneficence, nonmaleficence, and self-improvement.

To create an algorithm that operates according to this ethical system, ethics experts are given scenarios where they must rate each prima facie duty of two opposing actions with a numerical value with either -2 (serious violation), -1 (less serious violation), 0 (neutral), +1 (minimal satisfaction), or +2 (maximal satisfaction). The experts then choose which of the two actions they believe to be morally preferable. With enough scenarios, this data becomes a powerful training set that is fed into an *inductive-logic programming* (ILP) network, a machine learning technique that is able to represent far more complex relationships between the prima facie duties than simple linear coefficients.

Anderson demonstrates that for different ethical applications, different prima facie duties may be chosen (2006). For the purpose of evaluating how an autonomous vehicle should act during an accident, the following duties could be used: likelihood of survival, severity of injury, severity of property damage, and intensity of intervention (continuation versus swerving). Experts would then review thousands of self-driving trolley situations, choosing what they believe to be the correct action for the vehicle. Each situation would have a value assigned for each prima facie duty. As in Anderson's explanation, this data would then be used to train an ILP model that could produce decisions given new situations.

Anderson's highly specific technique makes it a prime candidate for implementation in autonomous vehicles. The flexibility of the prima facie duties allows Anderson's algorithm's focus to be narrowed to the application of self driving cars. This technique does require a large training dataset, something that was unnecessary in Rawls's algorithm, but the trade-off is the powerful results that this algorithm could yield. Without an actual implementation of the algorithm, it's difficult to conjure scenarios that the algorithm would yield counter-intuitive results for. However, due to the complex relationships that the ILP network would create between the prima-facie duties, it is likely that it would be more robust than the Rawlsian algorithm in producing results that align with human intuition. This result, together with algorithm's specificity, consistency, and completeness, yields is a promising candidate for implementation in self-driving vehicles.

Conclusion

Three ethical frameworks have been analyzed for their use in computational ethics in reference the self-driving trolley problem. Despite the efforts of Powers's "Prospects for a Kantian Machine", no practical algorithm has been outlined that implements a deontological ethical framework. Furthermore, deontological frameworks' reliance on sets of arbitrary rules provides an inherent issue for the completeness of the algorithm as shown using Rice's theorem. Leben demonstrates the implementation for a Rawlsian algorithm that is practical, complete, and consistent, but fails to produce results that universally agree with human intuition. Finally, Anderson's examination of utilitarian ethics through prima facie duties provides a promising algorithm that like Leben's is practical, complete, and consistent. However, due to the human-guided training and complex relationships that ILP allows between prima facie duties, Anderson's algorithm likely produces results that are more consistently in line with human intuition. Further steps in the analysis of ethical frameworks would be to look at a virtue based ethics approached, as Wallach and Allen have done in their book, "Moral Machines: Teaching robots right from wrong" (2010).

References

Anderson, M., et al. “An Approach to Computing Ethics.” *IEEE Intelligent Systems*, vol. 21, no. 4, 2006, pp. 56–63., doi:10.1109/mis.2006.64.

Kant, Immanuel, and James W. Ellington. *Grounding for the Metaphysics of Morals*. Hackett Pub. Co., 1981.

Leben, Derek. “A Rawlsian Algorithm for Autonomous Vehicles.” *Ethics and Information Technology*, vol. 19, no. 2, 2017, pp. 107–115., doi:10.1007/s10676-017-9419-3.

Powers, T.m. “Prospects for a Kantian Machine.” *IEEE Intelligent Systems*, vol. 21, no. 4, 2006, pp. 46–51., doi:10.1109/mis.2006.77.

Rawls, John. *A Theory of Justice*. Belknap Press of Harvard University Press, 1972.

Sipser, Michael. *Introduction to the Theory of Computation*. Course Technology Cengage Learning, 2015.

Sütfeld, Leon R., et al. “Using Virtual Reality to Assess Ethical Decisions in Road Traffic Scenarios: Applicability of Value-of-Life-Based Models and Influences of Time Pressure.” *Frontiers in Behavioral Neuroscience*, vol. 11, 2017, doi:10.3389/fnbeh.2017.00122.

Wallach, Wendell, and Colin Allen. *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, 2010.