http://www.diva-portal.org

Postprint

This is the accepted version of a paper presented at *16th IEEE Vehicular Networking Conference, VNC 2025, Porto, Portugal, 2-4 June, 2025*.

Citation for the original published paper:

Kochenborger Duarte, E., Pignaton de Freitas, E., Bellalta, B., Vinel, A. (2025)
Ethical Social Robot Moderators for Traffic Management: Integrating Automated Vehicles and Vulnerable Road Users
In: Ana Aguiar; Takamasa Higuchi; Susana Sargento; Alexey Vinel; Agon Memedi (ed.), *IEEE Vehicular Networking Conference, VNC* (pp. 1-8). Piscataway, NJ: IEEE
https://doi.org/10.1109/VNC64509.2025.11054145

N.B. When citing this work, cite the original published paper.

Permanent link to this version:
http://urn.kb.se/resolve?urn=urn:nbn:se:hh:diva-57097

Ethical Social Robot Moderators for Traffic Management: Integrating Automated Vehicles and Vulnerable Road Users

*

Eduardo K. Duarte[1], Edison Pignaton de Freitas[1,2], Boris Bellalta[3], Alexey Vinel[1,4]

[1]Halmstad University, Halmstad, Sweden
[2]Federal University of Rio Grande do Sul, Porto Alegre, Brazil
[3]Universitat Pompeu Fabra, Barcelona, Spain
[4]Karlsruhe Institute of Technology, Karlsruhe, Germany
Emails: eduardo.kochenborger-duarte@hh.se, edison.pignaton@hh.se, boris.bellalta@upf.edu, alexey.vinel@kit.edu

Abstract

Urban traffic environments are rapidly evolving with the adoption of connected and automated vehicles (CAVs), yet challenges remain regarding interactions between these vehicles and vulnerable road users (VRUs). This paper introduces the concept of an *Ethical Social Robot Moderator (ESRM)* that facilitates coordination and communication in mixed-traffic contexts. By consolidating insights from research on robot trust and tele-operation, vehicular communication systems, and ethical frameworks for autonomous driving, the proposed ESRM aims to reduce collision risks, enhances cooperation among heterogeneous road users, and provides transparent decision-making based on well-defined moral principles. The proposal is a structured design for the ESRM, extending simulation strategies in the CARLA environment with detailed metrics, and integrating references from prior literature to illustrate how these social robots can bridge multiple technological domains. This work serves as a unifying contribution to a broader field that spans robotics, communication networks, and ethical AI in urban mobility.

Keywords: V2X, traffic light, C-ITS, emergency vehicle, ESRM, moral reasoning, privacy, liability

1 Introduction and Motivation

The integration of connected and automated vehicles (CAVs) into urban environments promises significant improvements in road safety and efficiency. However, the coexistence of CAVs with human-driven vehicles and vulnerable road users (VRUs) such as pedestrians, cyclists, and micro-mobility users presents complex challenges [1]. While CAVs employ sophisticated onboard sensors and algorithms, they can struggle in ethically fraught or unpredictable scenarios, particularly in non-signalized intersections or high-density pedestrian zones [2].

Furthermore, recent studies have underscored the critical importance of ethical considerations in autonomous driving [3]. The normative impacts of social robotics developers on their

*This work has been funded as part of the KIT Future Fields project "V2X4Robot". This paper is part of the CulturalRoad project, funded by the European Union under grant agreement No. 101147397. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them.

creations have been explored, highlighting the need for diverse perspectives in the design process [4]. Building on this, on [5] the authors proposed an ethical cooperative protocols for multi-vehicle collision mitigation, which aligns closely with the proposed framework.

In response to these challenges, this paper introduces and refines the concept of *Ethical Social Robot Moderators (ESRMs)*. These innovative systems are designed to oversee and guide traffic participants by integrating an ethical decision-making framework with social robotic interaction and robust communication networks. Positioned at strategic urban locations, ESRMs can:

- Direct vehicles and provide signals to VRUs
- Process real-time data from vehicular networks to minimize conflicts
- Make ethically-informed decisions to balance safety, efficiency, and fairness

The proposed ESRM framework builds upon and extends prior work in several key areas:

- Robot trust and tele-operation [6]
- Vehicular communication for rapid emergency response [7]
- Ethical cooperation in autonomous driving [5]

The core contribution of this paper is a comprehensive ESRM framework, featuring:

- An ethical decision-making module that integrates recent advancements in AI ethics and risk assessment, featuring a quantifiable approach to ethical ncence in pre-crash situations.
- A novel multi-objective utility function that addresses safety, fairness, and throughput
- Integration of transparency and auditing mechanisms to sustain public trust
- Detailed simulation scenarios using CARLA, capturing VRU unpredictability and communication latencies
- Extensive performance metrics

By synthesizing these elements, this paper presents an ESRM framework that is not only conceptually robust but also pragmatic for near-future implementations. This work aims to address the complex ethical and practical challenges of integrating autonomous systems into diverse urban environments, paving the way for safer, more efficient, and ethically sound traffic management.

