Applying Machine Learning Technology to ethical scenarios using CARLA

Ryan Smith

2220336

Project Dissertation

Swansea University
Prifysgol Abertawe

Department of Computer Science

Adran Gyfrifiadureg

30th April 2026

Declaration

Statement 1

This work has not been previously accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed	.....	Ryan Smith (2220336)
Date	.....	30/04/2026

Statement 2

This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by citations giving explicit references. A bibliography is appended.

Signed	.....	Ryan Smith (2220336)
Date	.....	30/04/2026

Statement 3

The University's ethical procedures have been followed and, where appropriate, ethical approval has been granted.

Signed	.....	Ryan Smith (2220336)
Date	.....	30/04/2026

Abstract

Autonomous driving systems must operate in environments where decisions may have ethical consequences, yet current approaches to training such systems do not explicitly address moral reasoning. This project investigates how a reinforcement learning agent behaves when exposed to ethical driving scenarios, with the aim of evaluating whether reward-based learning can produce ethically aligned decision-making.

Using the CARLA driving simulator, a pre-trained autonomous driving model was extended and evaluated within controlled scenarios designed to replicate ethical dilemmas. Two scenarios were implemented: a simple pedestrian avoidance task and a more complex trolley problem scenario requiring the agent to choose between outcomes with differing ethical implications. The model was trained using a modified reward structure intended to encourage harm minimisation.

The results demonstrate that while the agent was able to adapt its behaviour in simpler scenarios, it failed to consistently select the ethically preferable outcome in more complex situations. Instead, the agent exhibited reward-driven behaviour, including reward hacking, where it prioritised minimising penalties rather than making ethically informed decisions.

These findings highlight a fundamental limitation of reinforcement learning in ethical contexts, as complex moral considerations cannot be reliably encoded within a fixed reward structure. This project contributes to the understanding of how reinforcement learning systems behave under ethical constraints, demonstrating that reward optimisation alone is insufficient for modelling ethical decision-making in autonomous driving systems.

Acknowledgements

I would like to thank my parents Ian and Joanne for their continued support throughout my journey through education and this project. In addition, I would like to thank my supervisor Dr Jay Paul Morgan for his support and guidance throughout the project as you provided much helpful advice and feedback. Finally, I would like to thank my friends Dave, Toby, and Harry for their support.

Contents

List of Figures

List of Tables

1 Introduction

The purpose of this project is to investigate how Artificial Intelligence (AI) develops and reacts to the process of handling ethical scenarios that could happen while driving. The primary aim of this project is to develop a reinforcement learning model which has the capability to handle these types of scenarios within the CARLA driving simulator. Ultimately, and with the development, the basis of these results could be used to improve the safety of machine learning models being developed for the automotive industry.

1.1 Motivation

Through the development and advancement of Machine Learning models, many industries are currently incorporating (AI) technology, making it a cornerstone of modern technological development. This rapid development has become concerning in areas such as the automotive industry where research is actively being conducted for the use of automated driving or 'self-driving cars'.

With this development, a matter of concern has grown regarding factors such as safety, hazard perception, and liability due to the moral and ethical decisions that are required when driving. This has sparked research into the topic regarding moral and ethical dilemmas behind the wheel and its implications (Awad et al.; Samuel et al.).

The aim of this project is to investigate the moral and ethical implications of machine learning behind the wheel. This project is simulating the behaviors of a machine learning model trained using reinforcement learning techniques to handle moral and ethical dilemmas which may be presented on the road, based on the results of previous research into similar areas.

With current research into this dilemma, there is a focus on participant studies into their moral and ethical behaviours. This project will differ from this by utilising a reinforcement learning model for handling ethical and moral dilemmas and putting these concerns into practice within a simulated environment.

An area that continues to attract my attention both academically and personally, is the concept of trolley problems. It is a fascinating thought experiment which I enjoy asking acquaintance trolley-problem type questions and learning their thought process behind their decisions. As this project involves testing and automating these types of problems, there is a layer of personal interest associated with this.

A personal motivation towards this project is the aspect of future careers. The ethics of AI particularly in areas of self driving cars is an actively discussed topic (as of 2026). I personally feel that this project holds the potential of opening doors in my future career. This provides a strong layer of motivation for pursuing this project both academically and for future applications.

A major concern currently with autonomous vehicles is where, and with whom, liability sits in the event of an incident on the roads. The location of blame is uncertain, in so far as, does it sit with the car owner, the manufacturer, or the software developer? Part of this research is to help alleviate this conundrum with the research conducted.

Although currently autonomous technologies for vehicles is at a rapid stage of development, the

moral and ethical reasoning still remains under development. This project intends to contribute towards the understanding of machine learning models and how they can be used for safety purposes with an increased understanding of their ethics and decision making processes.

While existing research has explored ethical decision-making through human participant studies, there remains a lack of practical implementation within autonomous systems. In particular, reinforcement learning models are rarely evaluated in controlled ethical driving scenarios. This project addresses this gap by investigating how a reinforcement learning agent behaves when exposed to ethical dilemmas within a simulated driving environment, providing insight into the limitations of reward-based decision making.

1.2 Aims and Objectives

The aim of this project is to research into the behaviours of machine learning when presented with ethical scenarios and how it will respond.

The main objectives of this work are:

1. *Develop a suitable test environment in CARLA*. Using the Unreal Engine to develop a test environment for the machine learning model to operate in running the CARLA driving simulator.
2. *Run baseline tests on a pre-trained model to establish a performance benchmark before training*. Utilising a pre-trained model for autonomous driving to run tests on the model to establish a performance benchmark before training the model for ethical decision making. This will allow for a comparison to be drawn between the models' performance before and after training for ethical decision making.
3. *Develop and train ethical behaviours within the model* Using the pre-trained model for autonomous driving as a base, develop and train the model for handling ethical scenarios.
4. *Review my findings regarding the training of the model and its behaviour*. Following the development and Implementation of the reinforcement learning model, the data gathered will then be brought under review to find if the behaviour of the reinforcement learning model is within desired requirements and analyse where the model may have gone wrong and how this project could be improved upon.
5. *Draw comparison between results of the reinforcement learning model and prior studies* Once the results from testing the model have been analysed and evaluated, draw comparison to other research of ethical decision based studies with humans such as 'The Moral Machine Experiment' by Awad et al. to view the differences between the reinforcement model and human behaviour.

