Journal of Artificial Intelligence Practice (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/jaip.2024.070313
ISSN 2371-8412 Vol. 7 Num. 3

Ethical Implications of AI in Autonomous Systems: Balancing Innovation and Responsibility

Yuanxi Xu[1,*], Yuhe Zhu[2]

[1]High School Affiliated to Northwest Normal University, 21 Shilidian South Street, Anning District, Lanzhou City, Gansu Province, 730030, China
[2]Forsyth Country Day School, 5501 Shallowford Road, Lewisville, North Carolina, 27023, United States
[*]Corresponding author

Keywords: Autonomous Systems, Safety Assurance, Semantic Gap, Responsibility Gap, Risk Management

Abstract: This study integrates insights from systems engineering, ethics, and law to create a unified framework for addressing the complex challenge of ensuring the safety of autonomous systems. The emphasis is on identifying the "gaps" that emerge throughout the development process: the semantic gap, where there is an absence of standard criteria for fully specifying intended functionalities; the responsibility gap, where typical conditions for attributing moral responsibility to human agents for potential harm are missing; and the liability gap, where the usual mechanisms for providing compensation to those affected by harm are inadequate. By categorizing these "gaps," we can more accurately identify critical sources of uncertainty and risk in autonomous systems, which can guide the creation of more comprehensive safety assurance models and enhance risk management strategies.

1. Introduction

Autonomous systems challenge traditional approaches to system design, moral responsibility, legal liability, and safety assurance. These systems, capable of making independent decisions, vary in autonomy: from fully autonomous systems handling both decision-making and action implementation to advisory systems where humans remain responsible for executing decisions.

This paper examines the "gaps" these systems create in the design and implementation phases: the semantic gap (a disconnect between intended and specified system functionality), the responsibility gap (difficulty in assigning moral responsibility for system decisions), and the liability gap (issues in holding manufacturers, operators, or users liable for harm caused by the system)[1].

These gaps stem from the complexity and unpredictability of both the systems and their operational environments, alongside the increasing delegation of decision-making to the systems themselves. Addressing these gaps is crucial for improving safety assurance in autonomous systems[2].

The paper focuses on autonomous systems using machine-learning models and discusses the implications of accidents and injuries caused by these systems. It also proposes solutions for narrowing these gaps, including case studies and approaches to enhance safety assurance[3].

2. Methodology and Data

This study adopts a multi-disciplinary approach, integrating perspectives from systems engineering, ethics, and law to address the challenges posed by autonomous systems. The methodology is structured around identifying and analyzing three key "gaps"—the semantic gap, the responsibility gap, and the liability gap—within the development and deployment of autonomous systems. The analysis involves:

Literature Review: A comprehensive review of existing literature in systems engineering, ethical frameworks, and legal standards related to autonomous systems. This review helps in defining the gaps and understanding their implications on safety assurance, moral responsibility, and legal liability [4].

Case Studies: Two case studies are conducted to illustrate the practical implications of the identified gaps. The first case study examines a highly automated driving system, highlighting issues related to the semantic gap during system design and deployment. The second case study focuses on a clinical advisory system, where the responsibility gap is explored in the context of decision-making processes and moral accountability.

Reflective Equilibrium: For narrowing the responsibility gap, the study employs the method of reflective equilibrium, a philosophical approach used to achieve coherence between moral principles and specific judgments. This method is applied during the design phase of autonomous systems to ensure that the assignment of moral responsibility is both reasonable and justifiable[5].

3. Results and Analysis

In recent years, many manufacturers have been advancing the development of highly automated driving systems. It provides an overview of the core components involved in such systems. These systems typically include sensing elements such as radar, lidar, and cameras, but may also incorporate indirect contextual data from sources like digital maps and vehicle-to-infrastructure networks. The "Understanding" components analyze sensory data to interpret the current driving environment, including vehicle position, trajectory, and the movements of other road users. Decision-making components then formulate driving strategies based on these interpretations and specific driving objectives (e.g., traveling from point A to point B). Finally, the "Action" components implement these strategies through various vehicle controls such as brakes, engine, and steering.