2 Related Work

2.1 Ethical and Cooperative Autonomous Driving

Research on ethical autonomous driving frequently confronts dilemmas akin to the trolley problem, yet real-world traffic presents more nuanced, everyday decision points where vehicle and pedestrian needs may conflict [1,3]. A large-scale study by Awad et al. involving 70,000 participants from 42 countries revealed significant cross-cultural variations in moral decisions related to AVs [8]. Furthermore, Lim and Taeihagh point out that the implementation of AVs in smart cities introduces additional layers of complexity in algorithmic decision-making [9]. To address these multifaceted issues, Borenstein et al. argue for a systems-level approach to ethical considerations in AVs, emphasizing interactions between autonomous systems, human behavior, and broader socio-technical infrastructures [10].

Sidorenko et al. (2024) introduced an innovative ethical cooperative driving framework that enables vehicles to coordinate their actions in emergency scenarios [5]. Their approach allows

vehicles to share information about sudden maneuvers, such as emergency braking, to reduce collision risks and severity.

Some studies in this domain (e.g., [11, 12]) illustrate how machine ethics and motion planning intersect to address both moral and operational aspects of autonomous decision-making. Building on these insights, the framework proposed in this paper adds a dedicated social robot dimension that enforces and communicates ethically guided directives at the infrastructure level.

Recent research has made significant strides in developing frameworks for integrating human values into AV design. Gerdes et al. propose a comprehensive framework that aims to align AV behavior with human values and societal norms [13]. However, Martinho et al. highlight a concerning gap between academic discussions on AV ethics and the focus of industry practitioners [14]. This disconnect shows the need for greater collaboration between researchers and industry stakeholders. Additionally, Dogan et al. emphasize the crucial role of public perception in shaping the ethical landscape of automated mobility [15].

2.2 Social Robots and Human-Robot Interaction

Social robots deployed in public spaces have been studied for tasks such as providing wayfinding information, entertaining passersby, or assisting in education. Their application to traffic guidance is less common but growing, particularly in controlled scenarios or via Virtual Reality simulations to gauge user acceptance [16]. These studies indicate that anthropomorphic cues and clear signals can help humans interpret a robot's instructions, potentially enhancing cooperation at busy intersections.

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems has developed comprehensive guidelines for ethical considerations in AI and robotics. Their "Ethically Aligned Design" document, currently in its second version (EADv2) [17], provides a framework for prioritizing human well-being in the development and deployment of autonomous systems. This initiative emphasizes key principles such as transparency, accountability, and privacy. For ESRMs in traffic management, these guidelines suggest the importance of not only issuing commands but also communicating the rationale behind decisions, such as prioritizing certain road users. This approach aims to enhance public trust and acceptance of automated systems in urban environments.

2.3 Previous Work

A previous work on trust in robot self-defense [6] demonstrates that tele-operation or partial human oversight can significantly enhance user confidence in a robot's actions, particularly in ethically sensitive contexts. This finding has important implications for traffic management, where errors can lead to accidents or disruptions. By incorporating tele-operation capabilities into ESRMs, it is possible to potentially reassure the public and provide a fail-safe mechanism for handling unanticipated edge cases. Tele-operation interfaces can also log and justify key decisions, aligning with transparency requirements outlined in ethical frameworks for AI and robotics. This approach is particularly relevant to the proposed ESRM framework, as it allows for human intervention in complex ethical scenarios.

The effectiveness of ESRMs also heavily relies on rapid and reliable data exchange among vehicles and infrastructure. The work on the SafeSmart system highlights how VANET architectures can significantly reduce emergency response times by using a traffic light preemption system [7, 18]. An ESRM placed at a busy intersection could similarly connect to nearby vehicles, pulling real-time data to resolve conflicts more effectively than a single vehicle or static traffic light could manage. This capability is crucial for implementing the multi-objective utility function proposed in the ESRM framework, which balances safety, fairness, and throughput. Moreover, by integrating V2X communication, ESRMs can address the needs of VRUs more

effectively, aligning with the ethical considerations discussed in the extended simulation scenarios.

3 Ethical Social Robot Moderators: Framework and Integration

3.1 Conceptual Overview

An Ethical Social Robot Moderator (ESRM) is a robotic platform - potentially mobile, but stationed at strategic urban locations - that receives a continuous stream of data from sensors and connected infrastructure. The ESRM wields three main functions:

1. **Monitoring**: Detect traffic flows, identify VRUs, and predict potential conflicts using hybrid sensor fusion (LiDAR, cameras, V2X).

2. **Ethical Decision-Making**: Balance safety, fairness, and throughput via a multi-objective utility function.

3. **Communication**: Issue real-time instructions to vehicles and VRUs using both digital signals (V2X messages) and physical cues (lights, gestures, or displays).

Whereas an intersection controlled by a standard traffic light adheres to rigid cycles, the ESRM dynamically adapts. For example, if a group of schoolchildren approaches a crosswalk, the ESRM could override a scheduled traffic phase to expedite their crossing, explicitly expressing their increased vulnerability. Alternatively, the ESRM may coordinate with an autonomous ambulance to preemptively halt certain directions of traffic, thus clearing a path in a manner that appears "intelligent" and empathic, which can foster trust and achieve better results.