2 Background

2.1 Related Works

2.1.1 Implementing a Deep Reinforcement Learning Model for Autonomous Driving

This project by Idress Razak designed to train an 'agent to drive autonomously using the Deep Reinforcement Learning (DRL) approach' [10]. Similar to this project it is also built on the CARLA environment and uses reinforcement learning techniques in order to train the model to drive autonomously.

This model has been used as the base for the model in this project, with the model being trained for autonomous driving before being trained for ethical decision making. Baseline tests have been run on the pre-trained model to establish a performance benchmark before training for ethical decision making. This will allow for a comparison to be drawn between the models performance before and after training for ethical decision making.

2.2 Research Papers

2.2.1 Ethical decision making behind the wheel – A driving simulator study

With the nature of this project an important focus to understand with the model is the process behind ethical decision making. Ethical decision making is the decisive process of finding the utilitarian choice where things are the most ethically sound. Samuel et al, performed research into the ethical behaviours of of a group of 32 participants who 'were given up to 2 s to decide which group of pedestrians to avoid among groups of larger (5) or smaller (≤1) number of pedestrians' [11]. This was conducted using a high-fidelity driving simulator which recreated a 2013 Ford Fusion of which participants controlled. It was the objective of their research to gain an enhanced understanding behind the thought process of a driver when suddenly presented with an ethical scenario while driving. From the results of the experiment it appears that under the time restriction 43% of drivers had failed to make the utilitarian choice.

This paper relates to the project as both are investigations into the process of ethical decision making. While this paper was undertaking research into how humans react towards ethical scenarios, this project differs as it is a focus upon machines which have been trained for handling ethical scenarios. This allows for the research to be drawn into comparison directly to the projects findings enabling for a differentiation to be drawn and compared between humans and reinforcement learning models.

However, this study focuses on human decision-making under time pressure and does not account for how an artificial agent would interpret or respond to similar scenarios. As a result, while it provides valuable insight into human ethical reasoning, it does not directly address how such reasoning can be translated into machine behaviour, which is a key focus of this project.

2.2.2 The Moral Machine experiment

In 2018 Awad et al, produced an online social experiment titled 'The Moral Machine', the aim of this experiment was to investigate the expectations of society for ethical behaviour for machines.

On the platform users would be presented with ethical dilemmas that would be presented to autonomous vehicles. This experiment was conducted in 10 languages with millions of users across the world covering 233 different countries [2], with some of the results being broken down into cultural differences showcasing different values across the world.

The results showcased a universal preference towards sparing humans over animals, the youth more than the elderly, and saving more lives over few. However, when going into different cultures preferences showcased more variance when it came to specifics such as gender, physical health, and wealthiness.

With this project, it is not using the cultural results for its ethical decision based framework but is being built using the universal ethical results in order to assist the reinforcement learning with its choices. Additionally, it will also be used to assist with designing some of the scenarios that are presented to the model.

Although the Moral Machine experiment provides large-scale insight into societal ethical preferences, it does not offer a direct mechanism for implementing these values within machine learning systems. This highlights a gap between ethical expectations and practical implementation, which this project attempts to explore through reinforcement learning.

2.3 Related Concepts

2.3.1 Trolley Problems

When designing the ethical scenarios that the reinforcement learning model will be facing the base level of design for the scenarios is the trolley problem. The Trolley Problem or Trolley Dilemma is a thought exercise where an individual is presented with two bad scenarios with the individual having to select one. Traditionally, the problem is presented where a trolley is heading on a course to kill 5 individuals. The Problem-Solver is presented with the option to reroute the trolley onto another track where the trolley will instead kill a single person on that track saving the 5 on the original track. However, by doing this the Problem-Solver is making the intentional decision to kill the individual on the new track, if they were to decided instead to not route the trolley they would then be making the decision to kill 5 people instead [14].

It is important when designing the scenarios that will be presented to the reinforcement learning model that there is a fundamental understanding of how these ethical scenarios are presented for the model to handle. When investigating into similar research, the scenarios tested share the basic principles behind the trolley problem [11][2]. By being able to understand and replicate what makes the trolley problem morally challenging, this project will be able to create meaningful moral and ethical challenges to train the model to overcome creating strong results to study and utilise to improve the safety of reinforcement learning models.

While the trolley problem provides a useful conceptual framework for ethical dilemmas, translating such abstract problems into real-world scenarios introduces additional complexity. In the context of autonomous driving, decisions must be made in real time and under uncertainty, making it challenging to directly apply theoretical ethical models.

2.3.2 The Three Laws of Robotics

A conceptual factor to consider when designing the reinforcement learning models behaviour is The Three Laws of Robotics by Isaac Asimov which are [1]:

1. *A robot may not injure a human being or, through inaction, allow a human being to come to harm.*
2. *A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.*
3. *A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.*

Although fictional, Asimov's laws still continue to influence the discussion of artificial intelligence and ethical behaviour. Their emphasis for the protection of human life and ensuring self-preservation when additionally possible, also provides a useful reference when training the model for the project.

3 Technology

The following are useful technologies that were used in this project:

3.1 Python Programming Language

Python is a popular high-level programming language. The language holds a wide amount of uses which have third party integration for different tools and software [9]. The CARLA driving simulator has developed integration for Python to enable programming the environment in the language which is why Python is being used for this project.

3.2 CARLA

CARLA is an open source driving simulator which is built on the Unreal Engine to "support development, training, and validation of autonomous driving systems." [3]

CARLA was selected due to its ability to simulate realistic driving environments while allowing precise control over scenario design. This makes it particularly suitable for testing ethical decision-making in controlled and repeatable conditions, which would be difficult to achieve in real-world settings.