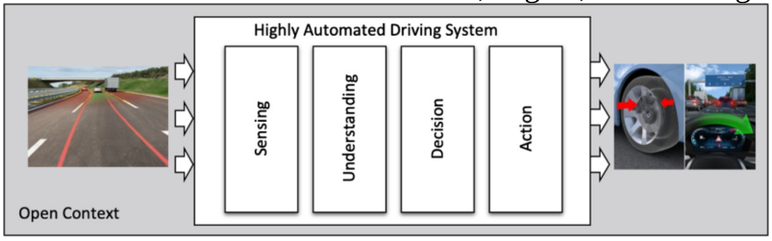


Figure 1: The Balance between Innovation and Ethics

Recent progress in machine learning and the availability of more powerful computing resources have enhanced the systems' ability to handle "Understanding" and "Decision" tasks in diverse environments. Deep neural networks now efficiently process unstructured data in real-time, achieving object recognition and classification accuracy that often exceeds human performance (Figure1)[6].

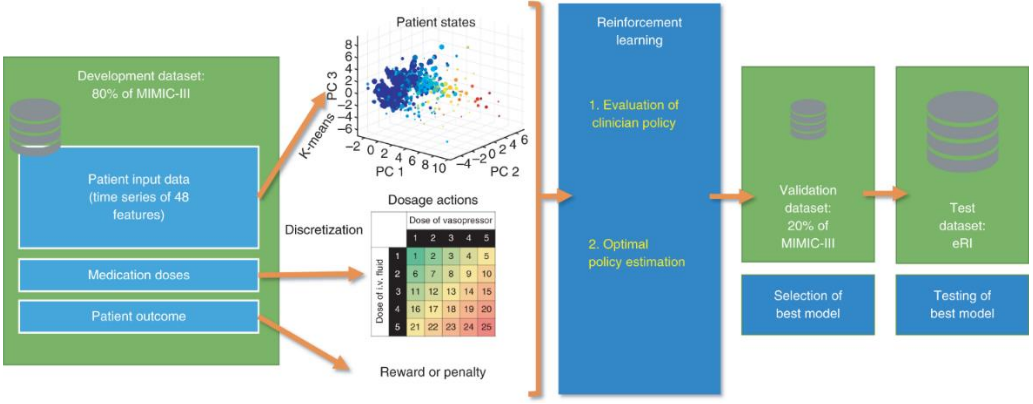
Currently, there are two main approaches to highly automated driving. The first is SAE Level 3, known as Conditional Driving Automation Systems. These systems manage vehicle control on highways but require the human driver to be ready to take over if needed or if the vehicle reaches the limits of its operational design domain. These systems represent an evolution of existing driver assistance technologies, with which users are somewhat familiar[7].

The second approach is SAE Level 4, or "High Driving Automation," which applies to urban environments. These systems handle Dynamic Driving Tasks (DDT) and Object and Event Detection and Response (OEDR) in complex settings, with fallback mechanisms in place for system failures. This approach signifies a more transformative shift in mobility, potentially introducing new vehicle types like autonomous shuttles or delivery vehicles. Although these vehicles will operate at lower speeds compared to SAE Level 3, they face greater technical and safety challenges.

The shift from manual driver assistance (SAE Levels 1-2) to fully autonomous driving (SAE Levels 3-5) underscores the challenge of the Semantic Gap. SAE Level 4 exemplifies how the three fundamental issues discussed in Section 2 contribute to this gap[8].

Firstly, urban environments are intricate and dynamic, populated by diverse and unpredictable traffic participants with varying capabilities, sizes, speeds, and trajectories. Additionally, as new modes of urban transport emerge and human behavior evolves in response to autonomous vehicles, these environments will continue to change. Consequently, it's impractical to define a comprehensive specification for urban automated driving.

Secondly, the complexity of systems required to handle urban driving tasks increases. Multiple sensor types—such as cameras and radars—must be integrated to counteract their individual limitations. These sensors' data must be synthesized and interpreted to create an accurate environmental model, which informs the development of an optimal driving strategy. Algorithms must also become more sophisticated to interpret the environment, predict traffic participants' intentions, and make decisions that minimize overall risk to occupants, other road users, and the environment (Figure2).



