Consequentialism & Machine Ethics

Towards a Foundational Machine Ethic to Ensure the Right Action of Artificial Moral Agents

Josiah Della Foresta

McGill University

1 Introduction

In this paper, I will argue that Consequentialism represents a kind of ethical theory that is the most plausible to serve as a basis for a machine ethic. First, I will outline the concept of an artificial moral agent and the essential properties of Consequentialism. Then, I will present a scenario involving autonomous vehicles to illustrate how the features of Consequentialism inform agent action. Thirdly, an alternative Deontological approach will be evaluated and the problem of moral conflict discussed. Finally, two bottom-up approaches to the development of machine ethics will be presented and briefly challenged.

When 23 of the world's most eminent artificial intelligence experts were asked for their predictions regarding the emergence of artificial general intelligence (AGI), the average date offered was 2099.[1] Further noting the promise and peril the development of AGI represents, this century may come to be defined by humanity's disposition towards and progress in the development of ethically-informed machine intelligence.

1. Ford 2018.

Of course, there exist many candidate normative principles and values that an AGI might be designed with, and it is not altogether obvious *which* principles and values ought to be used as opposed to others. Further complicating matters is the question of *how* to encode the chosen principles and values such that an AGI will consistently behave as prescribed. While the former question is of primary concern, the latter will also be considered.

2 Artificial Moral Agents & Consequentialism in Kind

It is typically clear what makes a machine good. If it functions as it is expected to, it is a good machine. Furthermore, this expected function is narrow such that the machine is not expected to deviate from set instructions and success or failure in accomplishing said instructions are readily determinable. Yet, an *artificial general intelligence* would be a sapient machine possessing an open-ended set of instructions where success or failure lies beyond the scope of merely completing a task. This machine would be autonomous and act in a world populated by moral agents: humans.

Indeed, an AGI would itself be a moral agent–an artificial moral agent (AMA). Assuming that an AGI would necessarily be an AMA, we are confronted with the pressing concerns introduced above: how ought an AMA behave in the social world, and how might it come to behave that way? Following the introduced thesis, this section will define what is meant by Consequentialism as a kind of ethical theory, and how it relates to AMAs.[2]

Consequentialism as a class of ethical theory can be said to be typified by two theses.[3]

- Value is independent of the right
  - It is possible to give an account of the value of states of affairs and thus a comparative ranking of alternative states of affairs without appeal to the concept of

2. In the interest of brevity, the moral agency of AGI was assumed. Sullins' criteria of autonomy, intentionality and responsibility are here assumed on the part of AGI. See Sullins (2011).
3. Timmons 2013, 112.

right action. The states of affairs relevant in a consequentialist theory of right action are consequences (or outcomes) related in some way to the action.

- Explanatory priority of value
  - A full and proper explanation of what makes an action right (or wrong) can be given (*solely*) in terms of the value of the consequences that are related in some way to the action.

Reduced to a single proposition, Consequentialist ethical theories would see that an act is right if and only if (and because) some action A possesses the best consequences of all available alternatives (*B, C, D...*) within the agent's power from an impartial standpoint.[4] Consequentialism is thus a kind of ethical theory that is *teleological*, *maximising*, and *impartial*. That is, theories of the kind assess the deontic status of agent action based on their consequences (*teleological*), the action with the best consequences is the obligatory action an agent must take to behave ethically (*maximising*), and the agent is to regard the consequences of their actions impartially.

Although speaking to the exact nature of any future AGI would be speculative at best, I believe it uncontroversial to assert that it would have to be in some manner similarly teleological, maximising, and impartial–just as Consequentialist ethical theories are. Extrapolating from contemporary narrow artificial intelligence (which is necessarily teleological at least, as are all machines that have a function), it is intuitive to see the other two characteristics as at least plausible potential characteristics of any future AGI.