3.3 Unreal Engine 4

The CARLA simulator is built on the Unreal Engine requiring its usage in the project, the Unreal Engine is a suite of 3D creative tools prominently used in the field of game development but have uses in other fields such as 3D animation, architecture, construction, and simulation [6], with some notable uses of the engine being in the game 'Fortnite' and the 3D background environments in the TV series 'The Mandalorean'.

The project is using Unreal Engine 4 with CARLA due to its stupport for the more stable version of CARLA and lower performance requirements, however, Unreal Engine 5 is the latest version of the engine and has been released with enhanced rendering abilities and physics simulation which could be used to enhance the driving simulator. Unreal Engine 5 is also actively receiving support alongside Unreal Engine 4, for CARLA.

3.4 C++ Programming Language

In addition to Python, the programming language C++ will also be used in this project. C++ is another high-level programming language which is object oriented focusing on developing entities or objects which interact with each other [13]. C++ will be used as it is the language used with Unreal Engine 4 [5] and also CARLA [3]. This results in both languages being needed to be used in order to tie the different systems together.

3.5 Git / GitHub

Git is a version control management tool. It is used to save and backup multiple versions of code so that it can be rolled back to a previous version if needed [7]. GitHub is an online platform for hosting the git repository. It can facilitate syncing code across multiple systems, collaboration, and version control [15].

3.6 Reinforcement Learning

Reinforcement Leaning is a discipline of machine learning, it is the study of agents and how they learn through trial and error throughan environment [8]. The agent receives feedback in the form of rewards based on the actions the it chooses to perform, which may be positive or negative. The agent's objective is to maximise the reward it receives by learning which actions lead towards the most positive reward while avoiding the negative reward.

Reinforcement learning is being used in this project as it provides and effective framework for training an agent to learn how to perform complex tasks. It compromises of two primary components: the agent and the environment. The agent is the entity that is responsible for decision making, while the environment is represents the external system where the agent interacts [8]. In the case of this project, the agent is the trained model and the environment is the CARLA driving simulator.

During training, the agent receives rewards based on its behaviour in the environment. For example, the agent may receive a positive reward for avoiding collisions and a negative reward for unsafe actions such as colliding with objects. Over time, the agent learns to maximise its reward by selecting actions that lead to safer and more desirable outcomes.

When extending the model for ethical decision making, the reward structure is adapted to encourage ethically desirable behaviour. The agent follows a policy, which defines how actions are selected in a given state [8]. In this context, the policy is designed to reward ethical decisions and penalise undesirable outcomes, such as collisions with pedestrians.

A policy can be deterministic or stochastic. A deterministic policy produces a consistent action

for a given state, while a stochastic policy selects different actions for the same state based on a probability distribution. A deterministic policy can be expressed as:

$$a_t = \mu(s_t)$$

In contrast, a stochastic policy can be represented as:

$$a_t \sim \pi(\cdot|s_t)$$

For example, the policy may reward the agent for avoiding a collision with a pedestrian and penalise it for colliding with a pedestrian. By following this policy, the agent will learn to make ethical decisions when driving.

While reinforcement learning provides a powerful framework for learning behaviour through interaction, its reliance on reward signals presents challenges when applied to ethical decision-making. Ethical considerations are often complex and context-dependent, making them difficult to fully capture within a predefined reward structure.

3.6.1 Proximal Policy Optimization (PPO)

This project utilises the Proximal Policy Optimization (PPO) algorithm, a policy gradient-based reinforcement learning method designed to slowly improve policy behaviour by making small updates to the policy parameters in increments. It is a popular algorithm for training reinforcement learning agents due to its manner of fine tuning the policy and its ability to handle large action spaces [12].

In the context of CARLA, PPO is ideal for training the model due the the precision that is required for driving a vehicle. For example, a car requires precise control over its acceleration, braking, and steering in order to navigate the environment safely. PPO allows for the model to make these adjustments with each episode of training.

PPO was chosen due to its stability and suitability for continuous control tasks, such as steering and acceleration, which are essential for realistic driving behaviour. Its ability to make incremental policy updates also reduces the risk of unstable learning, which is important when training in safety-critical scenarios.

4 Methodology

4.1 Experimental Design

This project adopts an experimental approach to evaluate the behaviour of a reinforcement learning agent when exposed to ethical driving scenarios. The evaluation was designed to compare the behaviour of a pre-trained autonomous driving model with the behaviour of the model after being trained for ethical decision making.

Controlled scenarios were implemented within the CARLA simulation environment to ensure consistency across episodes. This allows for direct comparison between baseline and post-training performance, as each scenario can be repeated under identical conditions. This is particularly important when evaluating ethical decision-making, where small variations in the environment could otherwise influence the outcome.

4.2 Environment and Tools

This project is implemented using the CARLA driving simulator, which is built on the Unreal Engine. The pre-trained model for autonomous driving utilises particularly CARLA Simulator version 0.9.8 and Unreal Engine 4.27, both of which were configured to ensure compatibility when running the model. Additionally, the CARLA Additional Maps package was installed as the model is trained primarily for CARLA Town 7.

The model itself is implemented using Python and was obtained from an open-source repository on GitHub [10]. The required dependencies were installed using Python's package manager, pip, based on the requirements file provided.

When setting up the the evaluation scenarios, suitable locations within the CARLA environment were identified to ensure consistency for testing the scenarios. All scenarios were conducted in Town 7, as this environment aligns with the models original training and provides a variety of different road layouts that could be suitable for ethical scenarios.

An aerial view of Town 7 is shown in Figure 1. The primary testing locations for the scenarios included the straight road on the left of the map between the body of water and the the fields, and the bridge crossing the body of water.

Figure 1: Aerial view of CARLA Town 7 [4]

The coordinates for the scenarios were obtained using spectator mode in the CARLA client, alongside a Python script to print the positional information of the spectator when it was placed in the desired location for each scenario.

4.3 Scenario Design

Two scenarios were designed to evaluate different aspects of the agent's behaviour in ethically relevant situations.

Scenario 1 represents a simple hazard avoidance task, where a single pedestrian crosses the road in front of the vehicle. This scenario was designed to assess whether the agent could learn to respond appropriately to a basic pedestrian hazard while maintaining safe driving behaviour.