	1	2	3	4	5
1	1	2	3	4	5
2	6	7	8	9	10
3	11	12	13	14	15
4	16	17	18	19	20
5	21	22	23	24	25

Figure 2: Key Ethical Challenges in Autonomous Systems

Machine learning techniques like Deep Learning and Reinforcement Learning could potentially address the Semantic Gap by enabling systems to process unstructured data from complex environments and continuously improve their functions based on real-world stimuli. However, these methods come with limitations. Machine learning often provides probabilistic rather than definitive answers—e.g., a video frame might show an 83% probability of a pedestrian being present, but the next frame could change this probability significantly. Moreover, the decision-making processes of these models are opaque, leading to a "no free lunch" scenario where the challenge shifts from specifying behavior to proving that the learned behavior aligns with intended outcomes[9].

The third issue related to SAE Level 4 systems is the significant shift in decision-making responsibility. Without a human backup driver, the system must be explicitly defined for every possible situation, even when a completely safe outcome is unattainable. This entails a substantial challenge due to the system's uncertainty and the complexity of its decision-making algorithms. Ensuring safety for highly automated systems requires developing and validating safety requirements that address scenarios traditionally managed by human drivers. This introduces a new class of safety requirements, including higher component reliability, as the system cannot be deactivated by a human driver in the event of a fault.

International safety standards, such as ISO 26262, address vehicle hazards resulting from malfunctions and ensure hardware reliability and fault tolerance. However, these standards are primarily focused on driver assistance rather than fully automated driving systems. Consequently, new approaches, including the "Safety of the Intended Functionality" (SOTIF), need to be developed to address the unique challenges of autonomous vehicles.

ISO 26262 mandates a safety case—a structured, evidence-based argument demonstrating that a system is safe for its intended function across its operational context. For autonomous driving systems, this safety case must address the coplexities and uncertainties of the driving domain, as well as potential system limitations. The recent fatal accidents involving such systems highlight the urgent need for consensus on acceptable residual risk levels and underscore the potential of autonomous vehicles to enhance road safety[10].

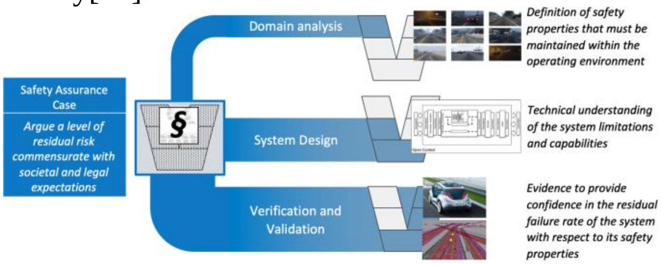

Domain analysis
System Design
Verification and Validation
Safety Assurance Case
Argue a level of residual risk commensurate with societal and legal expectations
§
Definition of safety properties that must be maintained within the operating environment
Technical understanding of the system limitations and capabilities
Evidence to provide confidence in the residual failure rate of the system with respect to its safety properties

Figure 3: Key Ethical Challenges in Autonomous Systems

Several strategies are currently being explored to bridge the semantic gap in the development of highly automated vehicles, addressing the gap's three fundamental causes (Figure3).

To manage the complexity of the operating environment, approaches are being implemented that restrict functionality to well-defined scenarios where safety risks and system capabilities are well understood. For SAE Level 3 systems, this might involve operation on specific highway sections with limited weather conditions and functional constraints, such as prohibiting overtaking. For SAE Level 4 systems, operations are confined to geo-fenced urban areas with detailed maps and validation data. These restrictions, however, limit the scope of the intended functionality.

To handle the complexity of the systems, the deployment of machine learning algorithms is being restricted to narrowly defined and constrained functions with minimal safety impact. This includes employing parallel sensing methods and plausibility checks. Additionally, greater use of infrastructure, such as traffic signal status transmission, is planned to enhance environmental sensing robustness. While this approach reduces the validation burden on machine learning components, it increases the number of system components and overall costs.