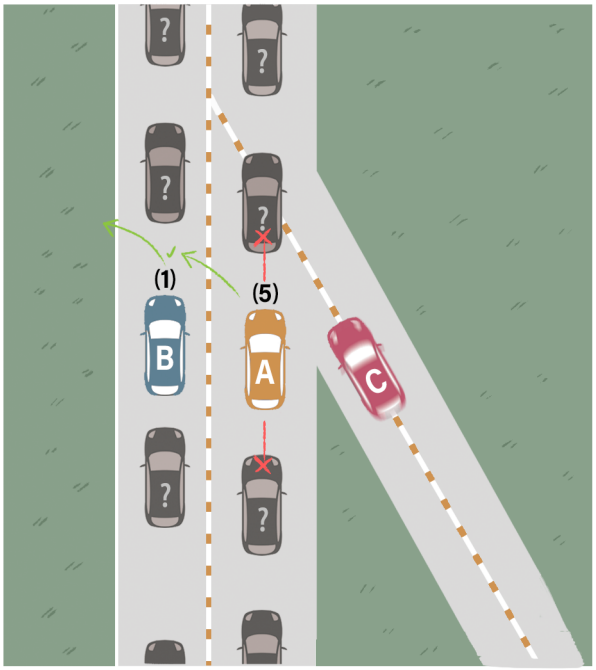
Perhaps the most salient relation Consequentialist-type ethical theories have in regards to AMAs is the inherently algorithmic nature of Consequentialism. For example, Anderson et. al. have demonstrated how Hedonistic Act Utilitarianism can be implemented to compute

4. Although the many competing tokens of Consequentialism (Act Utilitarian, Rule Utilitarian, Satisficing Consequentialist, to name a few) all have differing notions of what will ultimately count as "best consequences," Consequentialism as a kind of ethical theory is now sufficiently outlined for the purposes of this paper.

right action based on best consequences in a way that is consistent, complete and practical.[5] The computation of utility by a machine intelligence is more reliable than human moral arithmetic could ever be, as would its impartiality–both critical features of Consequentialism generally.

3 An Autonomous Vehicle 'Trolley-Problem'

To demonstrate the potential output of a Consequentialist foundation, consider the following scenario. There exist two vehicles travelling on a crowded motorway at high speed (A and B). Both vehicles are autonomously driven, they are aware of how many occupants they have, and are networked (can communicate with one another). Vehicle A in the right lane has five occupants, while vehicle B in the left lane has but one. The highway is populated with other non-autonomous vehicles such that both autonomous vehicles have many other vehicles in front of and behind them travelling at relative speed.



Suddenly, a non-autonomous vehicle (C) careens onto the crowded highway, ensuring a

5. Anderson, Anderson, and Armen 2005, 2.

high-speed collision will ensue between it and vehicle A at minimum. Indeed, the relative speeds of the vehicles in question are such that only a machine intelligence would be able to process information fast enough to make a decision of any consequence in what will soon become an ethical dilemma. In the seconds that vehicle A has before a fatal collision occurs, it begins moral deliberation to come to an ethically obligatory decision.

Vehicle A can rapidly accelerate or decelerate to escape collision with C, though this would endanger an unknown number of other lives in other vehicles, potentially starting a chain of further collisions. Label this decision S_1. It can take no evasive action, but this would assure a side-collision with C and would similarly endanger an unknown number of lives in the other vehicles behind A. Label this decision S_0. The critical injury of at least five individuals seems to be assured in either case, if not for a third option.

Recall that vehicle A and B are both fundamentally Consequentialist, and they are networked. If B were to run itself and its single occupant off of the road, then this would make room for A to swerve into the now-vacated left-most lane. A's collision with C would be avoided, and its five occupants spared harm. The single occupant of B would be sacrificed to not only save the five, but also avoid a multiple collision. Seeing as this joint decision between A and B results in the best state of affairs within their power, this course of action is morally obligatory and is therefore enacted by the two vehicles. Label this decision S_2.