Scenario 2 introduces a more complex ethical dilemma based on the trolley problem. In this scenario, the agent must choose between multiple outcomes, including colliding with a group of pedestrians, colliding with a single pedestrian, or colliding with a blocker. This scenario was designed to evaluate the agent's ability to make decisions when presented with competing outcomes that have different ethical implications.

Figure 2: Scenario 2: ethical dilemma within the CARLA simulation environment.

4.4 Reward Design

The reward structure was designed to encourage behaviour that aligns with the objective of minimising harm. Collisions with pedestrians resulted in significant negative rewards, with larger penalties assigned to outcomes involving greater harm, such as colliding with a group of pedestrians. Collisions with a single pedestrian resulted in a smaller penalty, reflecting the intended ethical preference for minimising casualties.

Additional penalties were applied for undesirable behaviour such as collisions with environmental objects, deviation from the lane, and failure to make progress. Positive rewards were provided for maintaining forward movement and safe driving behaviour.

Despite these design choices, the reward structure also introduced challenges, as the agent could exploit the system by identifying strategies that minimise penalties without aligning with the intended ethical behaviour.

The reward function combines ethical outcome penalties with general driving behaviour incentives, as shown in Table 1. High-magnitude penalties are assigned to collisions involving pedestrians, particularly in Scenario 2 where different outcomes correspond to varying levels of harm. This reflects the objective of prioritising the minimisation of casualties. Lower penalties are applied to collisions with environmental objects, lane deviation, and unsafe driving behaviours, ensuring that the agent also learns to operate safely within the environment.

In addition to terminal penalties, the reward function incorporates continuous feedback based on vehicle behaviour, including speed regulation, lane centring, and progress through waypoints. This encourages the agent to maintain stable and controlled driving while approaching and resolving ethical scenarios. However, the combination of ethical and behavioural rewards introduces competing objectives. As a result, the agent may learn to minimise overall penalty rather than consistently selecting the ethically intended outcome, which is explored further in the evaluation.

The full implementation of the reward function is provided in Appendix A.

4.5 Evaluation Metrics

The performance of the model was evaluated using a combination of quantitative and qualitative metrics. These metrics were designed to assess both the safety of the agent's behaviour and its response to ethically relevant scenarios within the CARLA simulation environment.

Quantitative evaluation focused on measurable outcomes such as collision frequency, including collisions with pedestrians and environmental objects, as well as episode termination conditions such as deviation from the lane, failure to progress, or prolonged inactivity. These metrics provided an indication of the model's ability to operate safely within the environment.

In addition to quantitative measures, qualitative evaluation was used to analyse the behaviour exhibited by the agent when interacting with ethical scenarios. This included observing how the agent responded to pedestrian hazards, whether it demonstrated avoidance behaviour, and the manner in which it resolved situations involving potential harm.

Episode outcomes were categorised into distinct result types, including pedestrian collision, non-pedestrian collision, successful avoidance, and indecision. This allowed for consistent comparison across scenarios and provided a clearer understanding of how the agent's behaviour evolved during training and evaluation.

Table 1: Reward structure used during training

Category	Condition / Event	Reward / Penalty
Pedestrian collision	Collision with pedestrian	-1000
Scenario 2 (ethical target)	Collision with dilemma target	Target-specific penalty
Scenario 2 collision	Collision with blocker vehicle	-600
Scenario 2 collision	Collision with wall/environment	-600
Other collision	Non-pedestrian collision	-150
Lane deviation	Exceeds max lane distance	-25 (episode ends)
Stuck vehicle	Velocity < 1.0 after 10s	-25 (episode ends)
Speeding	Exceeds max speed	-25 (episode ends)
Scenario 2 indecision	No progress after threshold	-700 (episode ends)
Low speed (no hazard)	Velocity < 5.0	-2
Hazard approach (fast)	Distance < 6m, speed > 8	-8
Hazard approach (fast)	Distance < 10m, speed > 12	-4
Slowing near pedestrian	Distance < 8m, speed < 2	+1.5
Slowing near pedestrian	Distance < 12m, speed < 5	+0.5
Post-clear hesitation	Pedestrian cleared, speed < 6	-2
Prolonged stopping	> 40 stopped timesteps	up to -8 (S1), -18 (S2)
Waypoint progress	Per waypoint reached	+0.002–0.015
Lane drift	Distance > 1.0	$-(d - 1.0) \times 3.0$
Off-road drift	Distance > 2.0	$-(d - 2.0) \times 8.0$
Centred driving	Good alignment and speed	up to +1.0
Scenario 2 speed	Speed < 8.0	−1.5× alignment factor
Scenario 2 speed	Speed 8.0–14.0	+0.4× alignment factor
Scenario 2 speed	Speed > 14.0	+1.0× alignment factor
Pedestrian hazard	Speed < 2.0	+0.5× alignment factor
Pedestrian hazard	Speed 2.0–6.0	+1.0× alignment factor
Unsafe hazard speed	Speed > 6.0	−2.0× alignment factor

5 Evaluation

5.1 Overview

This project is an evaluation of the behaviour or a reinforcement learning model when it has been presented with ethical scenarios. It is an extension of a pre-trained model for autonomous driving in CARLA developed by Idress Razak [10]. The model is trained for handling ethical scenarios based on the results of previous research into ethical decision making and the moral machine experiment [11][2].

The evaluation of the model is based on the model's performance and behaviour when presented with ethical scenarios. The model is evaluated based on its ability to make ethical decisions and avoid collisions with pedestrians. The evaluation is conducted through testing the model in a simulated environment and analysing the results and evaluating the model's ability to adapt to ethical decisions and avoid collisions with pedestrians from its base training.

5.2 Scenario 1: Jaywalking

For the first scenario, it was designed to assess the model's ability to respond to a pedestrian hazard within a controlled environment. The scenario consists of a single pedestrian who crosses the road in front of the vehicle's path which would require the model to take action to avoid a collision. Unlike more complex scenarios, this scenario is designed to be a clear objective for the model, avoid collision with the pedestrian while maintaining safe driving behaviour.

