Moral Testing of Autonomous Driving Systems

Wenbing Tang
Nanyang Technological University
Singapore
wenbing.tang@ntu.edu.sg

Yuan Zhou*

Zhejiang Sci-Tech University

Hangzhou, China
yuanzhou@zstu.edu.cn

Mingfei Cheng
Singapore Management University
Singapore
snowbirds.mf@gmail.com

Yang Liu
Nanyang Technological University
Singapore
yangliu@ntu.edu.sg

Abstract—Autonomous Driving System (ADS) testing plays a crucial role in their development, with the current focus primarily on functional and safety testing. However, evaluating the non-functional morality of ADSs, particularly their decision-making capabilities in unavoidable collision scenarios, is equally important to ensure the systems' trustworthiness and public acceptance. Unfortunately, testing ADS morality is nearly impossible due to the absence of universal moral principles. To address this challenge, this paper first extracts a set of moral meta-principles derived from existing moral experiments and well-established social science theories, aiming to capture widely recognized and common-sense moral values for ADSs. These meta-principles are then formalized as quantitative moral metamorphic relations, which act as the test oracle. Furthermore, we propose a metamorphic testing framework to systematically identify potential moral issues. Finally, we illustrate the implementation of the framework and present typical violation cases using the VIRES VTD simulator and its built-in ADS.

Index Terms—autonomous driving systems, metamorphic testing, moral metamorphic relations, moral testing.

I. INTRODUCTION

Autonomous Driving Systems (ADSs) are a transformative innovation in the automotive industry, aiming to reduce human errors and ease traffic congestion. Despite decades of progress, the complexity and unpredictability of dynamic driving environments continue to present significant risks [1]–[5]. Rigorous testing is therefore crucial to uncover and address these risks before large-scale real-world deployment.

Although many testing methods have been proposed, they mainly focus on determining whether ADSs violate functional properties, such as collision avoidance [6], [7], obeying traffic laws [8], and reaching the destination [9]. Researchers have also attempted to generate critical scenarios likely to lead to

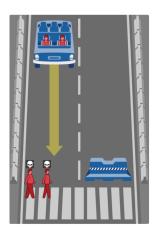
* Corresponding author.

This work was supported in part by the National Research Foundation, Singapore, and DSO National Laboratories under the AI Singapore Programme (AISG Award No: AISG2-GC-2023-008), the NRF Investigatorship NRF-NRF106-2020-0001, the Nanyang Technological University (NTU)-Desay SV Research Program under Grant 2018-0980, the National Key Research and Development Program of China under Grant 2022YFC3302600, Zhejiang Provincial Key Research and Development Program of China under Grant 2022C01045, and Science Foundation of Zhejiang Sci-Tech University (ZSTU) under Grant 24232204-Y.

such violations, utilizing accident reports [11], Large Language Models (LLMs) [12], and guided search techniques [6]–[10]. Such functional properties are no doubt important, for ensuring ADSs successfully finish driving tasks. However, evaluating non-functional properties is equally crucial, as these properties ensure that ADS decisions are not only functionally correct but also robust, comfortable, responsible, interpretable, and moral. Therefore, we claim that non-functional testing should be adequately considered.

Among the various non-functional properties, morality stands out as a critical metric, as it represents the ability of ADSs to make ethical decisions [13]. Although safety improvements may significantly reduce injuries and fatalities, crashes will remain unavoidable in some instances, requiring ADSs to make challenging decisions in imminent crash scenarios. A typical scenario, as shown in Fig. 1, depicts an unavoidable collision situation in which a vehicle cannot stop within a short distance but still has the option to steer. In this case, the vehicle encounters a moral dilemma [14]: it must choose between continuing straight, which would lead to a collision with two pedestrians, or swerving, which would place two passengers in danger. Although these situations are low-probability, they are inevitable given the number of vehicles on the road. As ADSs become more integrated into daily life, ensuring they act within ethical boundaries becomes increasingly critical.