Decision Label	**Agents Affected**	**Agents Harmed**
S_0	>5	>5
S_1	>5	>5
S_2	≥ 6	≥ 1

The AMA is, at minimum, concerned with the impact its decision has on all agents and those agents that will come to foreseeable harm. The lower the integer of either category, the better. This is an oversimplification, of course, since probability theory would inform all decision-making, and there would be other variables to consider, but the crucial point here is that there is always an unambiguous right action under Consequentialism which ought to

be effectuated that is computationally derivable.[6]

Again, the scenario here discussed is simplistic, but it is revealing. Consider the possibility that instead of vehicle A and B possessing the same Consequentialist ethic, Vehicle B as a rule always protected its occupants from harm, since–from a marketing perspective, say–partiality was more appealing to the consumer than a commitment to impartiality. The greater number would not be spared harm, since B would hold as a maxim to protect its occupant in all situations. Vehicle B would remain in place, forcing A to choose either S_0 or S_1. State of affairs are optimised if and only if (and because) *all* AMAs share a commitment to an impartial Consequentialism.

4 Challenges and Alternatives to Consequentialism

I am asserting that Consequentialist-type ethical theories are the most plausible to serve *as a basis* for a machine ethic. This is to say that, at minimum, for a machine to be ethical will entail a commitment to some token normative theory informed by Consequentialism in kind. Does this, however, leave open the possibility that certain features of other ethical theories (Deontological, Rossian Moral Pluralistic, Aretaic, to name a few) may find an important place in some 'complete' machine ethic on top of a Consequentialist foundation? Perhaps an AGI–just like a human intelligence–should look beyond the monism of Consequentialism in less morally fraught scenarios and lean on other articulated ethical frameworks to guide its behaviour? Perhaps the maximising and impartial nature of Consequentialism should only apply in scenarios of crisis, but Kantian maxims ought to guide AMA action otherwise, for example.

While I cannot definitively affirm or deny the validity, let alone the soundness, of such a suggestion, I do think that it introduces an additional layer of complexity in an already complex project. For instance, it seems to me that this thinking entails a theoretic com-

6. If S_2 were unavailable, then the AMA would either choose S_0 or S_1 depending on which would probabilistically endanger the least amount of agents. Even if the difference in probability is marginal, there will always be an unambiguous action for the AMA to effectuate.

mitment to a kind of Moral Particularism, which is itself controversial and may undermine the consistency we are striving for in the development of a foundational machine ethic. I believe that a foundational machine ethic would be at least computationally intelligible and consistent (read: predictable), and this latter feature seems to entail a Moral Generalism concerning any plausible machine ethic. So, if there is to be a plausible foundational machine ethic, it will strictly be *an* ethic–Consequentialist, Deontological, Aretaic, to name a few, and not interplay between some or all of ethics.

Thus, I now examine an alternative foundational ethic. Consider the classic rival to Consequentialism: Deontology. Powers argues that the first formulation of Kant's *Categorical Imperative* can be used to generate rules of right action which would inherently be computationally intelligible due to their purely formal (logical) nature. Instead of relying on an arithmetic approach and the empirical demands entailed by a commitment to "good consequences," a Deontological approach would entail testing against systematicity and universalisability for every circumstance, purpose and action an AMA might encounter according to a formalised deontic logic.[7]

I think that this Deontological approach is very promising. It is computationally intelligible and theoretically consistent, satisfying what I have argued are two minimal conditions for a plausible foundational machine ethic. However, and summarising substantially, this Kantian approach which relies on nonmonotonic logic fails to account for supererogation, cannot account for semidecidability of set membership (an important feature of first-order logic), and is typified by excessive specificity (not conducive to generalisability). Most detrimentally, though, it has no mechanism for prioritising one maxim (prescribed course of action) over another when two or more conflict.[8]

While most of the challenges faced by this Deontological approach do not disqualify it outright, I believe that its failure to account for moral conflict–a situation where two equally weighty moral precepts apply–does. An ethical theory is supposed to be, among other things,

7. Powers 2006, 46-47.
8. *ibid.*, 49-51.

action guiding. When a theory fails to prescribe an action, it fails in this vital sense, and is thus demonstrably lacking in an important regard. The problem of moral conflict is, of course, a general issue in normative ethics for all normative theories which are not monistic, but its consequences for AMA behaviour are particularly concerning.