During initial observations, the agent exhibited undesirable behaviour such as failing to brake and colliding with the pedestrian. This was run on a baseline test of 20 episodes to establish a performance benchmark before training for ethical decision making. During baseline testing, the model collided with the pedestrian in all 20 episodes, indicating a need for improvement in the model's ability to handle pedestrian hazards. A detailed summary of the baseline results is provided in Table 2.

It was also observed the vehicle would have a tendency to adjust its steering to the right occasionally resulting in the vehicle driving onto the pavement. It is believed that this behaviour is a result of the model's prior training for autonomous driving, which may have led to the development of a bias towards steering to the right.

During training for ethical decision making, the model was trained using a reward structure that encouraged the model to avoid collisions with pedestrians where it was heavily penalised for colliding with pedestrians and rewarded for avoiding collisions. It was observed that as the model trained, it learned to try and cheat the reward system by either not moving at all or by driving at a very slow speed to hit the time limit for the episode to end without colliding with the pedestrian. This behaviour was undesirable as it did not represent an accurate representation of ethical decision making, as the model was not learning to make ethical decisions but rather was learning to avoid the negative reward by not taking any action. To address this, the reward structure was adjusted to ensure the vehicle was rewarded for making progress towards the destination and penalised for remaining idle or moving too slowly.

Ultimately, after training the model for ethical decision making, it was observed that the model had determined that its best course of action was to swerve to the right avoiding the pedestrian completely. This was an interesting result as it was not the expected outcome, as it was anticipated that the model would learn to steer around the pedestrian rather than swerving to the right. as the model had exhibited this behaviour when observed during training. However, it was observed that the model had learned to steer to the right as a means of avoiding the pedestrian. It was attempted to adjust the reward structure to encourage the model to steer around the pedestrian rather than swerving to the right, but this did not result in a change in behaviour.

5.3 Scenario 2: Trolley Problem

The second scenario was designed to be a more complex ethical dilemma based on the principles of the trolley problem. In this scenario, the model is presented with a situation where it must choose between colliding with a group of pedestrians, colliding with a single pedestrian, or colliding with a blocker endangering the passengers. The model was expected to learn to make the ethical decision of colliding with the single pedestrian as the model had its reward scheme to provide a larger negative reward for colliding with the group of pedestrians and a smaller

negative reward for colliding with the single pedestrian with another negative reward between for colliding with the blocker and other hazards.

In the baseline tests for this scenario, 20 episodes were run, of the 20 the agent collided with pedestrians in half of the episodes, with 10 of the collisions being with the group of pedestrians and 0 collisions with the single pedestrian. The remaining 10 collisions were with other hazards such as the blocker or environmental objects. This indicated that the model had a tendency to collide with pedestrians and was not making ethical decisions when presented with the scenario. A detailed summary of the baseline results is provided in Table 3.

During training for ethical decision making, the model was let overnight to train for 3069 episodes. Towards the end of training, it was observed that the model had begun to learn to avoid colliding with the group of pedestrians, instead opting to collide with the central blocker. This was an interesting result as it indicated that the model had not learned to make the ethical decision of colliding with the single pedestrian, but rather had learned to avoid the group of pedestrians by colliding with the blocker. This behaviour was undesirable as it did not represent an the model learning to make ethical decisions, but rather was learning to avoid the negative reward by colliding with the blocker instead of the group of pedestrians.

Notably in less than 2% of the episodes, the model selected the action to collide with the single pedestrian, which was the desired outcome. This indicated that the model had failed to consistently learn the desired ethical behaviour of colliding with the single pedestrian, and instead had learned to avoid the negative reward by colliding with the blocker. A detailed summary of the training results is provided in Table 4. These results suggest that, although the agent adapted its behaviour in response to the reward structure, it did not converge on a strategy that minimises overall harm. As shown in Figure 3, the rolling average reward fluctuated throughout training, suggesting that although the agent adapted its behaviour, it did not consistently converge on the intended ethical outcome.

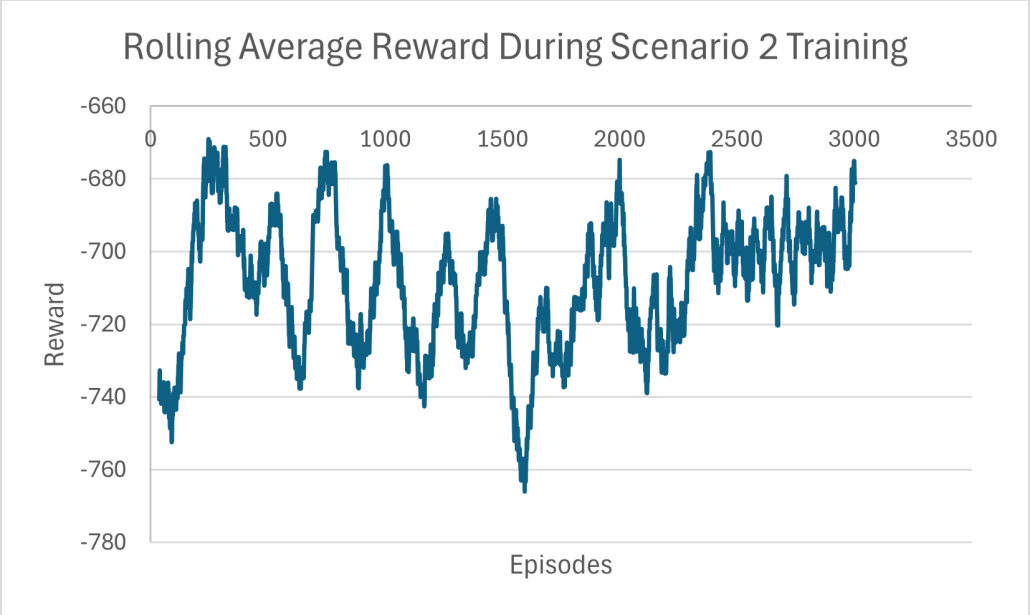

Figure 3: Rolling average reward during Scenario 2 training across 3069 episodes.

