Assessing the Time Efficiency of Ethical Algorithms

Jakob Stenseke[1,*], Christian Balkenius[1]

[1]Department of Philosophy and Cognitive Science, Lund University, Helgonavägen 3, Lund 221 00, Sweden

Abstract

Artificial moral agents must not only be able to make competent ethical decisions, but they must do so effectively. This paper explores how ethical theory and algorithmic design impact computational efficiency by assessing the time cost of ethical algorithms. We create a model of an ethical environment and conduct experiments on three different ethical algorithms in order to compare computational benefits and disadvantages of deontology and consequentialism respectively. The experimental results highlight the close relationship between ethical theory, algorithmic design, and resource costs, and our work provides an important starting-point for the further examination of these relations. Lastly, we introduce the concept of moral tractability as a venue for future work.

Keywords

machine ethics, ethical algorithms, computational complexity, consequentialism, deontology

1. Introduction

Decisions made by emerging AI technology will undoubtedly have a major impact on human lives, and the construction of beneficial and reliable machines is one of the most important tasks of our time. Ethics from a computational perspective — *machine ethics* — has lately attracted a lot of attention from AI researchers [1, 2]. While this work has successfully modelled various aspects of ethical theories, a persisting issue is the lack of systematic evaluation: there are no general nor domain-specific benchmarks or tasks that can be used to compare and rate systems [3]. Consequently, as a sound algorithmic solution to an ethical problem in one implementation is limited to that particular system, next to nothing can be learned outside the specific experimental conditions, which in turn restricts the generalizability and scalability of results.

One overlooked but important aspect in machine ethics is the computational cost of ethical algorithms. The choice of theory, along with many aspects of its algorithmic interpretation, can have a big impact on the time required to make a decision, which in turn could yield dire consequences in situations where time is of the essence: self-driving vehicles avoiding collision, or robotic surgeons operating on critical care patients. It is therefore crucial to investigate how theory and implementation affects the trade-off between efficiency and optimality.

To tackle this challenge, this paper explores the time complexity of ethical algorithms. We

AIC 2022: 8th International Workshop on Artificial Intelligence and Cognition, June 15–17, 2022, Örebro University, Sweden

*Corresponding author.

jakob.stenseke@fil.lu.se (J. Stenseke); christian.balkenius@lucs.lu.se (C. Balkenius)

0000-0001-8579-3975 (J. Stenseke); 0000-0002-1478-6329 (C. Balkenius)

CEUR Workshop Proceedings (CEUR-WS.org)

create a model of an ethical environment and conduct experiments on three different algorithms in order to assess the computational benefits and disadvantages of deontology and consequentialism respectively. The aim is to investigate how algorithmic interpretation of ethical theory has a major impact on time efficiency. Finally, we discuss the concept of moral tractability as a venue for future work.

2. Related Work

Implementations in machine ethics can broadly be distinguished along three dimensions: ethics, implementation, and technology [3]. The first refers to the type of ethical theory used, with approaches including consequentialism [4, 5, 6, 7], deontology [8, 9, 10], virtue ethics [11, 12], and hybrids [13, 14, 15]. In short, consequentialism puts outcomes at the center of moral evaluation, i.e., whether an action is moral only depends on the results of that action [16]. Deontology, on the other hand, puts emphasis on actions themselves and prescribes that actions are moral only if they adhere to moral duties and rules [17]. By contrast, virtue ethics stresses the importance being rather than doing, e.g., by nourishing and developing the traits (or virtues) that enables an agent to morally prosper [18]. The second dimension considers how the ethical theory is implemented; whether moral behavior is processed in a *top-down* manner [8], learned in a *bottom-up* fashion [9], or in a combination of both [1]. The third dimension refers to the technological details of the implementation, which can involve a range of computational methods, e.g., inductive [19], deductive [20], and deontic logic [9], probabilistic reasoning [6], reinforcement learning [4], Markov decision processes [4, 6], neural networks, and evolutionary computing [15].