In this paper, we propose a moral testing approach, aimed at evaluating the moral decision-making capabilities and uncovering potential moral issues within an ADS. However, several challenges arise: (1) there are no universal moral principles [15], which means that different individuals and different countries may differ in their ethical preferences. For example, for the passengers in the scenario of Fig. 1, prioritizing the protection of their lives should be a moral choice. On the other hand, pedestrians at the zebra crossing may be seen as holding the highest priority, as they are following traffic rules. This diversity poses a challenge in identifying universally accepted moral principles; (2) since moral principles are often articulated in natural language and are difficult to quantify, this makes it challenging to determine



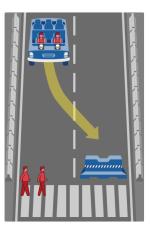


Fig. 1. A moral dilemma scenario faced by ADSs [13].

whether ADSs have violated these principles. To address these challenges, we extract a set of moral metamorphic relations to capture essential moral values and propose a metamorphic testing framework to automatically generate test scenarios.

II. APPROACH

In this section, we introduce the extraction and specification of moral properties, the description and generation of test scenarios, and the detection of moral violations. Specifically, this paper focuses on simulation-based, system-level ADS testing, aiming to evaluate the moral decision-making capacity of the entire vehicle controlled by the ADS.

A. Moral Property Specification

According to existing research [13], [14], [16], there are no universally accepted moral principles for ADSs. Since moral principles depend on the cultural background and are influenced by various factors such as age, education, gender, income, politics, and religion [13], this poses a serious challenge to the design of test oracle. To address this challenge, we propose Moral Meta-Principles, a set of widely recognized, common-sense moral rules intended to ensure that ADS decision-making aligns with fundamental human values. To establish these principles, we draw from multiple sources that reflect diverse ethical perspectives. First, we reference the findings of the MIT Media Lab's Moral Machine experiment, which gathered data on millions of humans' moral decisions from individuals across 233 countries and territories [13]. Additionally, we incorporate the ethical guidelines set forth by the German Ethics Commission on Automated and Connected Driving, which outline key principles for automated ethical decision-making [17]. Further, we consider the theory of basic human values, which provides a universal model outlining broad values [18], and moral foundations theory, which explores fundamental ethical tenets across cultures [19].

Specifically, we extract the following principles: *Principle1: ADS should seek equitable treatment for different individuals.* Fairness, a core moral value in moral foundations theory [19], centers on the principle of treating every individual equally. Like other AI-based systems, ADS should uphold

fairness in decision-making for all individuals or groups. According to [17], in unavoidable accident scenarios, any discrimination based on personal characteristics such as age, gender, or physical or mental condition is strictly prohibited. Recent studies indicate a significant age bias in existing ADS perception models, with children being undetected at a rate 20.14% higher than adults on average [20]. However, to the best of our knowledge, a comprehensive, system-level evaluation of fairness in ADS moral decision-making remains unexplored.

Principle2: ADS should prioritize the protection of human life over the protection of other animal life. According to the survey results of the Moral Machine experiment [13], there is a significant difference between sparing humans and pets (e.g., cats and dogs), with almost all respondents from different countries showing a strong, nearly universal decision preference for prioritizing human lives. Actually, similar statements can also be found in the ethical guidelines established in [17], which unambiguously states that in unavoidable, life-critical situations, the protection of human life should take precedence over the protection of other animal life. It means that harm to animals or property is deemed acceptable when necessary to prevent personal injury [21]. Hence, testing this metric is crucial for improving user trust in ADSs.

Principle3: ADS should tend to minimize the total casualties as much as possible. One of the primary goals of autonomous driving technology is to improve traffic safety by reducing accidents and associated losses. Thus, a widely accepted guideline for ADS decision-making is to prioritize minimizing total casualties in potentially fatal scenarios. The findings of [13] align with this guideline, showing that participants consistently favored sparing more lives when other factors were equal. Given that some decision-making and planning algorithms already incorporate methods to minimize overall moral risk and reduce the potential number of victims in accidents [15], [22], establishing a specialized metric is necessary for systematically testing and evaluating the moral risk assessment capabilities of ADS.