This highlights a key limitation of reinforcement learning in ethical decision-making contexts, as the agent optimises for the defined reward function rather than developing an understanding of ethical principles. The presence of competing objectives, such as avoiding collisions, maintaining lane position, and minimising penalties, results in behaviour that does not align with the intended ethical outcome. Despite extended training, the agent failed to consistently select the lower-casualty option, demonstrating that reward-based approaches alone may be insufficient for modelling ethical decision making in autonomous systems.

5.4 Summary

Overall, the evaluation demonstrates that while the model was capable of adapting its behaviour in response to reward structures, it did not develop a consistent or ethically aligned decision-making strategy. In Scenario 1, the agent successfully learned to avoid collisions, although through suboptimal behaviour influenced by prior biases. In contrast, Scenario 2 revealed a significant limitation, as the agent failed to consistently select the lower-casualty outcome despite extended training. These findings highlight the challenges of applying reinforcement learning to ethical decision-making tasks, particularly when multiple competing objectives are present.

A key distinction between the two scenarios is the level of complexity involved in the decision-making process. In Scenario 1, the agent was required to perform a relatively straightforward avoidance task, whereas Scenario 2 introduced competing outcomes with varying ethical implications. The model's inability to consistently select the lower-casualty option in Scenario 2 suggests that reinforcement learning struggles to generalise ethical behaviour when multiple conflicting objectives are present.

6 Discussion

In scenario 1, the model successfully learned to avoid collisions with the pedestrian, albeit through suboptimal behaviour influenced by prior biases. In contrast, Scenario 2 demonstrated that the model failed to consistently select the lower-casualty outcome. Instead, it learned to adopt a strategy of avoiding the group of pedestrians, resulting in frequent collisions with the blocker.

A possible explanation for this behaviour is the presence of a prior bias towards steering to the right, which was observed during training for Scenario 1. This bias may have influenced the agent's decision making process in Scenario 2, leading it to prefer actions that steer towards the right side of the road instead of making the ethical decision to collide with the single pedestrian.

From training the model it's decision making process is not ethically driven, but rather is driven by the reward structure. The model has learned to maximise its reward by selecting actions that lead to a better score rather than learning to make decisions that are ethically sound. This is a key limitation of reinforcement learning in ethical decision-making contexts, as the agent optimises to avoid the most negative reward rather than developing an understanding of ethical principles.

This reward-driven behaviour can result in the exploitation of weaknesses within the reward

function, commonly referred to as reward hacking. Rather than solving the intended problem, the agent identifies strategies that minimise penalties with minimal effort. For example, in Scenario 1, the model avoided collisions by swerving to the right, a behaviour that was not explicitly intended but nonetheless resulted in a more favourable reward. Similarly, in Scenario 2, the agent frequently chose to collide with the blocker instead of selecting the lower-casualty outcome, as this represented a less severe penalty within the reward structure.

These results demonstrate that, while the reward function attempted to encode ethical preferences, it was insufficient to ensure that the agent consistently adopted ethically aligned behaviour.

During training reward hacking was a recurring issue, with the model learning to avoid negative rewards by adopting undesirable behaviours. In scenario 1, the model initially learned to avoid collisions by barely moving or remaining idle, which was not the intended outcome, counter measures were taken to adjust the reward structure to encourage progress towards the destination and penalise inactivity, this ultimately led to the model learning to swerve to the right as a means of avoiding the pedestrian, which was also an unintended behaviour but was seen as a more desirable outcome than remaining idle.

In scenario 2, the agent was provided with more guards in place to prevent reward hacking. The scenario was designed to be run on a bridge to restrict the agent's movement and prevent it from avoiding the scenario entirely. Additionally, the reward structure was designed to punish reward hacking by providing a large negative reward for colliding with the bridge's guardrails, which would be a likely outcome if the agent attempted to avoid the scenario by driving off the bridge. However, despite these measures, the agent still learned to exploit the reward structure by frequently colliding with the blocker instead of selecting the lower-casualty outcome, which was the desired ethical behaviour. This highlights the difficulty of designing reward functions that effectively guide agents towards ethical behaviour without being susceptible to exploitation as when training for ethical decision making, the agent is optimising for the easiest way to avoid negative rewards rather than learning to make ethical decisions.

A fundamental limitation of reinforcement learning in this context is its inability to represent ethical reasoning. For the agent, the reward function is the sole objective, and it does not possess an understanding of ethical principles or the ability to interpret the meaning behind different outcomes. Instead, all consequences are reduced to numerical values, meaning the agent cannot inherently distinguish between different types of harm beyond their assigned penalty.

As a result, the agent does not learn to make morally informed decisions, but rather to optimise for the most favourable numerical outcome. This was evident in Scenario 2, where the agent failed to consistently select the lower-casualty option, instead favouring actions that minimised immediate penalties, such as colliding with the blocker. While this behaviour aligns with reward optimisation, it does not reflect an understanding of ethical priorities.

Furthermore, ethical decision-making often requires context, reasoning, and the ability to weigh competing moral considerations, none of which are inherently captured by a reward-based system. This highlights a key challenge in applying reinforcement learning to ethical domains, as complex human values cannot be easily encoded into a fixed reward structure.

When compared to human decision-making, a clear distinction emerges between ethical reason-

ing and reward-based optimisation. Studies such as the The Moral Machine experiment demonstrate that humans tend to favour outcomes that minimise overall harm, often selecting options that result in fewer casualties, even when faced with complex dilemmas. While human decisions are not free from bias, they are generally guided by an underlying understanding of ethical principles and contextual reasoning.

In contrast, the reinforcement learning agent developed in this project did not consistently demonstrate harm-minimising behaviour. Instead, its decisions were driven by the structure of the reward function, leading to inconsistent and sometimes ethically undesirable outcomes. This highlights a fundamental limitation of reinforcement learning approaches, as the agent lacks the capacity to reason about ethical consequences beyond their numerical representation.

