On Testing Ethical Autonomous Decision-Making

Michael E. Akintunde[1], Martim Brandão[1], Gunel Jahangirova[1], Hector Menendez[1], Mohammad Reza Mousavi[1], and Jie Zhang[1]

King's College London, London, United Kingdom
{michael.akintunde,martim.brandao,gunel.jahangirova,
hector.menendez,mohammad.mousavi,jie.zhang}@kcl.ac.uk

Abstract. We present an initial proposal for a testing framework for ethical decisions in autonomous agents, based on the well-known perception-action model. We identify three main components in our proposed framework for test-case generation, conformance analysis, and learning and adaption of ethical models based on examples from stakeholders. We define a number of templates formalising the main ethical theories in the literature that can be further instantiated for testing concrete systems according to such theories.

Keywords: Testing · Ethics · Autonomous Systems.

1 Introduction

Ethics, in our context, is a systematic description of principles for what is right and honourable [7,23]; examples of such behaviour include respecting the privacy of patients (of/by an agent), dealing fairly with patients, and not deceiving them in interactions. Autonomous systems are already taking decisions that are ethically charged: software that takes credit ratings for mortgages can be unfair or biased; a chatbot can be offensive and discriminatory; an assistive care robot can violate patients' privacy or make unethical decisions that damage patients' integrity.

Different stakeholders may have different ethical concerns and even subscribe to different meta-ethical frameworks. Even the opening, seemingly obvious, examples we gave above may be debated in different contexts and benefit from more scrutiny, specification, and discussion in their specific context. The diversity and ambiguity of these ethical concerns and meta-ethical frameworks have been a challenge in their system-level testing. There have been a number of recent approaches aiming to formalise specific meta-ethical frameworks for autonomous systems [24,27,41]. There are also works that focus on testing specific ethical concerns, such as fairness or bias [5,16–18]. However, we are not aware of any general framework that can be used to encode different stakeholders' ethical rules and concerns in order to test them.

Whether autonomous systems can be counted as fully-ethical agents is a philosophical concern which is outside the scope of our paper; regardless of one's

stance in this regard, it is helpful to have tools to evaluate autonomous decision-making.

Our aim is to automate system-level testing for ethical decision-making that is customisable to different meta-ethical frameworks. For this purpose, we propose a framework to: 1) generate challenging test scenarios/inputs, 2) analyse the test results via oracles, and 3) adjust the oracles through stakeholder engagement.

We do not aim to resolve disagreements among stakeholders; instead, we would like to give different stakeholders a tool to understand their ethical concerns better, use it to engage in a discussion with other stakeholders, and also test black-box or third-party autonomous systems against their concerns. For under-represented stakeholders and those with less power to scrutinise the design of such systems, we would like to provide a tool to rigorously capture their concerns and reveal any deviations from what they consider ethically significant. Our goal can hence be summarised as providing a tool for providing more transparency regarding ethical concerns in complex autonomous systems.

To illustrate the concepts presented in the remainder of this paper, we use the following scenario as our motivating example.

Motivating example. Consider an autonomous vehicle designed to drive autonomously in urban traffic. A function of this autonomous vehicle focuses on dealing with emergency vehicles such as ambulances and fire engines. Through a vehicle-to-vehicle communication method, it can learn about vehicles that are on a critical mission and their kinematics, and must react to this information. Such a function can make ethically-charged decisions, e.g., decisions that may help or harm the condition of a patient in an approaching ambulance at various costs, such as violating traffic regulations.

Considering the significant amount of studies in human ethics and the complexity of developing substantial frameworks that embed their knowledge, our work identifies the following challenges for the testing framework:

1. **Generating effective test scenarios**: there is a significant amount of scenarios, creating a wide input space for autonomous decision-making. Focusing on effective ones that are likely to reveal issues (or establish trust) is a major challenge. Another significant challenge to cope with this limitation is to define measures of effectiveness, both to steer the testing process and to evaluate and compare different techniques.
2. **Different meta-ethical frameworks**: Ethics have evolved differently in different contexts, creating a plethora of ethical frameworks. A major challenge in developing a discipline of testing is to choose one of them. However, since there is no single agreed-upon framework, this choice is significantly complex.
3. **Ethical oracles**: Even within a fixed ethical framework, defining a rigorous test oracle to judge ethical behaviours and pass verdicts about conflicting ethical concerns is a highly non-trivial challenge.
4. **Diversity and stakeholder engagement**: Developing a responsible regime of testing for ethical concerns requires interaction with a diverse population

of users. Overcoming this challenge involves gathering a truly diverse population of stakeholders (both representing the diversity in their demographics and backgrounds, but also in their roles and relationships with respect to each other and the system under test). Additionally, it requires a testing regime with artefacts (e.g., test models and test cases) that are meaningful and understandable to the diverse population.