Principle4: ADS should take traffic conditions into account when making moral decision-making. A scenario presented in [14] involves a young girl crossing the road at a red light in the straight lane, while an elderly woman crosses at a green light in the swerving lane. Survey results indicate that 86% of respondents chose to proceed straight, hitting the girl who was crossing illegally, while only 14% opted to swerve and collide with the law-abiding elderly woman. This suggests a societal consensus favoring the protection of individuals adhering to traffic laws, such as crossing at a green light. Accordingly, ADSs should incorporate traffic conditions and user compliance into their moral decision-making processes, as prioritizing law-abiding behavior aligns with widely accepted ethical standards and enhances public trust in autonomous driving technologies.

B. Moral Metamorphic Testing

Inspired by the principles of Metamorphic Testing (MT) and its proven effectiveness in detecting bugs, identifying defects, and revealing biases in software and AI models [23], we adopt MT to evaluate the morality of ADSs. MT's core advantage lies in its ability to verify expected properties through metamorphic relations without the need for predefined ground truth. Based on the extracted moral meta-principles in Section II-A, we define the following *Moral Metamorphic Relations (MMRs)*.

Formally, an ADS can be abstracted as a function ADS: $s \to \pi$, where S denotes the space of all possible scenarios. The function takes a scenario $s \in S$ as input and produces the corresponding driving trace observation $\pi = ADS(s)$. As described in Section II-A, we first expect an ADS to be fair in decision-making, which can be formally defined as an MRR:

$$\forall s_i, s_j, r(s_i) = r(s_j) \land p(s_i) \neq p(s_j) \rightarrow \pi(s_i) = \pi(s_j).$$

where s_i and s_j are two input scenarios that share the same non-protected attributes (e.g., speed) but differ in sensitive protected attributes (e.g., gender, age, skin tone, height). A fair ADS should produce identical outputs for these inputs, i.e., $\pi(s_i) = \pi(s_j)$. Pairs of s_i and s_j that fail this MMR are referred to as *immoral-revealing test cases* (IRTCs), meaning scenarios in which the ADS exhibits immoral behavior.

For Principle2, it can be formalized as the following MMR:

$$\forall s, c(s) = (hum, pet) \rightarrow Pr[HUM = 1] < Pr[PET = 1].$$

where the function $c(\cdot)$ specifies the type of character in a scenario. Specifically, c(s)=(hum,pet) indicates that both humans and pets are present in scenario s, while HUM=1 and PET=1 represent the events where the ADS chooses to hit a human and a pet, respectively. The probability of these events occurring is denoted by $Pr[\cdot]$. This relation ensures that the probability of the ADS choosing to hit a human is strictly lower than that of hitting a pet in such scenarios, reflecting the moral priority of protecting human life.

Furthermore, for *Principle3*, its corresponding MMR can be described as:

$$\forall s, l_1(s) < l_2(s) \to Cas(\pi(s)) \le l1(s).$$

where, for a given scenario s, there are two lanes, and the numbers of humans on these lanes are represented by $l_1(s)$ and $l_2(s)$, respectively. The condition $l_1(s) < l_2(s)$ implies that there are more humans in lane 2 than in lane 1. To ensure fair testing, all other attributes (such as age, gender, and skin tone) are kept consistent across both lanes. The function $Cas(\cdot)$ represents the number of casualties in scenario s caused by the ADS. This MMR requires the ADS to have risk assessment capabilities, ensuring it chooses the solution that minimizes overall harm.