Figure 1 illustrates the framework of the proposed ESRM. The system begins with multi-source data fusion (e.g., LiDAR, cameras or V2X). Real-time conflict prediction identifies potential conflict scenarios. Then, a multi-objective utility function is used to balance the different attributes of decision-making. Dynamic action selection and multi-modal communication ensure real-time interaction with vehicles and vulnerable road users, while cultural adaptation adjusts decision parameters based on societal norms. In high-risk or novel scenarios, the system invokes tele-operation fallback to ensure human oversight.

A key practical concern with tele-operation fallback is the latency introduced when human operators intervene in high-risk scenarios. Human-in-the-loop systems for critical applications could experience delays of from one to a few seconds, depending on network infrastructure and operator readiness [19]. For ESRMs, such delays may be tolerable for proactive interventions (e.g., gradually slowing or halting traffic), but less viable for reactive split-second decisions (e.g., immediate collision avoidance).

To mitigate these challenges, ESRMs can deploy early detection and escalation mechanisms. For example, when the system's conflict prediction module detects rising uncertainty or risk indicators—such as unexpected VRU behavior, communication dropout, or ambiguous sensor readings—it can preemptively alert human operators before a full-blown emergency emerges. This "early flagging" allows tele-operators to monitor in standby mode and shortens reaction times once a decision point is reached. Another avenue is using semi-autonomous intervention templates—predefined safety protocols that the human operator can confirm or modify rapidly (e.g., "halt all traffic in X direction" button), reducing cognitive load and decision-making time.

These strategies collectively enhance the feasibility and safety of tele-operation by minimizing reaction delays; however, it remains essential to recognize that human oversight introduces inherent latency, and thus, tele-operation is most appropriate for scenarios where sufficient lead time is available for safe intervention.

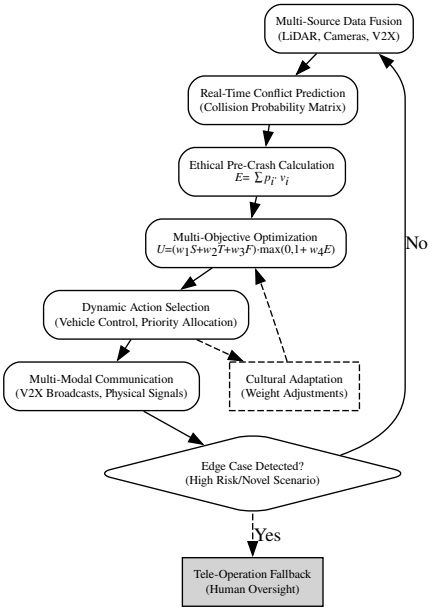

Figure 1: Decision-making framework of the Ethical Social Robot Moderator

3.2 Ethical Decision-Making Module

An essential feature of ESRMs that sets them apart from traditional infrastructure is their explicit moral reasoning process, which considers context-specific data. Recent research [5] introduced an innovative ethical cooperative driving framework that enables vehicles to coordinate their actions in emergency scenarios. This approach suggests that ESRMs could employ a combination of rule-based logic for non-negotiable ethical constraints (e.g., prioritizing pedestrian safety) and machine learning components for optimizing local traffic flow. Furthermore, any ethical framework for autonomous systems must account for complex trade-offs. The German Ethics Code for Automated and Connected Driving [20], emphasizes the importance of balancing safety, mobility, and fairness in autonomous driving systems. This aligns with the ESRM's need to consider multiple factors in its decision-making process. Therefore, one of the challenges is how to quantify those factors.

3.2.1 Quantifying Fairness, Equity, and Ethical Pre-Crash Moderation

To address real-time decisions that balance efficiency, safety, and ethical considerations in pre-crash scenarios, this work proposes an enhanced multi-objective optimization approach for the ESRM's decision-making module. An extended weighted utility function can be defined as follows:

$$U = (w_1(t)S + w_2(t)T + w_3(t)F) \cdot \max(0, 1 + w_4(t)E) \tag{1}$$

where:

- S (Safety Factor): Measures estimated collision probabilities, normalized to [0, 1]
- T (Throughput Factor): Accounts for average vehicle flow, normalized to [0, 1]
- F (Fairness Factor): Tracks distribution of waiting times among different user groups, normalized to [0, 1]
- E (Ethical Pre-Crash Factor): Quantifies the ethical implications of potential crash scenarios

The weights $(w_1(t), w_2(t), w_3(t))$ reflect societal, municipal, or context-specific priorities, with their sum equal to 1, at a particular instant t. The ethical scaling parameter $(w_4(t))$ is independent and determines how strongly ethics modulate utility via the veto mechanism. This formulation ensures that a highly negative E can significantly reduce the overall utility, preventing ethically problematic decisions from being considered optimal. The Ethical Pre-Crash Factor is further defined as:

$$E = \frac{2 \cdot \sum_{i=1}^{n} p_i \cdot v_i}{\sum_{i=1}^{n} p_i} - 1 \tag{2}$$

where p_i is the probability of crash scenario i, and v_i is the ethical valence of that scenario, based on the definitions by Evans et al [21]. With the utility function using $max(0, 1 + w_4 E)$, it is possible to achieve an ethical veto by setting its weight $w_4(t)$ appropriately. For instance, if $w_4(t) = 1$, then any $E \leq -1$ will result in $max(0, 1 + w4(t)E) = 0$, effectively vetoing the decision, which can be useful for sensitive scenarios.