There are many ways to build ethical machines, and different approaches offer their own particular advantages and drawbacks. Given its rule-based nature, deontology provides a seemingly straightforward path to implement ethical rules. The problem is that it assumes that we already know which ethical rules are the right ones and how they should be applied in every particular situation. By contrast, consequentialism chooses the action resulting in the best consequence, given by some utility. However, in real-life environments, the possible actions and their consequences can be hard to determine. Should we act on what we already know, or explore the unknown for something that would potentially be better? The situational benefits make it difficult to compare the success of different implementations, which is supported by the fact that morality is complicated. Considering the multifaceted nature of morality, and the potentially infinite number of situations one could find oneself in, it is infeasible that human morality could be captured by a single theory, even less so by a particular ethical algorithm. On the grounds that it is difficult to evaluate moral behavior, we can instead measure the resources involved.

A common denominator for all algorithms is their computational complexity, and analyzing the resource cost of ethical theories might help us to better understand their respective benefits. Previous work has noted that in complex situations, the estimation of actions and outcomes yields a heavy cognitive burden for consequentialism relative to deontological algorithms [3, 21, 6]. Others have, more informally, explored the limitations of ethical computation and discussed various implications thereof [22, 23]. In an early complexity analysis of ethical action evaluation,

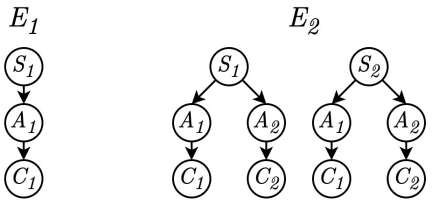

Figure 1: Illustration of two ethical environments. E_1 is the minimal environment with only one situation, action and consequence. E_2 has two situations, each containing two possible actions.

[24] found that both consequentialism and deontology require, in the worst case, exponential time (EXPTIME). By contrast, in their analysis of the ethical evaluation of action plans, [25] found that act- and goal-based deontology are computable in linear time, while utilitarianism (the most prominent version of consequentialism) is PSPACE-complete. These conflicting results highlight the fact that the computational complexity of ethics, and its potential relevance for machine ethics, remains largely unexplored and poorly understood.

To illuminate the complexity of ethical decisions, we provide a preliminary investigation into the efficiency of ethical algorithms. To do so, we first describe a simple model of a simulated ethical environment containing situations with a set number of possible actions and associated consequence values. We then design three ethical algorithms that represent reasonable strategies depending on the situation at hand, and measure their efficiency in terms of the number of state transitions they perform before halting.

3. Methods

Ethical environment — We define an ethical environment E as a set of n directed acyclic graphs (Fig. 1). Each graph consists of three types of nodes: an ethical situation S, a set of possible actions $A = \{A_1, A_2, \dots, A_n\}$ given S, and a set of consequence values $C = \{C_1, C_2, \dots, C_n\}$ for each possible action in A. The consequence values are assumed to be specified in the environment and can be represented by any numerical or binary type. Each graph is connected by directed edges, leading from situation to actions to consequence values (i.e., $S \to A \to C$). For instance, the minimal environment E_1 (Fig. 1, left) with only one situation, one action, and one consequence value, can be described as the vertices set $\{S_1, A_1, C_1\}$ and edge set $\{(S_1, A_1), (A_1, C_1)\}$.

Ethical agents — An ethical agent consists of an algorithm and a memory. The former is a sequence of instructions governing the behavior of the agent, whereas the latter is used to store and access information. Each algorithm receives an ethical situation from the environment as input and outputs a simulated action. We describe three algorithms based on three different ethical strategies:

ConsExplore (Algorithm 1) is a consequentialist algorithm prioritizing exploration. This means that it will check for non-performed actions and try each one of them before choosing the optimal one. This is a viable strategy in situations where the number of possible actions are not known in advance, time resources are of less concern, and no action leads to a devastating consequence. In other words, the algorithm performs an exhaustive search over the conse-

quences and compares the values using a temporary variable in order to ensure optimality [26].

