A Surrogate Model Framework for Explainable Autonomous Behaviour

Konstantinos Gavriilidis[1], Andrea Munafo[2], Wei Pang[1], Helen Hastie[1]

[1] Edinburgh Centre for Robotics, Heriot-Watt University, Edinburgh, UK

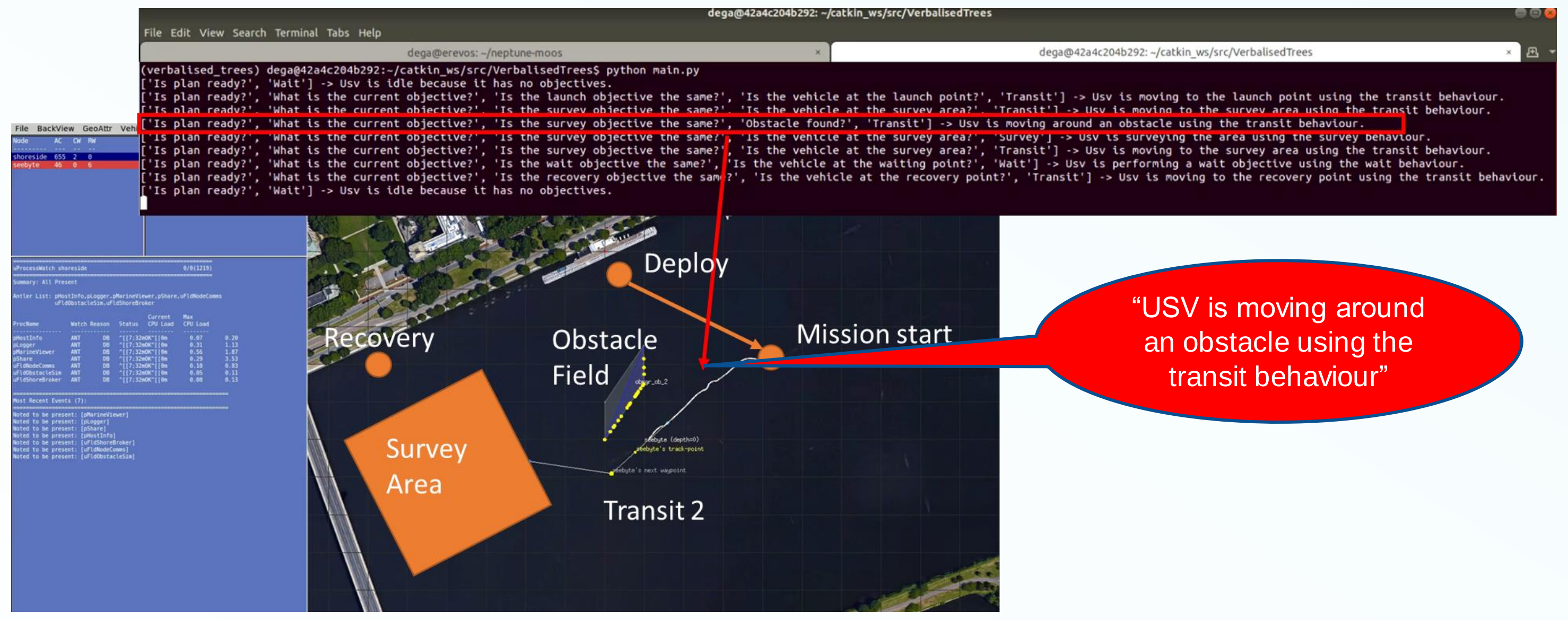
[2] SeeByte Ltd., Edinburgh, UK

Abstract

Adoption and deployment of robotic and autonomous systems in industry are currently hindered by the lack of transparency, required for safety and accountability. Methods for providing explanations are needed that are agnostic to the underlying autonomous system and easily updated. In this work, we use surrogate models to provide transparency as to the underlying policies for behaviour activation. We show that these surrogate models can effectively break down autonomous agents' behaviour into explainable components for use in natural language explanations.