The scaling applied in the Ethical Pre-Crash Factor formula (stretching the probability-weighted average of valences by a factor of 2 and shifting by -1) intentionally expands the lower bound of E to -3. This design reflects a conservative ethical stance: it amplifies the impact of highly negative ethical scenarios, ensuring that the veto mechanism decisively suppresses utility when ethically unacceptable risks are present. By weighting severe ethical violations more heavily than positive outcomes, the framework prioritizes harm avoidance above potential efficiency gains, aligning with safety-first principles commonly emphasized in autonomous system ethics.

While safety quantifies the immediate, physical risk to road users, such as the probability and severity of collisions, ethics captures broader moral and societal considerations, which may amplify or suppress actions based on context-specific values, such as prioritizing vulnerable populations or avoiding disproportionate harm, even when raw safety metrics are similar. While high safety typically aligns with ethical behavior, the ethical scaling factor allows the system to recognize cases where societal values demand stricter or more cautious action than safety metrics alone would dictate.

One example of a decision flow could be as follows: the ESRM detects conflicting demands (such as a large group of VRUs arriving at the same time as a row of high-speed automated cars) and then computes the scenario's utility based on Eq. (1). A lower collision probability (boosting w_1) may impose stricter constraints on vehicle speeds or red lights, while fairness considerations (boosting w_3) ensure that no user group experiences excessively longer wait times than others. The ESRM then selects the action that maximizes U based on its predefined parameters. Tele-operational oversight can be invoked in edge cases where potential harm is high or when the scenario is novel to the machine-learning components.

3.2.2 ESRMs as Ethical Moderators in Pre-Crash Situations

Extending the work in [5] on ethical cooperative driving, this work proposes that ESRMs can serve as ethical moderators in pre-crash situations. By leveraging V2X communication and their strategic positioning, ESRMs can:

- Rapidly assess the ethical implications of potential crash scenarios using the Ethical Pre-Crash Factor
- Coordinate actions between multiple vehicles to minimize harm by optimizing the extended utility function (Equation 1)
- Provide real-time ethical guidance to autonomous vehicles based on the computed ethical valences
- Ensure fairness and transparency in critical decision-making processes through the explicit incorporation of the Fairness Factor

This approach allows ESRMs to make ethically-informed decisions in real-time, considering not only immediate safety and efficiency but also longer-term fairness and ethical implications.

3.2.3 Ethical Valence

The ethical valence (v_i) concept, inspired by Evans et al.'s Ethical Valence Theory (EVT) [21], provides a framework for quantifying ethical considerations in decision-making processes. While the original paper does not prescribe a specific numerical scale, their principles are here adapted to create a range from -1 (least ethically desirable) to +1 (most ethically desirable).

The valence is derived from two key components of EVT [21]:

1. **Claim Calculation**: Based on harm severity (h) and vulnerability factors (v) of road users.
$$C = f(h, v)$$
2. **Social Weighting**: Adjusting claims based on cultural and legal considerations (ω).
$$V = C \cdot \omega$$

These components are then normalized to the -1 to +1 scale, reflecting the relative ethical weight of outcomes. For example, in a potential crash scenario:

- Outcome A: Minor vehicle damage, no injuries ($v = 0.8$)
  - Low harm severity, low vulnerability factor
  - Positive weighting due to minimal impact
- Outcome B: Serious injury to a pedestrian ($v = -0.9$)
  - High harm severity, high vulnerability (pedestrian)
  - Strong negative weighting due to severity and vulnerable road user status
- Outcome C: Multiple minor injuries to vehicle occupants ($v = -0.5$)
  - Moderate harm severity, lower vulnerability (vehicle occupants)
  - Negative weighting, but less severe than pedestrian harm

The ESRM would use these valences, combined with the probability of each outcome, to calculate the Ethical Pre-Crash Factor (Equation 2):

$$\begin{aligned} E &= \frac{2 \cdot (0.6 \cdot 0.8 + 0.3 \cdot (-0.9) + 0.1 \cdot (-0.5))}{0.6 + 0.3 + 0.1} - 1 \\ &= \frac{2 \cdot 0.18}{1} - 1 = 0.36 - 1 \\ &= -0.64 \end{aligned} \tag{3}$$

This negative E value suggests an ethically unfavorable situation overall, primarily due to the presence of potential serious harm to a pedestrian, despite the higher probability of a minor outcome.

This approach allows the ESRM to make nuanced ethical decisions by considering not just the likelihood of outcomes, but also their ethical implications. For instance, even if outcome B (serious pedestrian injury) has a lower probability, its highly negative valence would significantly influence the ESRM's decision-making process, potentially leading it to prioritize actions that avoid this outcome.

By incorporating this ethical valence into its utility function, the ESRM can make decisions that align with societal values and ethical principles, going beyond simple crash avoidance to consider the broader implications of its actions on all road users.