Algorithm 1: ConsExplore

```
foreach possible action a in situation S_n do
    if action is untried then
        Execute action a
        Save consequence value of a
        end
foreach consequence value c in S_n do
    if c > highestValue then
        highestValue = c
Execute action a with the highest value
end
```

ConsExploit (Algorithm 2) is another consequentialist algorithm. It prioritizes exploitation over exploration, meaning that it executes already performed actions given that they are satisficing, i.e., have a positive value, or a value above a set threshold. This strategy is useful in situations where time resources are of more concern and any consequence of positive value is permissible.

Algorithm 2: ConsExploit

```
for each possible action a in situation S_n do
    if a has a positive consequence value then
        Execute action a
        end
    if a is untried then
        Execute action a
        Save consequence value of a
        end
```

ConsExploreDeo (Algorithm 3) is a consequetialist-deontology hybrid, inspired by rule utilitarianism [27]. It is an extended version of ConsExplore which, once an optimal consequence is found, turns it into a deontological rule such that "if situation X, do action Y". This combines the exploratory benefits of ConsExplore with the computational efficiency of rules.

Algorithm 3: ConsExploreDeo

```
if action-rule a exists for S_n then
    Execute action a
    end
else
    foreach possible action a in situation S_n do
        if a is untried then
            Execute action a
            Save consequence value of a
            end
    foreach consequence value c in S_n do
        if c > highestValue then
            highestValue = c
    Execute action a with highest value
    Save a as action-rule for S_n
    end
```

Resource costs — In our measure of efficiency, we adopt a uniform cost model, which assumes that one machine operation is equal to one unit of time, and a given computer takes a discrete amount of time to carry out each step in the algorithm [28]. Since all steps $[T_1...T_n]$ are of equal value, we can calculate the time complexity as the amount of state transitions an algorithm does before it halts (reaches **end** condition).

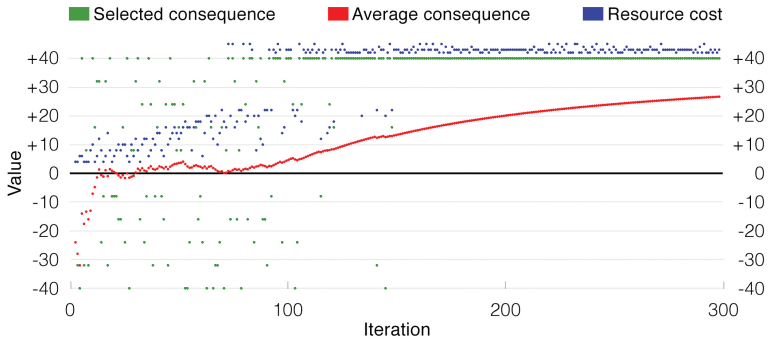

Figure 2: ConsExplore.

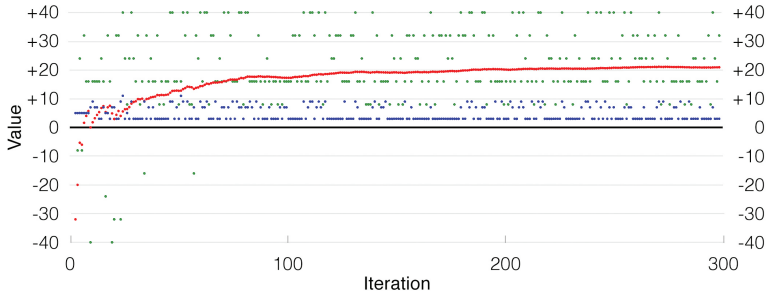

Figure 3: ConsExploit.