In current systems, the delegation of decision-making to the automated systems is also being limited. This is achieved through measures such as requiring driver supervision in Level 3 highway automation, employing safety drivers during Level 4 urban automated driving tests, or defaulting to a safe state, like stopping at the roadside, in ambiguous or critical situations. These measures impose constraints on the intended functionality.

Addressing the semantic gap will not be confined to the design phase alone; it will be an ongoing process. Adjustments to functionality will be made as systems are tested in real-world conditions and as understanding of cultural differences in safety attitudes and usage evolves. The compromises mentioned may be progressively relaxed as more validation evidence becomes available, technological advancements occur, and a more dynamic and robust safety case is developed. This safety case must cover hazard avoidance due to the automated driving function, responses to system failures, prevention or mitigation of misuse, and reactions to critical situations involving other road users.

Currently, there is no consensus on how to effectively reduce the semantic gap and develop a robust safety case for highly automated driving. Governments and regulators are still working to define the appropriate legal frameworks, and significant clarification is needed. Consequently, there is a pressing need for a unified language to articulate ethical and legal expectations for these systems.

4. Conclusion

Bridging the semantic gap in highly automated vehicles presents a complex challenge that involves addressing the intricate interplay between environment complexity, system design, and decision-making delegation. Current strategies aim to mitigate these challenges by constraining operational scenarios, managing system complexity through limited machine learning applications, and reducing decision-making delegation. While these approaches help in controlling the immediate risks and uncertainties, they also impose restrictions on the functionality and increase overall system costs. As the technology evolves and more real-world data becomes available, it will be crucial to iteratively refine these strategies. The process of addressing the semantic gap is not static but will require ongoing adjustments as our understanding of automated systems improves and as technological advancements are made. Furthermore, developing a robust safety case will involve not only demonstrating hazard avoidance and system reliability but also addressing cultural variations in safety expectations and usage. The current lack of consensus on the best methods to close the semantic gap and structure a compelling safety case underscores the need for continued dialogue among governments, regulators, and industry stakeholders. Establishing a common framework for articulating ethical and legal expectations will be essential for advancing the deployment of highly automated vehicles. Future discussions and developments must focus on creating a unified language to navigate these complexities and ensure that safety and functionality are upheld in the evolving landscape of automated driving.

References

[1] Lin, P., & Abney, K. Robot Ethics: The Ethical and Social Implications of Robotics. In The Cambridge Handbook of Artificial Intelligence. Cambridge University Press, 2017, pp. 281-307.

[2] Binns, R., Veale, M., Shadbolt, N., & Shadbolt, N. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. ACM, 2018, pp. 1-14.

[3] Calo, R. The Case for a Federal Robotics Commission. Harvard Journal of Law & Technology, 2016, 29(2), 429-466.

[4] Crawford, K., & Paglen, T. Excavating AI: The Politics of Images in Machine Learning Training Data. AI & Society, 2019, 34(1), 135-150.

[5] Dastin, J. Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women. Reuters. 2018. Retrieved from https://www.reuters.com/article/us-amazon-com-recruitment-insight-idUSKCN1MK08G

[6] Gogoll, J., & Müller, J. F. Legally and Ethically Admissible AI: Challenges and Solutions. Artificial Intelligence Review, 2017, 50(3), 227-249.

[7] Heath, J. Ethical Implications of AI in Autonomous Vehicles: Balancing Innovation with Responsibility. Journal of Business Ethics, 2019, 160(2), 343-356.

[8] Kumar, P., & Singh, P. Machine Learning in Autonomous Systems: Ethics and Regulation. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(6), 2110-2122.

[9] Lin, P., & Bekey, G. A. The Ethics of Autonomous Cars. The Atlantic. 2012. Retrieved from https://www. theatlantic.com/technology/archive/2012/10/the-ethics-of-autonomous-cars/263786/

[10] Raji, I. D., & Buolamwini, J. Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. ACM. 2019, pp. 1-15.