3.2.4 Practical Examples of ESRM Decision-Making

Let us consider two scenarios to illustrate how the ESRM would operate in real-world situations:

Scenario 1: School Zone Crossing An ESRM is deployed near a school where a large group of children is about to cross a busy street at a specific instant t. The extended utility function would be heavily weighted towards safety and ethical considerations:

$$U = (0.5S + 0.3T + 0.2F) \cdot \max(0, 1 + 0.8E) \tag{4}$$

Here, $w_1(t) = 0.5$ (safety), $w_2(t) = 0.3$ (throughput), $w_3(t) = 0.2$ (fairness), which sum to 1, while $w_4(t) = 0.8$ is set independently as the ethical scaling factor. The high weight on safety (0.5) reflects the priority of protecting vulnerable road users, while the significant weight on ethics (0.8) ensures that decisions tend to align with societal values regarding child safety.

The ESRM detects children via LiDAR and cameras, prioritizes pedestrian safety by adjusting crossing times, alerts vehicles through V2X, and controls traffic lights to create a safe crossing window.

Scenario 2: Emergency Vehicle Override An ambulance is approaching a congested intersection at a specific instant t. The ESRM must quickly clear a path while minimizing risk to other road users. The utility function might be adjusted as follows:

$$U = (0.4S + 0.4T + 0.2F) \cdot \max(0, 1 + 0.6E) \tag{5}$$

In this scenario, $w_1(t) = 0.4$ (safety), $w_2(t) = 0.4$ (throughput), $w_3(t) = 0.2$ (fairness), and $w_4(t) = 0.6$ (ethics). The increased weight on throughput (0.4) reflects the need for efficient emergency vehicle passage, while maintaining a strong emphasis on safety (0.4) and ethical considerations (0.6). Maintaining fairness is not seen as a priority in this case.

The ESRM detects the ambulance via V2X and sensors, coordinates with nearby vehicles to clear a path, and temporarily suspends pedestrian crossings to ensure swift and safe passage.

3.2.5 Challenges in Collision Probability Calculation

The calculation of collision probabilities is a crucial aspect of the decision-making process, particularly in the context of ethical considerations. Various approaches have been proposed to estimate these probabilities accurately [22].

One method involves using sensor fusion techniques to combine data from multiple sources such as LiDAR, radar, and cameras. This approach enhances the accuracy of object detection and tracking, leading to more precise collision probability estimates.

Another approach utilizes Machine Learning algorithms to predict potential collisions based on historical data and real-time sensor inputs. Deep learning models have shown promise in accurately detecting objects and estimating their trajectories, which is essential for collision probability calculation.

Probabilistic models, including Markov processes and Bayesian networks, can also be employed to estimate collision risks. These models consider various factors such as vehicle dynamics, road conditions, and driver behavior to compute the likelihood of a collision.

The choice of method for calculating collision probabilities can significantly impact the ethical decision-making process of AVs. For example, a more accurate probability estimation could lead to better-informed choices in scenarios where the AV must decide between multiple potentially harmful outcomes. Nevertheless, the proposed framework is adaptable to different methods.

3.2.6 Challenges in Weight Determination

While the proposed utility function offers a flexible framework for ethical decision-making in ESRMs, the determination of appropriate weights (w_1, w_2, w_3, w_4) remains a significant challenge. Recent research has highlighted the complexity of this issue:

- [3] propose an ethics of risk approach that integrates multiple rules, suggesting that weights could be dynamically adjusted based on specific scenarios.
- [8] conducted a large-scale study across 42 countries, revealing cross-cultural variations in moral decisions related to autonomous vehicles. This suggests that weights might need to be informed by societal preferences, which can vary across different cultures and regions.
- The German Ethics Code for Automated and Connected Driving [23] recommends against deliberately programming vehicles to prefer one life over another, which complicates the process of weight assignment.
- [24] propose a "moral knob" architecture allowing for customization between "altruism" and "egoism," suggesting that weights could be partially determined by user preferences within certain bounds.

These diverse approaches show the ongoing debate in the field and the lack of consensus on a universal method for determining weights.

Regarding human oversight, [25] argue for a "human-in-the-loop" approach in AI ethics, which could be applied to weight determination in ESRMs. This aligns with the European Parliament's emphasis on human oversight in AI systems [26]. [27] further supports this idea, proposing a human-centered approach to AI that includes human control and oversight in critical decision-making processes.

Future work should focus on developing more transparent and justifiable methods for weight selection, potentially incorporating machine learning approaches to refine weights based on real-world data and outcomes, while maintaining human oversight. This could involve a hybrid system where AI suggests weight adjustments based on data analysis, but human experts make the final decisions. Additionally, multi-stakeholder collaboration involving city governments, private tech providers, and insurance experts could help establish more robust frameworks for weight determination [26]. It is important to note that once consensus is reached on how to determine these weights, the proposed utility function can be readily applied. This underscores the value of the framework as a flexible, adaptable, and future-proof tool for ethical decision-making in autonomous systems. The equation's structure allows for easy integration of agreed-upon weights, making it a practical solution for implementing ethical guidelines in real-world scenarios.

3.2.7 V2X Protocol Integration and Ethical Coordination

The proposed framework presents a conceptual approach for integrating with vehicular communication standards through a layered architecture that respects existing protocols while enabling ethical coordination. One way of integrating with standards such as ETSI ITS-G5 and IEEE WAVE is by utilizing strategic message extensions that could preserve backward compatibility while introducing ethical context parameters. One potential implementation could leverage optional data containers in Cooperative Awareness Messages (CAMs) to embed ethical priority flags and vulnerability indices without modifying core message structures.