In the remainder of the paper, we propose an architecture for testing ethical decision-making in Section 2. Then in Section 3, we focus on the formalisation of ethical theories that can be used for test-case generation, test oracles, and learning from stakeholder engagement. In Section 4, we review some of the related work and in Section 5, we conclude the paper and present the directions of our ongoing research.

2 Architecture for Testing Ethics

Figure 1 demonstrates our proposed architecture to test ethics in autonomous systems. The overall workflow starts from the selected *ethical model* which along with the System Under Test (SUT) is used to generate the test cases. The generated test cases are executed on the SUT, and the resulting system traces are passed to the *conformance analyser* which checks whether the test cases pass or fail the ethical tests according to the selected oracle. A subset of both passing and failing test cases is then passed to the stakeholders for their consideration and validation. It is important at this step to perform a meaningful *test selection* so that these test executions are representative of the system's overall behaviour from the ethical perspective. The stakeholders/ethical experts can indicate the test cases for which they do not agree with the outcome provided by the conformance analyser. Such test cases become counter-examples to the adopted ethical model. The *learning module* component can use generated counter-examples to adjust the ethical model. This adjustment can be performed by applying the ideas behind existing works on online and offline learning [14, 19, 31, 37, 49], as well as search-based approaches to oracle improvement [29,30,46]. The presented iterative process can continue until no more counter-example test cases can be identified. This does not mean that the SUT eventually passes all tests but that the stakeholder(s) eventually agree with the pass/fail decision of tests generated by the system. Each stakeholder may be using the testing system separately from other stakeholders and thus learning a different ethical model that reflects their concerns. The test cases they generate can then be used for discussion and negotiation with other stakeholders.

2.1 Test Input Generation

The test cases need to capture the scenarios where the decision taken by the system puts various ethical principles of the ethical model into conflict. The aim of test case generation in this architecture is to generate such test cases out of