What is an AMA to do when its action guiding foundation offers it no prescription? In narrow implementations, does the AMA throw an exception and simply cease operation? Perhaps, since that might be warranted in supervised domains where a human intelligence can be called upon, but this (in)action seems to fail in unsupervised domains where the performance of right action is arguably even more crucial. Or, does the AMA simply continue towards its goal as if no moral dilemma was encountered at all? Indeed, how a simple software program (let alone a narrow machine intelligence) treats an encountered error is itself as much of an ethical concern as it is a technical one when people and machines interact.

When it comes to a general machine intelligence–a level of artificial intelligence where a lack of direct human supervision is entailed–perhaps a novel bottom-up solution will be generated if and when a scenario that entails moral conflict arises. After all, AGI is an intelligence indistinguishable from our own, and human beings do not throw errors when confronted with a moral challenge. Yet, as the designers of AGI, I think it uncontroversial to suggest that while it may be fascinating to see what a given intelligence might decide to do when its literal moral code fails to offer a decisive action, such novelty is by definition unpredictable, and therefore the AMA fails to be consistent.

Ultimately, though, an AGI is a hypothetical entity, so I can only speculate. The most I will suggest here is that the designers of AGI ought not set up its AMA to fail by giving it what might be described as an incomplete ethic: one which cannot account for moral conflict. Why have as a design feature uncertainty, and in such a crucial domain as ethical behaviour? The good AMA is one which, when confronted with a moral dilemma, adheres to an ethical theory which has it return a morally obligatory action every time.

What is required is the encoding of an ethic that, if it cannot *avoid* moral conflict, can at

least resolve it predictably. Even in conceivably narrow implementations of AI–such as the autonomous vehicle example above–stopping and calling for help, or ignoring the dilemma entirely, would result in the harm of many. Since moral conflict avoidance is so demonstrably important, and the monism of Consequentialism ensures that such cannot arise in the first place, this further indicates that a Consequentialist theory would be a sound foundation for a machine ethic.

Of course, the Consequentialist approach is not without its own theoretical problems. One of the more powerful objections is the suggestion that the set of empirical data required by Consequentialism to come to a moral verdict would be prohibitively large. So large, in fact, that it would require the AMA to be virtually omniscient.[9] I believe this is overselling the size of the set of what is morally relevant–that which would be required for a moral decision. Bearing in mind that when we are designing an AMA, the goal first and foremost is to have an agent who is ethically greater than or equal to a human moral agent. Humans are far from ideal moral agents, yet they can be perfectly consistent welfare-Consequentialists even though they are further from omniscience than a machine.

Like humans, AMAs would be capable of adopting heuristics to come to decisions. "One kind of ethical heuristic might be to follow rules that are expected to increase *local* utility. For instance, rather than analyse all the consequences of her possible actions, a person may choose between them solely on the basis of benefits to her local community."[10] Applied to AMAs, this line of reasoning would see the features of its immediate environment privileged over historical or anticipated environmental states, with distant consequences in future times heavily discounted, for example. Beyond this, though, any further objections would require treatment by whichever token Consequentialist moral theory is ultimately chosen.

9. Wallach 2009b, 88.
10. *ibid.*, 90.

5 Bottom-Up Approaches

In what other ways can an AGI be made to think ethically? Thus far, we have considered two major top-down approaches to encoding morality, but we will now briefly consider what I believe to be the two most promising bottom-up approaches I am aware of. Even though we lack full knowledge of how human agents come to develop a moral character as they mature, this need not disqualify its inquiry nor its potential applications in machine ethics. One bottom-up approach simulates evolutionary pressure on artificial entities in the hopes that each successive generation of entities evolve and refine complex behaviours from initially simple ones. 20th century developments in sociobiology and mathematics revealed that, under certain environmental conditions (such as those informed by iterated game-theoretic situations like the Prisoner's Dilemma), cooperative behaviour could be conducive to successful genetic propagation intraspecially and could therefore emerge evolutionarily.[11]