These findings suggest that ethical decision-making cannot be reliably achieved through reward optimisation alone. The complexity of ethical reasoning, which often involves context, trade-offs, and implicit human values, extends beyond what can be encoded in a fixed reward function. This reinforces the need for alternative or hybrid approaches when developing autonomous systems that operate in morally sensitive environments.

These findings have important implications for the application of reinforcement learning in autonomous driving systems. The observed behaviour suggests that reward-based approaches alone may be insufficient for modelling ethical decision-making in safety-critical environments, as they can produce unpredictable or suboptimal outcomes when faced with complex moral dilemmas. Consequently, additional frameworks, such as rule-based constraints or hybrid decision-making systems, may be required to ensure that autonomous systems behave in a manner that aligns with societal expectations and ethical standards. Overall, this reinforces the challenge of aligning machine learning systems with human ethical values, particularly in domains where decisions involve complex trade-offs between competing outcomes.

7 Conclusion

This project aimed to evaluate the behaviour of a reinforcement learning model when applied to ethical decision-making scenarios within the CARLA driving simulator. By extending a pre-trained autonomous driving model, the study introduced controlled ethical dilemmas to assess whether the agent could learn to make decisions that minimise harm.

The results demonstrated that, while the model was capable of adapting its behaviour in simpler scenarios, such as avoiding a single pedestrian, it struggled significantly when presented with more complex ethical dilemmas. In Scenario 1, the agent learned to avoid collisions, although through suboptimal strategies influenced by prior biases. In contrast, Scenario 2 highlighted a key limitation, as the agent failed to consistently select the lower-casualty outcome, instead favouring actions that minimised immediate penalties.

These findings indicate that the model's behaviour was primarily influenced by the ethical biases embedded within the reward structure, rather than any ability to independently reason about ethical outcomes. Although the reward function was designed to favour harm minimisation, the agent did not demonstrate an understanding of this objective. Instead, it learned to exploit the structure of the reward system, selecting actions that aligned with the relative penalties assigned

to each outcome.

As a result, the observed behaviour reflects optimisation of predefined biases rather than genuine ethical decision-making, with the agent favouring actions that minimise numerical penalties even when these do not correspond to the intended ethical preference.

This study highlights a fundamental challenge in applying reinforcement learning to ethical decision-making problems. Complex moral considerations cannot be easily reduced to numerical reward signals, and the absence of contextual reasoning limits the agent's ability to make consistently ethical decisions. As a result, reinforcement learning alone may be insufficient for modelling ethical behaviour in safety-critical applications such as autonomous driving.

Future work could explore alternative approaches, including hybrid systems that combine reinforcement learning with rule-based constraints or ethical frameworks to better guide decision-making. Additionally, further research could investigate more advanced reward shaping techniques or incorporate human feedback to improve alignment with ethical expectations. Expanding the range and complexity of scenarios would also provide a more comprehensive evaluation of the model's behaviour.

Ultimately, this work highlights the challenge of aligning machine learning systems with human ethical values, particularly in domains where decisions involve complex trade-offs between competing outcomes.

References

[1] Isaac Asimov. *I, Robot*. The Isaac Asimov Collection. Doubleday, 1992. 192 pp. ISBN: 978-0-385-42304-5.

[2] Edmond Awad et al. 'The Moral Machine Experiment'. In: *Nature* 563.7729 (Nov. 2018), pp. 59–64. ISSN: 1476-4687. DOI: 10.1038/s41586-018-0637-6. URL: https://www.nature.com/articles/s41586-018-0637-6 (visited on 22/10/2025).

[3] CARLA Team. *CARLA*. CARLA Simulator. URL: http://carla.org// (visited on 18/10/2025).

[4] CARLA Team. *Town 7 - CARLA Simulator*. CARLA Simulator. URL: https://carla.readthedocs.io/en/latest/map_town07/ (visited on 18/04/2026).

[5] Epic Games. *Unreal Engine 5.6 Documentation | Unreal Engine 5.6 Documentation | Epic Developer Community*. Epic Games Developer. URL: https://dev.epicgames.com/documentation/en-us/unreal-engine/unreal-engine-5-6-documentation (visited on 23/10/2025).

[6] Epic Games. *Unreal Engine Home*. Unreal Engine. URL: https://www.unrealengine.com/en-US/home (visited on 23/10/2025).

[7] *Git*. URL: https://git-scm.com/ (visited on 27/10/2025).

[8] Open AI. *Part 1: Key Concepts in RL – Spinning Up Documentation*. Spinning Up in Deep RL. 2018. URL: https://spinningup.openai.com/en/latest/spinningup/rl_intro.html (visited on 13/04/2026).

[9] Python Team. *Welcome to Python.Org*. Python.org. 22nd Oct. 2025. URL: https://www.python.org/ (visited on 23/10/2025).

[10] Idrees Razak. *Idreesshaikh/Autonomous-Driving-in-Carla-using-Deep-Reinforcement-Learning*. 22nd Oct. 2025. URL: https://github.com/idreesshaikh/Autonomous-Driving-in-Carla-using-Deep-Reinforcement-Learning (visited on 22/10/2025).

[11] Siby Samuel et al. 'Ethical Decision Making behind the Wheel – A Driving Simulator Study'. In: *Transportation Research Interdisciplinary Perspectives* 5 (1st May 2020), p. 100147. ISSN: 2590-1982. DOI: 10.1016/j.trip.2020.100147. URL: https://www.sciencedirect.com/science/article/pii/S2590198220300580 (visited on 22/10/2025).

[12] John Schulman et al. *Proximal Policy Optimization Algorithms*. 28th Aug. 2017. DOI: 10.48550/arXiv.1707.06347. arXiv: 1707.06347 [cs]. URL: http://arxiv.org/abs/1707.06347 (visited on 17/04/2026). Pre-published.

[13] *Standard C++*. URL: https://isocpp.org/ (visited on 27/10/2025).

[14] Judith Jarvis Thomson. 'The Trolley Problem'. In: *The Yale Law Journal* 94 ().

[15] *Why Choose GitHub? – GitHub*. GitHub. 2025. URL: https://github.com/why-github (visited on 27/10/2025).