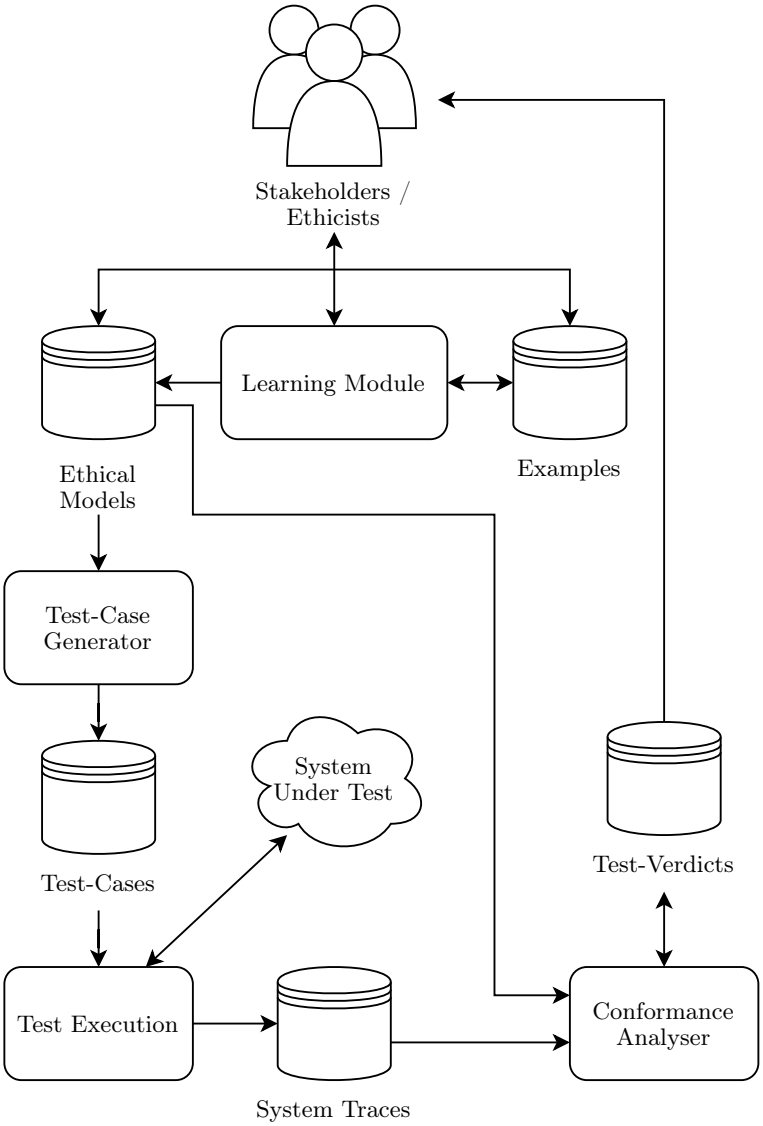

Fig. 1. Proposed Architecture for Testing Ethics

the large space of all possible scenarios. One way to achieve this task is to cast it as a *search-based* testing problem [28, 34]. The search process can be guided by two different goals: (1) detecting ethical faults more effectively (2) checking the soundness of the ethical model.

The first goal concerns generating scenarios in which the violation of an ethical model can be effectively established; in other words, we are looking for scenarios in which the choices of the agent can clearly and quickly reveal the violation of an ethical model. When generating such scenarios, we would like to maximise the diversity between them as well as the coverage of the ethical model and the space of possible scenarios. The target measure for diversity can be uniformity which can be measured by using a statistical test related to the uniform distribution, such as the L2-test [26]. As a target measure for coverage a straightforward measure is the t-wise coverage of possible choices among actions.

The second goal aims to put the agent into ethical dilemmas, e.g., to force it to make a choice between different actions that have similar ethical status. For this, we can first identify test objectives that determine how far candidate tests are from taking different ethically-charged actions. We then can guide the test generation process towards test scenarios that lead to undesired interactions between these test objectives [3] by using *many-objective optimisation* algorithms such as NSGA-II [20], HypE [12], and MOSA [36]. Game-theoretic formulations [25] and synthesising scenarios towards an expected equilibrium can provide another alternative approach.

For our motivating example, we need to first guide the scenario-generation towards situations 1) where a particular rule of the road is applicable (e.g., as specified in the Highway Code), and 2) where there is a potential conflict in the rights of way and/or the rules of the road. Regarding case 1, a generated test scenario should for example demonstrate whether a vehicle gives way to an ambulance. If it does not, this can reveal faults in ethical decision-making. A scenario pertaining to case 2 is when giving way to an ambulance involves damaging another vehicle or hitting a curb or a road user. Moreover, we would like to diversify the set of generated scenarios by considering substantially different situations and also cover various possible conflicts, e.g., all possible pairwise choices among conflicting actions.

2.2 Test Oracle Identification

Test oracle [13] identification is one of the key problems in testing ethics. Autonomous systems are typically stochastic and the ground truth is not specified a priori. Moreover, for ethics itself, even human ethics are often faced with difficult dilemmas without easy answers, in which each side might have valid arguments. In our architecture, the *conformance analyser* needs to automatically decide which decisions are acceptable and which are not. The ethical models needed for conformance analysis can be extracted from various sources, discussed below.

Ethical Models from Laws and Policies Researchers and companies have recommended various laws and regulations from government or non-profit insti-

tutions as a means of ensuring machine ethics. There are some widely acknowledged regulations such as the "Ethics Guidelines for Trustworthy AI" from the European Commission [44] and "Recommendation on the Ethics of AI" from the UNESCO Ad Hoc Expert Group [48]. IEEE developed an extensive process model standard for incorporating ethical concerns during system design [1]. Some of these standards have been translated into domain-specific guidelines, e.g., in the domain of autonomous vehicles [24]. The information from these sources can be extracted, in terms of rules or utility functions, to build the basis for an ethical model to be used in our automated conformance analyser. In Section 3, we provide the templates that can be used to encode these informal descriptions into a formal specification.