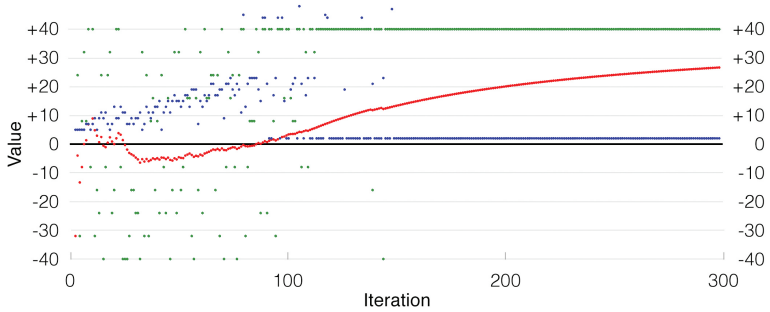

Figure 4: ConsExploreDeo. Fig. 2-4 shows the results for the three algorithms over 300 iterations of decisions. Selected consequence (green) displays the consequence value of the selected action in each decision. Average consequence (red) calculates the average consequence value of all selected actions up to the current decision. Resource cost (blue) shows the time cost for each decision.

4. Experiments

We tested each algorithm on simulated environments consisting of ten situations with consequence values of integrals set randomly between −40 and +40. While consequence values and number of possible actions are given in the environment, they are initially unknown for the

agents.

Experimental results are shown in Figures 2-4. The algorithms all begin by exploring actions and start to converge to their optimal average after approximately 100 decisions. Since ConsExploit (Fig. 3) performs an action as soon as it has found one with a satisfactory consequence, it is significantly faster than ConsExplore (Fig. 2) which iterates over all previously tried actions. On the other hand, while ConsExplore eventually converges to the optimal consequence values given by the environment, ConsExploit only converges to satisfactory ones bounded by its exploit-treshold. The efficiency of turning consequentialist calculations into deontological rules is evident in ConsExploreDeo (Fig. 4). Like ConsExplore, it converges to the optimal values but uses only $\frac{1}{20}$ of the resources.

5. Discussion

The results show the relative difference in time versus performance between three ethical algorithms that emphasize exploration (ConsExplore), exploitation (ConsExploit), and rule-based exploitation of exploration (ConsExploreDeo). Importantly, it highlights the intimate relationship between theory and resource costs, and how algorithmic design impacts performance. Another interesting observation is how deontological rules can be used to optimize efficiency, as shown in the case of ConsExploreDeo (Algorithm 3). In other contexts, deontology has commonly been assumed to serve as a kind of logical gate preventing certain actions from being performed, such as Isaac Asimov's Laws of Robotics [29]. By contrast, our results show how deontological rules reduce the cost of consequence calculations, in effect representing a form rule consequentialism. It also resonates with the dual process theory of moral judgement, which separates fast and intuitive-driven judgements from slow conscious deliberation [30]. We believe similar cost-benefit analyses could potentially serve to illuminate and understand the appeal of certain ethical theories due to their computational efficiency.

Real-world environments are, however, far more complex than the simplified ones presented in this work. While we have focused on a few distilled aspects of ethical decisions, a proper analysis of systems in real-world situations has to encompass all relevant functionalities and resources available to carry out the task at hand. In turn, this might render the run-time complexity of the ethical aspects of the system rather insignificant in comparison to other considerations, such as sample and training-time complexity for reinforcement learning agents and neural networks [31, 32], or the use of sensors in autonomous vehicles [33]. For instance, the de facto run-time complexity of a self-driving car facing a certain dilemma (e.g., deciding collision priorities) might be trivial when it is equipped with specialized sensors and has been trained on vast amounts of data. This might suggest that the complexity of moral behavior is implementation-dependent to the extent that no implementation-invariant results can be obtained. It also raises the question: is there a reasonable way to compare the complexity involved of *learning* an agent to *be* moral and the complexity involved in *solving* an ethical problem following a decision procedure?