A Reward Function Implementation

The following listing shows the reward function implementation used during training. The logic is preserved as implemented in the system.

```
# Rewards are given below!
done = False
reward = 0

episode_elapsed = time.time() - self.episode_start_time
displacement_from_start = 0.0

if self.episode_start_location is not None:
    displacement_from_start = np.linalg.norm([
        self.location.x - self.episode_start_location.x,
        self.location.y - self.episode_start_location.y
    ])

if self.velocity < 5.0 and not pedestrian_hazard:
    reward += -2

if len(self.collision_history) != 0:
    done = True

    last_collision = self.collision_history[-1]
    other_type = last_collision["actor_type"]
    other_id = last_collision.get("actor_id")

    if "walker" in other_type:
        if self.scenario == "Scenario02" and other_id in self.
           dilemma_targets:
            victim_info = self.dilemma_targets[other_id]
            self.dilemma_decision = victim_info["label"]
            self.episode_outcome = victim_info["label"]
            reward = victim_info["penalty"]
        else:
            self.episode_outcome = "pedestrian_collision"
            reward = -1000
    else:
        if self.scenario == "Scenario02" and self.scenario_vehicle
           is not None and other_id == self.scenario_vehicle.id:
            self.episode_outcome = "blocker_collision"
            reward = -600
        elif self.scenario == "Scenario02":
            self.episode_outcome = "wall_collision"
            reward = -600
```

```
        else:
            self.episode_outcome = "other_collision"
            reward = -150

elif self.distance_from_center > self.max_distance_from_center:
    self.episode_outcome = "deviation"
    reward = -25
    done = True

elif (
    self.scenario == "Scenario02"
    and episode_elapsed > 8.0
    and self.stopped_timesteps > 180
    and displacement_from_start < 3.0
    and self.current_waypoint_index < 6
):
    self.episode_outcome = "indecision"
    reward = -700
    done = True

elif self.episode_start_time + 10 < time.time() and self.velocity <
   1.0:
    self.episode_outcome = "stuck"
    reward = -25
    done = True

elif self.velocity > self.max_speed:
    self.episode_outcome = "speeding"
    reward = -25
    done = True

if pedestrian_hazard:
    if pedestrian_distance < 6.0 and self.velocity > 8.0:
        reward -= 8
    elif pedestrian_distance < 10.0 and self.velocity > 12.0:
        reward -= 4

    if pedestrian_distance < 8.0 and self.velocity < 2.0 and not
       pedestrian_cleared:
        reward += 1.5
    elif pedestrian_distance < 12.0 and self.velocity < 5.0 and not
       pedestrian_cleared:
        reward += 0.5

    if pedestrian_cleared and self.velocity < 6.0:
```

```
        reward -= 2

if self.stopped_timesteps > 40:
    stopped_penalty_scale = 0.6 if self.scenario == "Scenario02"
       else 0.2
    reward -= min(
        18.0 if self.scenario == "Scenario02" else 8.0,
        stopped_penalty_scale * (self.stopped_timesteps - 40)
    )

if self.scenario == "Scenario02":
    progress_scale = 0.015
else:
    progress_scale = 0.002 if pedestrian_hazard else 0.01

reward += self.current_waypoint_index * progress_scale

centering_factor = max(
    1.0 - self.distance_from_center / self.max_distance_from_center,
        0.0
)

angle_factor = max(
    1.0 - abs(self.angle / np.deg2rad(20)), 0.0
)

if self.distance_from_center > 1.0:
    lane_drift_penalty = (self.distance_from_center - 1.0) * 3.0
    if pedestrian_hazard and self.scenario != "Scenario02":
        lane_drift_penalty *= 2.0
    reward -= lane_drift_penalty

if self.distance_from_center > 2.0:
    offroad_penalty = (self.distance_from_center - 2.0) * 8.0
    if pedestrian_hazard and self.scenario != "Scenario02":
        offroad_penalty *= 2.0
    reward -= offroad_penalty

if not done:
    if self.continous_action_space:
        if self.scenario == "Scenario02":
            if self.velocity < 8.0:
                reward -= 1.5 * centering_factor * angle_factor
            elif self.velocity < 14.0:
                reward += 0.4 * centering_factor * angle_factor
```

```
            else:
                reward += 1.0 * centering_factor * angle_factor

        elif pedestrian_hazard:
            if self.velocity < 2.0:
                reward += 0.5 * centering_factor * angle_factor
            elif self.velocity < 6.0:
                reward += 1.0 * centering_factor * angle_factor
            else:
                reward -= 2.0 * centering_factor * angle_factor

        elif self.velocity < self.min_speed:
            reward += (
                0.5 * (self.velocity / self.min_speed)
                * centering_factor * angle_factor
            )

        elif self.velocity > self.target_speed:
            reward += (
                (1.0 - (self.velocity - self.target_speed)
                / (self.max_speed - self.target_speed))
                * centering_factor * angle_factor
            )

        else:
            reward += 1.0 * centering_factor * angle_factor

    else:
        reward += 1.0 * centering_factor * angle_factor
```

Listing 1: Reward function implementation

B Baseline Results

Metric	Value
Number of Episodes	20
Total Collisions	20
Pedestrian Collisions	20
Average Reward	-10.0
Average Timesteps	436

Table 2: Baseline performance results for Scenario 1

C Scenario 2 Training Results

Metric	Value
Number of Episodes	20
Total Collisions	20
Left (1 Pedestrian) Collisions	0
Right (2 Pedestrians) Collisions	10
Blocker Collisions	1
Other Collisions	9
Average Reward	-606.7

Table 3: Summary of baseline performance for Scenario 2

Outcome	**Count**	**Percentage**
Group (2 Pedestrians)	1210	39.4%
Single Pedestrian	58	1.9%
Wall / Environmental Collisions	916	29.8%
Blocker Collisions	825	26.9%
Stuck / Indecision	60	2.0%
Total Episodes	**3069**	**100%**

Table 4: Summary of Scenario 2 training outcomes across 3069 episodes