Test Oracles from Stakeholders and Human Experts The information about the expected behaviour of the ethical model can be provided in the form of a number of examples by the stakeholders or human experts. These examples can then be generalised into ethical models using learning algorithms (such as automata learning or neural networks) in our templates for ethical theories specified below. Our ethical theories specify a relative value for different types of behaviour; when coming up with a complete model is challenging (due to lack of sufficient examples or conflict among stakeholders), abstractions of such models such as metamorphic properties can be used.

For example, for our motivating example, a specific ethical theory can 1) specify the relative value of different actions, such as giving way to an ambulance, damaging another vehicle, and hitting a road user, or 2) specify the total utility of each course of actions for the other road users (including any quantification of the resulting damage) by taking a weighted sum of the utility of actions for the individual road users involved in a scenario.

The next section describes the ethical theories that can be used as templates for ethical models.

3 Ethical Theories for Conformance Analysis

Passing a verdict on the decision-making scenarios (bottom-right corner of Figure 1) requires defining or learning an ethical model (top-left corner of Figure 1), e.g., through a set of rules. Modelling the ethics of autonomous agents has been the subject of machine ethics for the past few decades. It aims to understand the consequences of machine behaviour on either other machines or people [7]. The main goal of this field is to study and help construct systems that act under a specific ethical theory. Instead of committing to a particular ethical theory, in this paper, we develop a general semantic framework that can be used to define different ethical models. These models can be learned using the examples provided by the stakeholders by fitting the parameters of our semantic model.

In this section, we review the three major ethical theories, namely, deontological, consequentialist (or utilitarian), and virtue ethics [47], and present the semantic templates for them. Our templates use the simple perception-action

agent model by Russel and Norvig [40], depicted in Figure 2. This model postulates two sets of *Prc* and *Act*, for percepts and actions, respectively. Our framework is designed for black-box testing and can be further extended to take the details of the agent implementation into account. Furthermore, depending on the ethical theory we may need to model the environment or not.

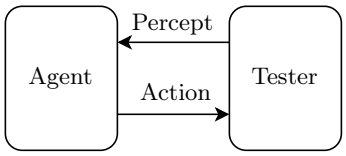

Fig. 2. A Model of Agent and Tester Interaction

3.1 Deontological Ethics

In deontological ethics, an action is morally good if it follows a predefined set of moral values or rules [45]. There are two perspectives for defining the values of actions: agent-centred and patient-centred. The former focuses the ethics on the agent's actions, while the latter focuses the ethics on the agent who receives the action and the consequences [32]. Following the moral theory of deontological ethics, our semantic template for a deontological oracle involves defining a value for each action; the values are taken from any (pseudo-) metric space on a discrete or continuous set, i.e., a set of values with a defined distance between them. A basic domain is the discrete domain {Forbidden, Neutral, Obligatory}, with a unit distance from Forbidden to Neutral and from Neutral to Obligatory. However, more sophisticated domains define a spectrum of actions and their relative values to each other. The objective will be used to define the oracle as well as to steer the test-case generation in a direction that is more likely to reveal ethical issues. The test case input is a non-empty sequence of percepts. The considered output is the last action in the sequence of actions produced as the result of inputs. The oracle checks that the output action is not more than ϵ apart from the ideal action after the input (first row in Table 1).

Consider our running example; assume that the agent has three moral duties:

1. an agent shall respect human lives,
2. an agent shall give way to ambulances, and
3. an agent shall not damage other cars.

Obviously these three moral duties can be in conflict and we model the conflict, and the relative priority of these rules as follows. We define a model in which the value of not giving way to the ambulance is -1, damaging a car is -0.5 and hitting a human road user is -2. Note that this choice of values are a rough indication of the relative importance of the three rules; however, often

the domain of values needs to be more complex, e.g., be multi-dimensional. This allows for comparing the outcomes of different scenarios based on deontological ethics. For example, based on this oracle, a scenario in which a vehicle blocks the ambulance because it may damage another vehicle while giving way to the ambulance is considered a failure, while blocking the ambulance when giving way will lead to killing or seriously injuring a pedestrian will be considered a pass.