The upshot is that complex ethical behaviour might be capable of emerging evolutionarily given the proper evolutionary pressures. In this way, AMAs would develop morality through iterative interactions with other AMAs without the need of some top-down encoding of a moral theory. The suggestion is that the intelligence required to successfully participate in iterated game-theoretic environments is all that is required for the emergence of morality. So, what would begin as a simple choice to cooperate or not would, over time, lead to complex ethical reasoning. However, it is wholly unclear if mere evolutionary pressures in a simulated environment are enough to result in ethical agents, let alone ethical agents that are guided consistently enough for their behaviour to be foundational.[12]

Another bottom-up approach recognises that human morality is–at least to some degree–learned, and if it can be learned, then it can be taught. Much like how a child learns to read, perhaps an AGI might learn to be an AMA through varieties of reinforcement learning and case training. Marcello Guarini, for example, outlined his attempt to use

11. Wallach 2009a, 101-102.
12. *ibid.*, 103.

different neural network configurations to classify moral propositions as either permissible or impermissible.[13] While learning approaches are theoretically promising, and have solid applications in other domains, they have yet to result in anything resembling a foundational ethical theory. Perhaps all that is required is more data, and more time. Common to both of these bottom-up strategies, though, are their difficulty and uncertainty. More work needs to be done before any bottom-up approach can be counted on to supply a consistent and plausible basis for a machine ethic.

6 Conclusion

Recall the two fundamental theses which underly Consequentialism. Value-independence and the explanatory priority of value. The former allows an AMA to offer a value judgement on the state of its environment without appeal to the concept of right action. What is of concern are probabilistic outcomes of actions. The latter ensures the AMA knows when it has behaved ethically since what makes an action right is given solely in terms of the value of the consequences that are related to the action. Thus, a Consequentialist AMA would see that an act is right if and only if (and because) that act would result in the best state of affairs of all available alternatives within its power, impartially.

The two theses of Consequentialism are computationally intelligible and offer theoretically consistent moral verdicts. Moreover, the monism of Consequentialism escapes the thorny issue of moral conflict. I believe that top-down approaches are still the most promising towards the development of an AMA, though bottom-up approaches may nevertheless play a role. Of the set of normative theories, I submit that Consequentialist-type theories are the most plausible to serve as a basis for a machine ethic precisely because they can be counted on to come to a moral conflict-free verdict based on computationally intelligible principles of welfare-maximisation in a consistent manner.

13. Guarini 2011, 321-328.

References

Allen, Colin, Gary Varner, and Jason Zinser. 2000. "Prolegomena to any Future Artificial Moral Agent." *Journal of Experimental & Theoretical Artificial Intelligence* 12, no. 3 (July): 251–261. doi:10.1080/09528130050111428. http://www.tandfonline.com/doi/abs/10.1080/09528130050111428.

Anderson, Michael, and Susan Leigh Anderson. 2018. "GenEth: a General Ethical Dilemma Analyzer." *Paladyn, Journal of Behavioral Robotics* 9, no. 1 (November): 337–357. ISSN: 2081-4836. doi:10.1515/pjbr-2018-0024. http://www.degruyter.com/view/j/pjbr.2018.9.issue-1/pjbr-2018-0024/pjbr-2018-0024.xml.

Anderson, Michael, Susan Leigh Anderson, and Chris Armen. 2005. "Toward Machine Ethics: Implementing Two Action-Based Ethical Theories." In *Machine Ethics: Papers from the AAAI Fall Symposium. Technical Report FS-05-06,* 7. Association for the Advancement of Artificial Intelligence, Menlo Park, CA.