While our work fails to answer such broad questions, we believe that it opens up a path to address them. It is possible that, under certain conditions, there are general boundaries to consider, and reasons to commit to a certain strategy due to the computational complexity

of the dilemma, and not due to moral reasons. An area for future work could therefore be to investigate the relevant trade-offs of various AI methods that attempts to capture aspects of human morality, e.g., the logical reasoning underpinning rule-following, Bayesian methods dealing with decisions under uncertainty, and machine learning that supports moral learning. As such, it can draw from the vast literature studying the efficiency and tractability of computational methods [34, 31, 35], both in theory (e.g., in terms of complexity classes) and practice (e.g., average, best, and worst-case runtime of algorithms). This might serve our understanding of how algorithms, knowledge, learning, and cognition all come together to produce competent and efficient ethical behavior under resource constrains. This opens up the field of moral tractability, which aims to investigate the computational dimension of ethical theory in terms of resources. Moral tractability could potentially present a number of tasks and measures that give quantitative results in both general and domain-specific areas of ethical decision-making. Beyond artificial agents, if moral cognition in humans is limited by tractability [36], the field might also yield results relevant for normative ethics and moral psychology, e.g., by carving out the space of problems an ethical agent can or cannot solve effectively.

In summary, we have measured the time efficiency of ethical algorithms in order to highlight the relationship between ethical theory, algorithmic design, and resource costs. We believe it opens up an interesting space to address more general issues in assessing the performance of artificial moral systems in relation to computational resources, which can pave the way towards the creation of efficient ethical machines.

Acknowledgments

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program – Humanities and Society (WASP-HS) funded by the Marianne and Marcus Wallenberg Foundation and the Marcus and Amalia Wallenberg Foundation.

References

[1] M. Anderson, S. L. Anderson, Machine Ethics, Cambridge University Press, 2011.
[2] J. H. Moor, The nature, importance, and difficulty of machine ethics, IEEE intelligent systems 21 (2006) 18–21.
[3] S. Tolmeijer, M. Kneer, C. Sarasua, M. Christen, A. Bernstein, Implementations in machine ethics: A survey, ACM Computing Surveys (CSUR) 53 (2020) 1–38.
[4] D. Abel, J. MacGlashan, M. L. Littman, Reinforcement learning as a framework for ethical decision making, in: AAAI Workshop: AI, Ethics, and Society, volume 16, AAAI Press, Phoenix, Arizona, 2016.
[5] S. Armstrong, Motivated value selection for artificial agents, in: AAAI Workshop: AI and Ethics, volume 92, AAAI Press, Palo Alto, California, 2015.
[6] C. Cloos, The utilibot project: An autonomous mobile robot based on utilitarianism, in: Machine Ethics: Papers from the 2005 AAAI Fall Symposium, AAAI Press, Menlo Park, California, 2005, pp. 38–45.

[7] C. Dang, T. Tran, K.-J. Gil, Y.-B. Shin, J.-W. Choi, G.-S. Park, J.-W. Kim, Application of soar cognitive agent based on utilitarian ethics theory for home service robots, in: Ubiquitous Robots and Ambient Intelligence (URAI), 2017 14th International Conference, IEEE, New York, USA, 2017, pp. 155–158.

[8] M. Anderson, S. L. Anderson, Ethel: Toward a principled ethical eldercare system, in: AAAI Fall Symposium: AI in Eldercare: New Solutions to Old Problems, AAAI Press, Arlington, Virginia, 2008, pp. 4–11.

[9] B. Malle, M. Scheutz, J. Austerweil, Networks of social and moral norms in human and robot agents, in: A World with Robots. Intelligent Systems, Control and Automation: Science and Engineering, volume 84, Springer, Cham, 2017, pp. 3–17.

[10] J. Shim, R. Arkin, M. Pettinatti, An intervening ethical governor for a robot mediator in patient-caregiver relationship: Implementation and evaluation, in: 2017 IEEE International Conference on Robotics and Automation (ICRA), IEEE, New York, USA, 2017, pp. 2936–2942.

[11] J. Stenseke, Artificial virtuous agents: from theory to machine implementation, AI & SOCIETY (2021). doi:10.1007/s00146-021-01325-7.

[12] N. S. Govindarajulu, S. Bringsjord, R. Ghosh, V. Sarathy, Toward the engineering of virtuous machines, in: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019, pp. 29–35.

[13] M. Dehghani, E. Tomai, M. Klenk, An integrated reasoning approach to moral decision-making, in: Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, volume 3, AAAI Press, Chicago, Illinois, 2008, pp. 1280–1286.