Table 1. Model Templates for Different Theories of Ethics

Framework	Objective	Test input	Test output	Test oracle
Deontological	$Obj : Act \to Val$	$\alpha : Prc^+$	$a : Act$, last observed action	Obj(ideal after α) - $Obj(a$ after $\alpha) \leq \epsilon$
Consequentialism	$Obj : Act \times Env \to Val$	$\alpha : Prc^+$	$\beta : Act^+$, sequence of observed actions	Obj(ideal after α, env) - $Obj(\beta$ after α, env$) \leq \epsilon$
Virtue	$S \subseteq Prc^+ \times Act^+$, Robotic saint	$A \subseteq Prc^+$	$B \subseteq Prc^+ \times Act^+$	(1) $\forall \alpha \in B \ \exists \beta \in S \cdot dist(\alpha, \beta) \leq \epsilon \wedge \forall \beta \in B \ \exists \alpha \in S \cdot dist(\alpha, \beta) \leq \epsilon$, (2) Convergence to S

3.2 Consequentialism

Consequentialism focuses on consequences, e.g., aiming at maximising global well-being [42]. According to this moral theory, the value of an action is determined by its global utility. Two of the sub-fields of consequentialism are act utilitarianism and rule utilitarianism [42]. The former establishes that every single act must focus on maximising utility. The latter focuses on social rules. It establishes that the only rules that need to be applied are those that maximise well-being.

The semantic framework for a consequentialist oracle (second row in Table 1) involves an objective function that defines the value of actions through their effects on the environment Env. The form of such an objective is a weighted sum of the effect of actions (e.g., happiness) of agents in the environment. The input and output, similar to the deontological case, are non-empty traces of percepts and actions. The oracle asserts that the distance of the accumulative value of all actions is not farther than ϵ from that of the ideal sequence.

In our running example, an ethical model assigns a value to the relative damage to the different patients caused by an action (i.e., ego vehicle, other vehicles, road users, and the emergency vehicle patient) and a weighting to calculate the total utility of the scenario.

3.3 Virtue Ethics

In virtue ethics, an action is good if the agents manifest virtuosity when they act [43]. A distinct form of virtue ethics is the exemplarist virtue theory [52] where the morality of an agent is measured in terms of its similarity to an exemplary agent (a moral saint).

Virtue ethics is a challenging ethical theory to test; we model the semantic model of virtue ethics (third row of Table 1) as a robotic saint which is modelled as a set of pairs of percept and action sequences. The test input is a set of percept sequences and the test output is a set of pairs of percept- and action sequences. We propose two types of oracles for testing virtue ethics; the first type of oracle is a conformance oracle that checks for each behaviour of the saint, there is a similar behaviour of the agent that is at most ϵ apart and vice versa. The notion of distance is calculated based on a notion of action similarity akin to the objective function of deontological ethics. The second type of oracle measures the convergence of the agent's behaviour towards the behaviour of the saint, i.e., measures how much more conforming the longer traces of behaviour become compared to the shorter traces. In our running example, an ethical model is defined by learning the behaviour of an idealised driver, e.g., exploiting driving logs, taking accidents as negative examples; subsequently, measures of conformance [2] and convergence [4] can be used to measure the conformance of an agent to the model.

4 Related work

4.1 Ethical Oracle Identification

Test oracles that rely on human judgment One way to identify test oracles is to rely on the judgement of human (stakeholders and ethicists). Pontier and Widdershoven [39] use human judges to provide oracles to find unethical issues. Allen et al. [6] conducted a "Moral Turing Test" in which a "blind" observer is asked to compare the behaviour of a machine to humans. Similarly, Anderson et al. [8] provided a self-made Ethical Turing Test to evaluate their ethical principles: If the system performs as an ethical expert would, then it passes the test. They also argued that ethically significant behaviour of autonomous systems should be guided by ethical principles determined by ethics experts. Wu et al. [51] employed ordinary human data to derive human policies and help to learn ethical behaviours as test oracles.

Test oracles that rely on laws and policies Asaro [11] suggests the existing legal system can be used as a starting point for deriving AI ethics. Vanderelst and Winfield [50] test robot behaviours based on Asimov's laws of robotics. The use of Asimov's laws has been extensively criticised, e.g., in [9].