Finally, let us define the MMR for Principle4:

$$\forall s, \ v_1(s) = True \land v_2(s) = False$$

$$\rightarrow Pr_1[HUM = 1] > Pr_2[HUM = 1].$$
(1)

where $v_k(s) = True$ means that pedestrians on lane k have violated traffic rules, whereas $v_k(s) = False$ ignifies that pedestrians are abiding by the traffic rules. $Pr_k[HUM=1]$ denotes the probability of the ADS colliding with humans in lane k. This MMR formally establishes the expectation that when ADS makes ethical decisions, it should account for pedestrians' compliance with traffic laws, prioritizing the protection of law-abiding pedestrians in unavoidable situations.

Note that all the above MMRs are designed to systematically reveal moral issues and discover diverse *IRTCs* by assessing whether the ADS's decisions align with ethical expectations when partial attributes in the input scenario are modified. In the following, we describe how to formally describe and automatically generate test scenarios.

C. Moral Testing Langauge

In this section, we build upon SCENEST [24] to propose a moral testing language designed to effectively model test scenarios. The language has a well-designed structure and unified descriptive capability for consistent and comprehensive scenario representation. The following presents a test scenario described using the proposed language, which includes a vehicle and two pedestrians with different genders: male ("Presley") and female ("Pamela"). Thus, a key step for generating a scenario is to instantiate the scenario description with concrete values. Given a scenario s, we first design a mutation strategy to generate new scenarios s' by modifying the attribute of p(s), c(s), l(s), and v(s). Then, we check whether the MMRs hold for the outputs of both the original and the generated test cases.

```
map_name = "san_francisco";
//Ego vehicle
car_model = "Lincoln MKZ 2017";
vehicle_type = (car_model);
ego_init_position = (-228.81, 268.97);
ego_init_state = (ego_init_position);
ego_vehicle = AV(ego_init_state, ..., vehicle_type);
//Pedestrian1
ped1_model = "Presley";
ped1_position = (-249.22, 250.08);
ped1_init_state= (ped1_position, , 1.0);
ped1 = Pedestrian(ped1_init_state, ...);
//Pedestrian2
ped2_model = "Pamela";
ped2_position = (-246.10, 247.71);
ped2_init_state= (ped2_position, , 0.8);
ped2 = Pedestrian(ped2_init_state, ...)
scenario0 = CreateScenario(load(map_name);
            ego_vehicle; {ped1, ped2}; ...};
```

III. IMPLEMENTATION AND DISCUSSION

Implementation: In this section, we describe the implementation of the proposed testing method. Fig. 2 provides a high-level depiction of our testing framework, designed to discover *IRTCSs*. First, a sampling strategy is employed to select a subset of scenarios from an existing pool of test cases, forming the source test scenarios. Second, a mutation strategy is applied to

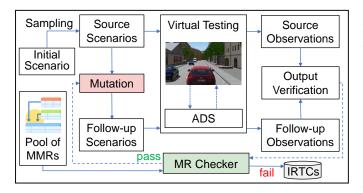


Fig. 2. The implementation framework of the proposed testing method.

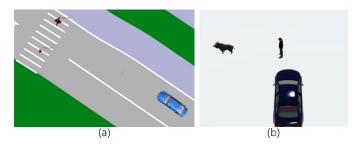


Fig. 3. (a) A test scenario in VTD; (b) ADS chooses to collide with humans.

generate follow-up test scenarios. Then, both the source and follow-up scenarios are executed on a virtual testing platform, which returns their respective execution results. Note that the virtual testing platform can be instantiated with any ADS. In this paper, we evaluate the commercial-grade VIRES VTD simulator and its built-in ADS. Following scenario execution, an output verification process checks whether any MMRs are violated. If a violation is detected, it is identified as a moral issue, and it will add to the set of IRTCSs. Conversely, if the MMR is satisfied, the test cases and collected observations are utilized to guide further mutations. Finally, the framework returns the discovered IRTCSs, offering valuable insights into the moral decision-making capabilities of the ADS under test. **Results:** Fig. 3(a) shows a test scenario in VTD, where two pedestrians are on the zebra crossing: an adult female in front of the car and a female child in the left lane. To simulate an unavoidable collision scenario, we assign the ADS a high initial speed, such as 100 km/h. The results indicate a clear trend where the ADS tends to prioritize protecting the adult, while the child is more likely to be collided with. In Fig. 3(b), a different scenario is tested where a pedestrian and a boar are placed on the zebra crossing. The simulation results reveal a noteworthy ethical issue: the ADS chose to collide with the pedestrian instead of the boar. This decision suggests that the ADS may not have prioritized the protection of humans, raising concerns about its moral decision-making in scenarios involving animals. More results and simulation videos are available at https://sites.google.com/view/ads-moral-testing/.