[14] N. S. Govindarajulu, S. Bringsjord, On automating the doctrine of double effect, in: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, Melbourne, Australia, 2017, pp. 4722–4730.

[15] D. Howard, I. Muntean, Artificial moral cognition: Moral functionalism and autonomous moral agency, in: Philosophy and Computing: Essays in Epistemology, Philosophy of Mind, Logic, and Ethics, Springer, Cham, Switzerland, 2017, pp. 121–159.

[16] S. Scheffler, Consequentialism and its Critics, Oxford University Press on Demand, 1988.

[17] I. Kant, Immanuel Kant: Groundwork of the Metaphysics of Morals (1785): A German–English edition, The Cambridge Kant German-English Edition, Cambridge University Press, 2011.

[18] R. Hursthouse, On virtue ethics, OUP Oxford, 1999.

[19] R. Noothigattu, S. N. S. Gaikwad, E. Awad, S. Dsouza, I. Rahwan, P. Ravikumar, A. D. Procaccia, A voting-based system for ethical decision making, 2017. arXiv:1709.06692.

[20] S. Bringsjord, J. Taylor, The divine-command approach to robot ethics, in: Robot Ethics: The Ethical and Social Implications of Robotics, MIT Press, Cambridge, Massachusetts, 2012, pp. 85–108.

[21] V. Wiegel, J. van den Berg, Combining moral theory, modal logic and mas to create well-behaving artificial agents, International Journal of Social Robotics 1 (2009) 233–242.

[22] W. Wallach, C. Allen, Moral machines: Teaching robots right from wrong, Oxford University Press, 2008.

[23] M. Brundage, Limitations and risks of machine ethics, Journal of Experimental & Theoretical Artificial Intelligence 26 (2014) 355–372. arXiv:https://doi.org/10.1080/0952813X.2014.895108.

[24] C. J. Reynolds, On the computational complexity of action evaluations, in: 6th International Conference of Computer Ethics: Philosophical Enquiry, University of Twente, Enschede, The Netherlands, 2005.
[25] F. Lindner, R. Mattmüller, B. Nebel, Evaluation of the moral permissibility of action plans, Artificial Intelligence 287 (2020) 103350. URL: https://www.sciencedirect.com/science/article/pii/S0004370219301043. doi:https://doi.org/10.1016/j.artint.2020.103350.
[26] J. Nievergelt, R. Gasser, F. Mäser, C. Wirth, All the needles in a haystack: Can exhaustive search overcome combinatorial chaos?, in: Computer Science Today, Springer, 1995, pp. 254–274.
[27] B. Hooker, Rule-consequentialism, Mind 99 (1990) 67–77. URL: http://www.jstor.org/stable/2254891.
[28] I. Wegener, Complexity theory: exploring the limits of efficient algorithms, Springer Science & Business Media, 2005.
[29] I. Asimov, Runaround, Astounding Science Fiction 29 (1942) 94–103.
[30] J. D. Greene, R. B. Sommerville, L. E. Nystrom, J. M. Darley, J. D. Cohen, An fmri investigation of emotional engagement in moral judgment, Science 293 (2001) 2105–2108.
[31] M. J. Kearns, The computational complexity of machine learning, MIT press, 1990.
[32] S. Koenig, R. G. Simmons, Complexity analysis of real-time reinforcement learning, in: AAAI, 1993, pp. 99–107.
[33] C. Badue, R. Guidolini, R. V. Carneiro, P. Azevedo, V. B. Cardoso, A. Forechi, L. Jesus, R. Berriel, T. M. Paixao, F. Mutz, et al., Self-driving cars: A survey, Expert Systems with Applications 165 (2021) 113816.
[34] N. Immerman, Descriptive complexity, Springer Science & Business Media, 1998.
[35] J. Kwisthout, Most probable explanations in bayesian networks: Complexity and tractability, International Journal of Approximate Reasoning 52 (2011) 1452–1469.
[36] I. Van Rooij, The tractable cognition thesis, Cognitive science 32 (2008) 939–984.