Test Oracles with Simulations The survey by Nallur [35] does not specifically mention testing ethics. However, it discusses the evaluation of ethics by simulation of ethical dilemmas. In this formulation, ethical dilemmas serve as test cases, and if the autonomous system can resolve a dilemma in a particular manner, then ethics were successfully implemented in this system.

4.2 Ethical Representation

Arkoudas et al. [10, 15] propose to use Horty logic to compose ethical semantics. Dennis et al. [21] developed a framework for representing the context of ethical reasoning, which involves encoding user values as a set of rules. An ethical reasoner can then be embedded in a reasoning cycle to gather contextual information and update its ethical encoding. Dennis et al. [22] apply the AJPF model-checker to verify the behaviour of the consequence engine in a robot system.

There are also several works that mentioned the importance of ethics testing. Pontier and Hoorn talked about the importance of making ethics measurable [38]: "Ethics must be made computable in order to make it clear exactly how agents ought to behave in ethical dilemmas". Madl and Franklin [33] discuss the necessity of a set of moral tests to guarantee AI ethics and propose the idea of moral test-driven development.

AI ethics considerations can be various, including privacy, fairness, accountability, explainability and others. The testing efforts on these aspects mainly focus on fairness testing. We refer to Chen et al. [18] for a comprehensive survey of the literature on testing AI fairness. As far as we know, there is no general framework that is specially designed for testing ethical decision-making.

5 Conclusion and Future Research Roadmap

In this paper, we propose a framework for testing ethical aspects of decision-making in autonomous systems. Our framework comprises three major parts: 1) a test-case generation algorithm, 2) a conformance analyser, and 3) a learning algorithm to learn an ethical model for the former two parts and adjust it based on stakeholders' feedback. We presented three formalisations of the major ethical theories that can be used as templates for the ethical models in test-case generation, conformance analysis, and learning from examples. The purpose of our proposed framework is to provide a tool for various stakeholders, and particularly under-represented and less powerful ones, to specify their concerns and test complex autonomous systems against them.

We plan to instantiate our framework with concrete algorithms in the domain of autonomous vehicles. We are currently developing a simulation environment in order to execute the generated scenarios, present the stakeholders with tangible examples, and receive their feedback. Our framework focuses on a black-box perspective on the system under test; extensions of our framework can be developed by using more information from the agents' state, including aspects such as belief, desire, and intention as well as specifications of neuro-symbolic agents.

Acknowledgments

The support of the UKRI Trustworthy Autonomous Systems Hub (reference EP/V00784X/1) and Trustworthy Autonomous Systems Node in Verifiability (reference EP/V026801/2) is gratefully acknowledged. The authors are grateful to Peta Masters and to the anonymous reviewers for their constructive comments.

References

1. IEEE standard model process for addressing ethical concerns during system design. IEEE Std 7000-2021 pp. 1–82 (2021). https://doi.org/10.1109/IEEESTD.2021.9536679
2. Abbas, H.: Test-Based Falsification and Conformance Testing for Cyber-Physical Systems. Ph.D. thesis, Arizona State University, Tempe, USA (2015), https://hdl.handle.net/2286/R.I.29861
3. Abdessalem, R.B., Panichella, A., Nejati, S., Briand, L.C., Stifter, T.: Testing autonomous cars for feature interaction failures using many-objective search. In: 2018 33rd IEEE/ACM International Conference on Automated Software Engineering (ASE). pp. 143–154. IEEE (2018)
4. Abdessalem, R.B., Panichella, A., Nejati, S., Briand, L.C., Stifter, T.: Testing autonomous cars for feature interaction failures using many-objective search. In: Huchard, M., Kästner, C., Fraser, G. (eds.) Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering, ASE 2018, Montpellier, France, September 3-7, 2018. pp. 143–154. ACM (2018). https://doi.org/10.1145/3238147.3238192
5. Aggarwal, A., Lohia, P., Nagar, S., Dey, K., Saha, D.: Black box fairness testing of machine learning models. In: Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. pp. 625–635 (2019)
6. Allen, C., Varner, G., Zinser, J.: Prolegomena to any future artificial moral agent. Journal of Experimental & Theoretical Artificial Intelligence **12**(3), 251–261 (2000)
7. Anderson, M., Anderson, S.L.: Machine ethics. Cambridge University Press (2011)
8. Anderson, M., Anderson, S.L.: Geneth: A general ethical dilemma analyzer. Paladyn, Journal of Behavioral Robotics **9**(1), 337–357 (2018)
9. Anderson, S.L.: The Unacceptability of Asimov's Three Laws of Robotics as a Basis for Machine Ethics, p. 285–296. Cambridge University Press (2011). https://doi.org/10.1017/CBO9780511978036.021
10. Arkoudas, K., Bringsjord, S., Bello, P.: Toward ethical robots via mechanized deontic logic. In: AAAI fall symposium on machine ethics. pp. 17–23. The AAAI Press Menlo Park, CA, USA (2005)
11. Asaro, P.M.: What should we want from a robot ethic? In: Machine Ethics and Robot Ethics, pp. 87–94. Routledge (2020)
12. Bader, J., Zitzler, E.: Hype: An algorithm for fast hypervolume-based many-objective optimization. Evolutionary computation **19**(1), 45–76 (2011)
13. Barr, E.T., Harman, M., McMinn, P., Shahbaz, M., Yoo, S.: The oracle problem in software testing: A survey. IEEE transactions on software engineering **41**(5), 507–525 (2015)