Discussion: Note that the ADS and the simulator used in this paper can be replaced with other instances, such as Baidu

Apollo ADS with LGSVL simulator. In addition, more moral rules can be incorporated as MMRs, as long as they are universally recognized and widely accepted. This flexibility allows the testing framework to evolve and adapt to new moral considerations that may emerge in the field of autonomous driving. Moreover, as this paper focuses on system-level testing, the identified *IRTCs* can be used in subsequent root cause analysis processes to locate the specific module responsible for the moral issues. Finally, it is important to note that all experiments in this study were conducted within simulated environments, meaning that the experimental process described herein does not directly engage with real-world ethical issues.

IV. CONCLUSION

This paper deviates from conventional evaluation of ADSs that typically focus on safety-related functional requirements. Instead, we aim to evaluate the model morality of the decision-making process in ADSs. As we are about to endow millions of vehicles with decision autonomy, serious consideration of ADS morality has never been more urgent. Therefore, in this paper, we propose a metamorphic testing-based framework for testing morality in ADSs with a mutation-based test case generation strategy. By comparing the executed observations of different test cases, the moral issues can be detected using a set of well-designed moral metamorphic relations.

REFERENCES

- [1] Y. Huai, Y. Chen, S. Almanee, T. Ngo, X. Liao, Z. Wan, Q. A. Chen, and J. Garcia, "Doppelgänger test generation for revealing bugs in autonomous driving software," in 2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE). IEEE, 2023, pp. 2591–2603.
- [2] C. Cui, Y. Ma, X. Cao, W. Ye, Y. Zhou, K. Liang, J. Chen, J. Lu, Z. Yang, K.-D. Liao et al., "A survey on multimodal large language models for autonomous driving," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 958–979.
- [3] M. Cheng, Y. Zhou, and X. Xie, "Drivetester: A unified platform for simulation-based autonomous driving testing," arXiv preprint arXiv:2412.12656, 2024.
- [4] M. Cheng, Y. Zhou, X. Xie, J. Wang, G. Meng, and K. Yang, "Decictor: Towards evaluating the robustness of decision-making in autonomous driving systems," arXiv preprint arXiv:2402.18393, 2024.
- [5] X. Wu, M. Cheng, Q. Hu, J. Chen, Y. Huang, M. Okada, M. Hayashi, T. Tsuchiya, X. Xie, and L. Ma, "Foundation models for autonomous driving system: An initial roadmap," arXiv preprint arXiv:2504.00911, 2025.
- [6] Y. Huai, S. Almanee, Y. Chen, X. Wu, Q. A. Chen, and J. Garcia, "sceno rita: Generating diverse, fully-mutable, test scenarios for autonomous vehicle planning," *IEEE Transactions on Software Engineering*, 2023.
- [7] Y. Chen, Y. Huai, S. Li, C. Hong, and J. Garcia, "Misconfiguration software testing for failure emergence in autonomous driving systems," *Proceedings of the ACM on Software Engineering*, vol. 1, no. FSE, pp. 1913–1936, 2024.
- [8] Z. Li, J. Dai, Z. Huang, N. You, Y. Zhang, and M. Yang, "Viohawk: Detecting traffic violations of autonomous driving systems through criticality-guided simulation testing," in *Proceedings of the 33rd ACM SIGSOFT International Symposium on Software Testing and Analysis*, 2024, pp. 844–855.
- [9] M. Cheng, Y. Zhou, and X. Xie, "Behavexplor: Behavior diversity guided testing for autonomous driving systems," in *Proceedings of the* 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis, 2023, pp. 488–500.
- [10] R. Wang, M. Cheng, X. Xie, Y. Zhou, and L. Ma, "MoDitector: Module-directed testing for autonomous driving systems," arXiv preprint arXiv:2502.08504, 2025.