Anderson, Susan Leigh. 2011. "Philosophical Concerns with Machine Ethics." In *Machine Ethics,* edited by Michael Anderson and Susan Leigh Anderson, 162–167. Cambridge: Cambridge University Press. ISBN: 978-0-511-97803-6. doi:10.1017/CBO9780511978036.014. https://www.cambridge.org/core/product/identifier/CBO9780511978036A022/type/book_part.

Bostrom, Nick, and Eliezer Yudkowsky. 2014. "The Ethics of Artificial Intelligence." In *The Cambridge Handbook of Artificial Intelligence,* edited by Keith Frankish and William M. Ramsey, 316–334. Cambridge: Cambridge University Press. ISBN: 978-1-139-04685-5. doi:10.1017/CBO9781139046855.020. https://www.cambridge.org/core/product/identifier/CBO9781139046855A027/type/book_part.

Ford, Martin R. 2018. "Architects of Intelligence: the truth about AI from the people building it." (Birmingham, UK).

Guarini, Marcello. 2011. "Computational Neural Modeling and the Philosophy of Ethics." In *Machine Ethics,* edited by Michael Anderson and Susan Leigh Anderson, 316–334. Cambridge: Cambridge University Press. ISBN: 978-0-511-97803-6. doi:10.1017/CBO9780511978036.023. https://www.cambridge.org/core/product/identifier/CBO9780511978036A032/type/book_part.

McLaren, Bruce M. 2011. "Computational Models of Ethical Reasoning." In *Machine Ethics,* edited by Michael Anderson and Susan Leigh Anderson, 297–315. Cambridge: Cambridge University Press. ISBN: 978-0-511-97803-6. doi:10.1017/CBO9780511978036.022. https://www.cambridge.org/core/product/identifier/CBO9780511978036A031/type/book_part.

Powers, T.M. 2006. "Prospects for a Kantian Machine." *IEEE Intelligent Systems* 21, no. 4 (July): 46–51. ISSN: 1541-1672. doi:10.1109/MIS.2006.77. http://ieeexplore.ieee.org/document/1667953/.

Santos-Lang, Christopher Charles. 2015. "Moral Ecology Approaches to Machine Ethics." In *Machine Medical Ethics,* edited by Simon Peter van Rysewyk and Matthijs Pontier, 74:111–127. Cham: Springer International Publishing. ISBN: 978-3-319-08107-6 978-3-319-08108-3. doi:10.1007/978-3-319-08108-3_8. http://link.springer.com/10.1007/978-3-319-08108-3_8.

Sullins, John P. 2011. "When Is a Robot a Moral Agent?" In *Machine Ethics,* edited by Michael Anderson and Susan Leigh Anderson, 151–161. Cambridge: Cambridge University Press. ISBN: 978-0-511-97803-6. doi:10.1017/CBO9780511978036.013. https://www.cambridge.org/core/product/identifier/CBO9780511978036A021/type/book_part.

Timmons, Mark. 2013. *Moral Theory: An Introduction.* 2nd ed. Elements of Philosophy. Lanham, Md: Rowman & Littlefield Publishers. ISBN: 978-0-7425-6491-6.

Wallach, Wendell. 2009a. "Bottom-Up and Developmental Approaches." In *Moral Machines,* 100–116. Oxford University Press, February. ISBN: 978-0-19-537404-9. doi:10.1093/acprof:oso/9780195374049.001.0001. http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780195374049.001.0001/acprof-9780195374049.

———. 2009b. "Top-Down Morality." In *Moral Machines,* 84–98. Oxford University Press, February. ISBN: 978-0-19-537404-9. doi:10.1093/acprof:oso/9780195374049.001.0001. http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780195374049.001.0001/acprof-9780195374049.

Whitby, Blay. 2011. "On Computable Morality." In *Machine Ethics,* edited by Michael Anderson and Susan Leigh Anderson, 138–150. Cambridge: Cambridge University Press. ISBN: 978-0-511-97803-6. doi:10.1017/CBO9780511978036.012. https://www.cambridge.org/core/product/identifier/CBO9780511978036A020/type/book_part.