14. Bottou, L.: Online algorithms and stochastic approxima-p tions. Online learning and neural networks (1998)
15. Bringsjord, S., Arkoudas, K., Bello, P.: Toward a general logicist methodology for engineering ethically correct robots. IEEE Intelligent Systems **21**(4), 38–44 (2006)
16. Chakraborty, J., Majumder, S., Menzies, T.: Bias in machine learning software: why? how? what to do? In: Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. pp. 429–440 (2021)
17. Chen, Z., Zhang, J., Sarro, F., Harman, M.: Maat: A novel ensemble approach to addressing fairness and performance bugs for machine learning software. In: The ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE) (2022)
18. Chen, Z., Zhang, J.M., Hort, M., Sarro, F., Harman, M.: Fairness testing: A comprehensive survey and analysis of trends. arXiv preprint arXiv:2207.10223 (2022)
19. Damasceno, C.D.N., Mousavi, M.R., da Silva Simão, A.: Learning to reuse: Adaptive model learning for evolving systems. In: Ahrendt, W., Tarifa, S.L.T. (eds.) Integrated Formal Methods - 15th International Conference, IFM 2019, Bergen, Norway, December 2-6, 2019, Proceedings. Lecture Notes in Computer Science, vol. 11918, pp. 138–156. Springer (2019). https://doi.org/10.1007/978-3-030-34968-4_8
20. Deb, K., Jain, H.: An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, part i: solving problems with box constraints. IEEE transactions on evolutionary computation **18**(4), 577–601 (2013)
21. Dennis, L.A., Bentzen, M.M., Lindner, F., Fisher, M.: Verifiable machine ethics in changing contexts. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 11470–11478 (2021)
22. Dennis, L.A., Fisher, M., Winfield, A.: Towards verifiably ethical robot behaviour. In: Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence (2015)
23. Dubber, M.D., Pasquale, F., Das, S.: The Oxford Handbook of Ethics of AI. Oxford University Press (2020). https://doi.org/10.1093/oxfordhb/9780190067397.001.0001
24. Evans, K., de Moura, N., Chauvier, S., Chatila, R., Dogan, E.: Ethical decision making in autonomous vehicles: The AV ethics project. Sci. Eng. Ethics **26**(6), 3285–3312 (2020). https://doi.org/10.1007/s11948-020-00272-8
25. Gogoll, J., Zuber, N., Kacianka, S., Greger, T., Pretschner, A., Nida-Rümelin, J.: Ethics in the software development process: from codes of conduct to ethical deliberation. Philosophy & Technology **34**(4), 1085–1108 (2021). https://doi.org/10.1007/s13347-021-00451-w
26. Goldreich, O., Ron, D.: On testing expansion in bounded-degree graphs. In: Studies in Complexity and Cryptography. Miscellanea on the Interplay between Randomness and Computation, pp. 68–75. Springer (2011)
27. Govindarajulu, N.S., Bringsjord, S., Ghosh, R., Sarathy, V.: Toward the engineering of virtuous machines. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. p. 29–35. AIES '19, Association for Computing Machinery, New York, NY, USA (2019)
28. Harman, M., Jia, Y., Zhang, Y.: Achievements, open problems and challenges for search based software testing. In: 2015 IEEE 8th International Conference on Software Testing, Verification and Validation (ICST). pp. 1–12. IEEE (2015)