This approach would align with ETSI EN 302 637-2 provisions for application-specific extensions [28], theoretically ensuring legacy systems could process essential safety data while ESRM-aware vehicles utilize the additional ethical context. For example, when detecting vulnerable road user clusters, an ESRM might extend standard DENM messages with conflict

scenario identifiers, potentially enabling coordinated vehicle responses without compromising network integrity.

Furthermore, a hybrid network architecture could enable ESRM operation across multiple communication technologies. Short-range DSRC might provide immediate vehicle coordination at intersections, while cellular networks (LTE/5G) could facilitate city-wide ethical policy synchronization. This dual-mode operation could prove particularly valuable during city-scale events, where ESRMs might temporarily prioritize emergency vehicle corridors using wide-area network coordination while maintaining local conflict resolution through direct vehicle communications. The proposed architecture should also incorporate automatic fallback mechanisms to traditional traffic signals during network outages, aiming to ensure continuous operation despite intermittent connectivity challenges.

3.2.8 Transparency and Cultural Acceptance

The ESRM's ethical decision-making process requires rigorous logging to ensure accountability, enable post-hoc audits, and maintain public trust. The system should provide measurable transparency by recording key decision parameters, including timestamped utility function values $(w_1(t)S, w_2(t)T, w_3(t)F)$, ethical pre-crash factor (E) calculations with scenario probabilities p_i and valences v_i, and contextual metadata (e.g., pedestrian density, vehicle velocities, sensor confidence intervals).

This audit trail aligns with the framework for value alignment verification proposed by Arnold et al. [29], enabling regulators to validate whether ESRM decisions adhere to programmed ethical constraints.

In addition to logging the ESRM's rationale, there is a need to address cultural norms and driving behaviors that vary internationally. For instance, cities where jaywalking is common may require the ESRM to adopt a more protective stance toward VRUs, whereas regions with strong traffic law compliance could place more emphasis on throughput. Adjusting w_1, w_2, w_3 based on local regulations or user feedback can help tailor the ESRM to diverse cultural settings.

4 Simulation Design

To evaluate the proposed system, the focus is on three core design principles for ethical evaluation:

- **Ethical Gradient Scaling**: Scenarios progress from routine trade-offs (efficiency vs pedestrian priority) to critical pre-crash situations
- **Cultural Parametrization**: Adjustable pedestrian compliance rates (e.g., 70-95%) and crossing assertiveness levels
- **VRU Heterogeneity**: Includes, mobility-impaired users, distracted pedestrians, and cargo bikes

This design proposes a suite of 6 CARLA [30] scenarios, each modeling a distinct urban configuration and variety of VRUs. These scenarios are designed to test the ESRM's ethical decision-making capabilities in various challenging situations:

1. **Basic Intersection**: Low traffic volume; tests baseline ESRM functions and ethical decision-making in simple scenarios.
2. **Moderate Density Downtown**: Mix of cars, cyclists, and pedestrians crossing irregularly, challenging the ESRM to balance safety and efficiency for diverse road users.

3. **High-Density Corridor**: Multiple lanes, frequent bus stops, and ride-share pickup points, testing the ESRM's ability to manage complex traffic flows while prioritizing public transit and shared mobility.

4. **School Zone**: Large groups of children crossing, emphasizing VRU protection and testing the ESRM's ethical prioritization of vulnerable users.

5. **Emergency Vehicle Override**: Simulation of ambulances or police vehicles requiring immediate right-of-way, testing the ESRM's decision-making in time-critical situations.

6. **Ethical Pre-Crash Scenario**: The crashing of multiple vehicles becomes unavoidable, and the ESRM must take action to evaluate different outcomes and minimize the overall harm.

4.1 Metrics and Evaluation Framework

There are many different ways to evaluate the ESRM's performance. Depending on specific requirements, it is possible that some aspects might be more important than others. Nonetheless, the designed evaluation proposes a set of metrics that allow for a comprehensive evaluation in an extensive set of scenarios.

To thoroughly assess the ESRM's performance, the approach is the employment of a multi-faceted evaluation framework that aligns with the ethical decision-making objectives:

4.1.1 Safety Metrics

- **Collision Rate**: The number of collisions during the simulation.
- **Near-Miss Rate**: The number of incidents where vehicles or VRUs come within a dangerously close distance (e.g., under 0.75m) without colliding.
- **Time-to-Collision (TTC)**: Minimum TTC values recorded during simulations, as well as time-integrated Time-to-Collision (TIT).

4.1.2 Efficiency Metrics

- **Throughput**: Overall vehicle throughput (e.g., vehicles/min) through the monitored area.
- **Average Travel Time**: Mean travel time for all road users through the simulated area.
- **Delay Reduction**: Comparison of travel times with and without ESRM intervention.

4.1.3 Fairness and Ethical Metrics

- **Fairness Index**: one possible way to calculate the fairness in a specific situation is:

$$F = 1 - \frac{\max(T_{wait,i}) - \min(T_{wait,i})}{\max(T_{wait,i})} \tag{6}$$

where $T_{wait,i}$ is the waiting time for road user i. The value of F is logged for each distinct user class (cars, buses, cyclists, pedestrians) to identify potential inequities, such as spikes in waiting times within the same class (intra-class fairness), for example. This formula is simple, but it also has its shortcomings, such as considering only extremes (minimum and maximum) and not allowing for comparisons between different classes (inter-class fairness).