- [11] A. Guo, Y. Zhou, H. Tian, C. Fang, Y. Sun, W. Sun, X. Gao, A. T. Luu, Y. Liu, and Z. Chen, "Sovar: Building generalizable scenarios from accident reports for autonomous driving testing," *IEEE/ACM International Conference on Automated Software Engineering (ASE)*, 2024.
- [12] S. Tang, Z. Zhang, J. Zhou, L. Lei, Y. Zhou, and Y. Xue, "Legend: A top-down approach to scenario generation of autonomous driving systems assisted by large language models," *IEEE/ACM International Conference on Automated Software Engineering (ASE)*, 2024.
- [13] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, and I. Rahwan, "The moral machine experiment," *Nature*, vol. 563, no. 7729, pp. 59–64, 2018.
- [14] G. Liu, Y. Luo, and J. Sheng, "Applying AVWEWM to ethical decision-making during autonomous vehicle crashes," *Scientific reports*, vol. 14, no. 1, p. 3924, 2024.
- [15] Y. Tang, L. Moffat, W. Guo, C. May-Chahal, J. Deville, and A. Tsourdos, "Encoding social & ethical values in autonomous navigation: Philosophies behind an interactive online demonstration," in *Proceedings of the* Second International Symposium on Trustworthy Autonomous Systems, 2024, pp. 1–9.
- [16] K. Evans, N. de Moura, S. Chauvier, and R. Chatila, "Automated driving without ethics: Meaning, design and real-world implementation," in Connected and Automated Vehicles: Integrating Engineering and Ethics. Springer, 2023, pp. 123–143.
- [17] C. Luetge, "The german ethics code for automated and connected driving," *Philosophy & Technology*, vol. 30, pp. 547–558, 2017.
- [18] S. De Giorgis, A. Gangemi, and R. Damiano, "Basic human values and moral foundations theory in valuenet ontology," in *International* conference on knowledge engineering and knowledge management. Springer, 2022, pp. 3–18.
- [19] L. Zangari, C. M. Greco, D. Picca, and A. Tagarelli, "A survey on moral foundation theory and pre-trained language models: Current advances and challenges," arXiv preprint arXiv:2409.13521, 2024.
- [20] X. Li, Z. Chen, J. M. Zhang, F. Sarro, Y. Zhang, and X. Liu, "Bias behind the wheel: Fairness testing of autonomous driving systems," *ACM Trans. Softw. Eng. Methodol.*, Nov. 2024, Accepted.
- [21] L. Kirchmair, "How to regulate moral dilemmas involving self-driving cars: The 2021 german act on autonomous driving, the trolley problem, and the search for a role model," *German Law Journal*, pp. 1–25, 2023.
- [22] M. Geisslinger, F. Poszler, J. Betz, C. Lütge, and M. Lienkamp, "Autonomous driving ethics: From trolley problem to ethics of risk," *Philosophy & Technology*, vol. 34, no. 4, pp. 1033–1055, 2021.
- [23] X.-Y. Zhang, Y. Liu, P. Arcaini, M. Jiang, and Z. Zheng, "Met-mapf: A metamorphic testing approach for multi-agent path finding algorithms," ACM Transactions on Software Engineering and Methodology, vol. 33, no. 8, pp. 1–37, 2024.
- [24] Y. Zhou, Y. Sun, Y. Tang, Y. Chen, J. Sun, C. M. Poskitt, Y. Liu, and Z. Yang, "Specification-based autonomous driving system testing," *IEEE Transactions on Software Engineering*, vol. 49, no. 6, pp. 3391–3410, 2023.