29. Jahangirova, G., Clark, D., Harman, M., Tonella, P.: Test oracle assessment and improvement. In: Proceedings of the 25th International Symposium on Software Testing and Analysis. pp. 247–258 (2016)
30. Jahangirova, G., Clark, D., Harman, M., Tonella, P.: An empirical validation of oracle improvement. IEEE Transactions on Software Engineering **47**(8), 1708–1728 (2019)
31. Jain, L.C., Seera, M., Lim, C.P., Balasubramaniam, P.: A review of online learning in supervised neural networks. Neural computing and applications **25**(3), 491–509 (2014)
32. Kant, I.: Groundwork for the Metaphysics of Morals. Yale University Press. Commented by Jerome B Schneewind (1785)
33. Madl, T., Franklin, S.: Constrained incrementalist moral decision making for a biologically inspired cognitive architecture. In: A Construction Manual for Robots' Ethical Systems, pp. 137–153. Springer (2015)
34. McMinn, P.: Search-based software test data generation: a survey. Software testing, Verification and reliability **14**(2), 105–156 (2004)
35. Nallur, V.: Landscape of machine implemented ethics. Science and engineering ethics **26**(5), 2381–2399 (2020)
36. Panichella, A., Kifetew, F.M., Tonella, P.: Reformulating branch coverage as a many-objective optimization problem. In: 2015 IEEE 8th international conference on software testing, verification and validation (ICST). pp. 1–10. IEEE (2015)
37. Parisi, G.I., Kemker, R., Part, J.L., Kanan, C., Wermter, S.: Continual lifelong learning with neural networks: A review. Neural Networks **113**, 54–71 (2019)
38. Pontier, M., Hoorn, J.: Toward machines that behave ethically better than humans do. In: Proceedings of the annual meeting of the cognitive science society. vol. 34 (2012)
39. Pontier, M.A., Widdershoven, G.A.: Robots that stimulate autonomy. In: IFIP International Conference on Artificial Intelligence Applications and Innovations. pp. 195–204. Springer (2013)
40. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach. Pearson, 4th edition edn. (2020)
41. Shea-Blymyer, C., Abbas, H.: Algorithmic ethics: Formalization and verification of autonomous vehicle obligations. ACM Trans. Cyber-Phys. Syst. **5**(4) (sep 2021). https://doi.org/10.1145/3460975
42. Sinnott-Armstrong, W.: Consequentialism. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, Fall 2021 edn. (2021)
43. Slote, M.: Agent-based virtue ethics. Handbuch Tugend und Tugendethik pp. 1–10 (2020)
44. Smuha, N.: Ethics guidelines for trustworthy ai. In: AI & Ethics, Date: 2019/05/28-2019/05/28, Location: Brussels (Digityser), Belgium (2019)
45. Tännsjö, T.: Understanding ethics. Edinburgh University Press (2013)
46. Terragni, V., Jahangirova, G., Tonella, P., Pezzè, M.: Evolutionary improvement of assertion oracles. In: Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. pp. 1178–1189 (2020)
47. Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M., Bernstein, A.: Implementations in machine ethics: A survey. ACM Computing Surveys (CSUR) **53**(6), 1–38 (2020)
48. UNESCO: Recommendation on the Ethics of Artificial Intelligence. United Nations Educational, Scientific and Cultural Organization (2022)

49. Vaandrager, F.W.: Model learning. Commun. ACM **60**(2), 86–95 (2017). https://doi.org/10.1145/2967606
50. Vanderelst, D., Winfield, A.: An architecture for ethical robots inspired by the simulation theory of cognition. Cognitive Systems Research **48**, 56–66 (2018)
51. Wu, Y.H., Lin, S.D.: A low-cost ethics shaping approach for designing reinforcement learning agents. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
52. Zagzebski, L.: Exemplarist virtue theory. Metaphilosophy (2010)