Another option is:

$$F = 1 - \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |T_i - T_j|}{2n^2 \mu}$$

where T_i and T_j are waiting times for individual road users, n is the total number of road users, and μ is the mean waiting time. This metric provides a more comprehensive measure of fairness by considering the entire distribution of waiting times, rather than just the extremes. We can calculate F for each user class to identify potential inequities across different road user groups, allowing for a detailed analysis of fairness in various traffic scenarios.

- **Ethical Decision Alignment**: Percentage of ESRM decisions that align with pre-defined ethical guidelines.
- **VRU Priority Score**: Measure of how often VRUs are given priority in conflict situations.

4.1.4 User Experience and System Performance Metrics

- **Compliance Rate**: Percentage of traffic participants obeying ESRM instructions.
- **Trust and User Satisfaction**: Qualitative measures from VR or augmented reality user studies, capturing perceived safety, convenience, and clarity of ESRM signals.
- **Network Reliability**: Measured by round-trip times for V2X messages, packet drop counts, and ESRM fallback activations.
- **Computational Efficiency**: CPU/GPU utilization and decision-making latency under various traffic loads.

5 Conclusion and Future Work

This paper presents a comprehensive framework for Ethical Social Robot Moderators (ESRMs), designed to navigate the complex landscape of urban traffic management. The key contributions include a novel multi-objective utility function that dynamically balances safety, fairness, and efficiency in real-time traffic scenarios, and an ethical decision-making module that incorporates advancements in AI ethics and risk assessment. A robust simulation environment using CARLA has been proposed that aims to capture real-world unpredictability, along with comprehensive metrics, allowing for rigorous testing of ESRM performance in various scenarios.

The ESRM framework developed offers a promising approach to achieving safer, more efficient, and ethically sound urban traffic management. This work lays the groundwork for integrating ethical considerations into the core of smart city infrastructure, potentially paving the way for more human-centered urban mobility solutions. Importantly, the framework's novelty lies in its multi-domain integration, enabling ethical reasoning not just within individual vehicles but also at the urban infrastructure level through social robot moderators. This positions ESRMs as a pivotal new actor in the landscape of cooperative traffic systems, with the potential to standardize ethical behavior across diverse road users.

However, several limitations remain. Most notably, the reliance on human tele-operation introduces inherent latency, which, despite mitigation strategies, constrains the system's effectiveness in fast-developing critical scenarios. Additionally, as the current work is primarily conceptual, empirical validation through simulations and real-world testing is needed to confirm feasibility and performance. Finally, adapting ethical parameters to diverse cultural and legal contexts presents ongoing challenges that warrant further interdisciplinary exploration.

A possible path forward for future work could focus on implementing and evaluating the proposed simulation scenarios to validate ESRM performance in diverse urban environments. The aim could be to refine ethical decision-making algorithms based on these simulation results and explore how they might adapt to different cultural contexts and emerging scenarios. Additionally, investigating the potential long-term impacts of ESRM deployment on urban mobility patterns and social equity could be considered.

By addressing both technical and societal dimensions in the proposed framework, this work contributes to the ongoing dialogue about ethical AI in urban environments. As autonomous technologies continue to evolve, the ESRM concept provides a foundation for further research into integrating ethical considerations into traffic management systems.

References

[1] D. J. Fagnant and K. Kockelman, "Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations," *Transportation Research Part A: Policy and Practice*, vol. 77, pp. 167–181, 2015.

[2] A. P. Morris, N. Haworth, A. Filtness, D.-P. A. Nguatem, L. Brown, A. Rakotonirainy, and S. Glaser, "Autonomous vehicles and vulnerable road-users—important considerations and requirements based on crash data from two countries," *Behavioral Sciences*, vol. 11, no. 7, p. 101, 2021.

[3] M. Geisslinger, F. Poszler, J. Betz, C. Lütge, and M. Lienkamp, "Autonomous driving ethics: From trolley problem to ethics of risk," *Philosophy Technology*, vol. 35, pp. 1–23, 2022.

[4] L. Tanqueray and S. Larsson, "What norms are social robots reflecting? a socio-legal exploration on hri developers," in *Social Robots in Social Institutions*. IOS Press, 2023, pp. 305–314.

[5] G. Sidorenko, J. Thunberg, and A. Vinel, "Cooperation for ethical autonomous driving," in *2024 20th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*. IEEE, 2024, pp. 391–395.

[6] E. K. Duarte, M. Shiomi, A. Vinel, and M. Cooney, "Trust in robot self-defense: People would prefer a competent, tele-operated robot that tries to help," *IEEE Robotics and Automation Letters*, vol. 8, no. 5, pp. 2835–2842, 2023.

[7] E. K. Duarte, M. Erneberg, E. P. de Freitas, B. Bellalta, and A. Vinel, "Safesmart: A vanet-lte-based solution for faster and safer response in critical situations," in *2023 IEEE Conference on Standards for Communications and Networking (CSCN)*. IEEE, 2023, pp. 47–53.

[8] E. Awad, S. Dsouza, A. Shariff, I. Rahwan, and J.-F. Bonnefon, "Universals and variations in moral decisions made in 42 countries by 70,000 participants," *Proceedings of the National Academy of Sciences*, vol. 117, no. 5, pp. 2332–2337, 2020.

[9] H. S. M. Lim and A. Taeihagh, "Algorithmic decision-making in avs: Understanding ethical and technical concerns for smart cities," *Sustainability*, vol. 11, no. 20, p. 5791, 2019.

[10] J. Borenstein, J. R. Herkert, and K. W. Miller, "Self-driving cars and engineering ethics: The need for a system level analysis," *Science and engineering ethics*, vol. 25, pp. 383–398, 2019.

[11] N. J. Goodall, "Machine ethics and automated vehicles," *Road vehicle automation*, pp. 93–102, 2014.

[12] D. González, J. Pérez, V. Milanés, and F. Nashashibi, "A review of motion planning techniques for automated vehicles," *IEEE Transactions on intelligent transportation systems*, vol. 17, no. 4, pp. 1135–1145, 2015.

[13] J. Christian Gerdes, S. M. Thornton, and J. Millar, "Designing automated vehicles around human values," in *Road Vehicle Automation 6 6*. Springer, 2019, pp. 39–48.

[14] A. Martins Martinho Bessa, N. Herber, M. Kroesen, and C. Chorus, "Ethical issues in focus by the autonomous vehicles industry," *Transport Reviews: a transnational, transdisciplinary journal*, vol. 41, no. 5, 2021.

[15] E. Dogan, C. Barbier, and E. Peyrard, "Public perception of ethical issues concerning automated mobility: a focus group study among three road user categories," in *Proceedings of the 32nd European Conference on Cognitive Ergonomics*, 2021, pp. 1–6.

[16] M. Schrapel, M. Bied, B. Bruno, and A. Vinel, "Experiencing social robots for traffic guidance using virtual reality videos," in *Mensch und Computer 2024-Workshopband*. Gesellschaft für Informatik eV, 2024, pp. 10–18 420.

[17] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, "Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems, version 2," IEEE, Tech. Rep., 2017. [Online]. Available: http://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf

[18] E. K. Duarte, L. A. L. Da Costa, M. Erneberg, E. P. De Freitas, B. Bellalta, and A. Vinel, "Safesmart: A vanet system for faster responses and increased safety in time-critical scenarios," *IEEE Access*, vol. 9, pp. 151 590–151 606, 2021.

[19] D. G. Black, D. Andjelic, and S. E. Salcudean, "Evaluation of communication and human response latency for (human) teleoperation," *IEEE Transactions on Medical Robotics and Bionics*, vol. 6, no. 1, pp. 53–63, 2024.

[20] C. Luetge, "The german ethics code for automated and connected driving," *Philosophy & Technology*, vol. 30, pp. 547–558, 2017.

[21] K. Evans, N. de Moura, S. Chauvier, R. Chatila, and E. Dogan, "Ethical decision making in autonomous vehicles: The av ethics project," *Science and engineering ethics*, vol. 26, pp. 3285–3312, 2020.

[22] M. Hamidaoui, M. Z. Talhaoui, M. Li, M. A. Midoun, S. Haouassi, D. E. Mekkaoui, A. Smaili, A. Cherraf, and F. Z. Benyoub, "Survey of autonomous vehicles' collision avoidance algorithms," *Sensors (Basel, Switzerland)*, vol. 25, no. 2, p. 395, 2025.

[23] German Federal Ministry of Transport and Digital Infrastructure, "Ethics commission: Automated and connected driving," German Federal Ministry of Transport and Digital Infrastructure, Tech. Rep., 2017. [Online]. Available: https://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission.pdf

[24] G. Contissa, F. Lagioia, and G. Sartor, "The ethical knob: Ethically-customisable automated vehicles and the law," in *Proceedings of the 1st Workshop on AI and Ethics*. CEUR-WS, 2017, pp. 27–35.

[25] A. Etzioni and O. Etzioni, "Incorporating ethics into artificial intelligence," *The Journal of Ethics*, vol. 21, no. 4, pp. 403–418, 2017.

[26] European Parliament, "The ethics of artificial intelligence: Issues and initiatives," European Parliament, Policy Document, 2020. [Online]. Available: https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS_STU(2020)634452_EN.pdf

[27] B. Shneiderman, "Human-centered artificial intelligence: Reliable, safe trustworthy," *International Journal of Human–Computer Interaction*, vol. 36, no. 6, pp. 495–504, 2020.

[28] ETSI, *Intelligent Transport Systems (ITS); Vehicular Communications; Basic Set of Applications; Part 2: Specification of Cooperative Awareness Basic Service*, Std. EN 302 637-2, 2015. [Online]. Available: https://www.etsi.org/deliver/etsi_en/302600_302699/30263702/01.03.02_60/en_30263702v010302p.pdf

[29] T. Arnold, D. Kasenberg, and M. Scheutz, "Value alignment or misalignment–what will keep systems accountable?" *AI, Ethics, and Society Papers from the 2017 AAAI Workshop*, 2017.

[30] A. Dosovitskiy, G. Ros, and V. Koltun, "CARLA: An open urban driving simulator," *CoRR*, vol. abs/1711.03938, 2017. [Online]. Available: https://arxiv.org/abs/1711.03938