Introduction

1. There is a need to convey rationale behind the decision-making of an autonomous system to operators.
2. We explore various vehicles that can exhibit a variety of behaviours, including:
   1. Autonomous Underwater Vehicles (AUVs) for pipeline inspection and Autonomous Surface Vehicles (USVs) for surveys at designated areas.
3. With this work, we attempt to answer the following **research questions**:
   1. **RQ1**: How robust are surrogate models in policy approximation for behaviour activation?
   2. **RQ2**: Can these surrogate models be used to effectively generate explanations?
   3. **RQ3**: How is the performance affected when going from simulated data to real trials with real vehicles tested in a realistic environment?



```
(verbalised_trees) dega@42a4c204b292:~/catkin_ws/src/VerbalisedTrees$ python main.py
['Is plan ready?', 'Wait'] -> Usv is idle because it has no objectives.
['Is plan ready?', 'What is the current objective?', 'Is the launch objective the same?', 'Is the vehicle at the launch point?', 'Transit'] -> Usv is moving to the launch point using the transit behaviour.
['Is plan ready?', 'What is the current objective?', 'Is the survey objective the same?', 'Is the vehicle at the survey area?', 'Transit'] -> Usv is moving to the survey area using the transit behaviour.
['Is plan ready?', 'What is the current objective?', 'Is the survey objective the same?', 'Obstacle found?', 'Transit'] -> Usv is moving around an obstacle using the transit behaviour.
['Is plan ready?', 'What is the current objective?', 'Is the survey objective the same?', 'Is the vehicle at the survey area?', 'Survey'] -> Usv is surveying the area using the survey behaviour.
['Is plan ready?', 'What is the current objective?', 'Is the survey objective the same?', 'Is the vehicle at the survey area?', 'Transit'] -> Usv is moving to the survey area using the transit behaviour.
['Is plan ready?', 'What is the current objective?', 'Is the wait objective the same?', 'Is the vehicle at the waiting point?', 'Wait'] -> Usv is performing a wait objective using the wait behaviour.
['Is plan ready?', 'What is the current objective?', 'Is the recovery objective the same?', 'Is the vehicle at the recovery point?', 'Transit'] -> Usv is moving to the recovery point using the transit behaviour.
['Is plan ready?', 'Wait'] -> Usv is idle because it has no objectives.
```

dega@42a4c204b292: ~/catkin_ws/src/VerbalisedTrees
File Edit View Search Terminal Tabs Help
dega@erevos: ~/neptune-moos
File BackView GeoAttr Vehi
Deploy
Recovery
Obstacle Field
Mission start
Survey Area
Transit 2

"USV is moving around an obstacle using the transit behaviour"

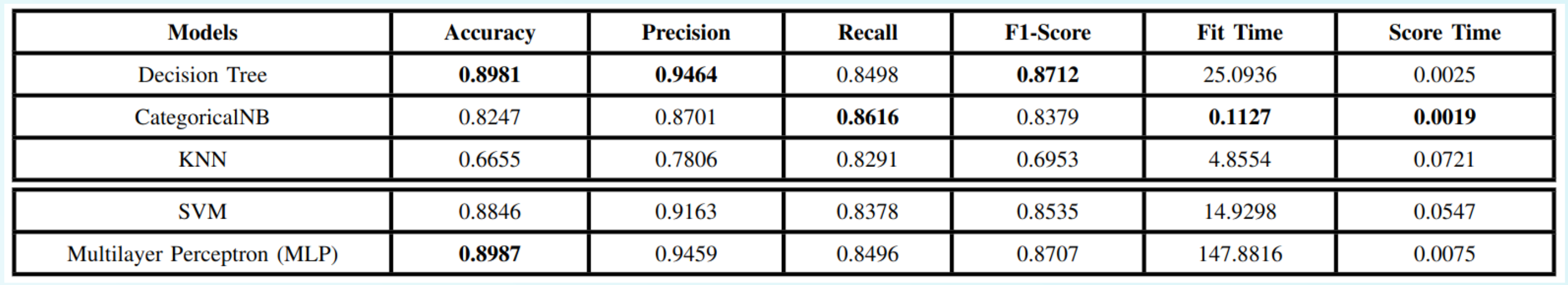
Acknowledgements

We would like to thank MIT's AUV Lab and Laurence Boe from SeeByte Ltd for their assistance with the simulator and the real trial. This work was also funded and supported by the EPSRC Prosperity Partnership (EP/V05676X/1), the UKRI Node on Trust (EP/V026682/1), EPSRC CDT on Robotics and Autonomous Systems (EP/S023208/1), and Scottish Research Partnership in Engineering.

Contact Info

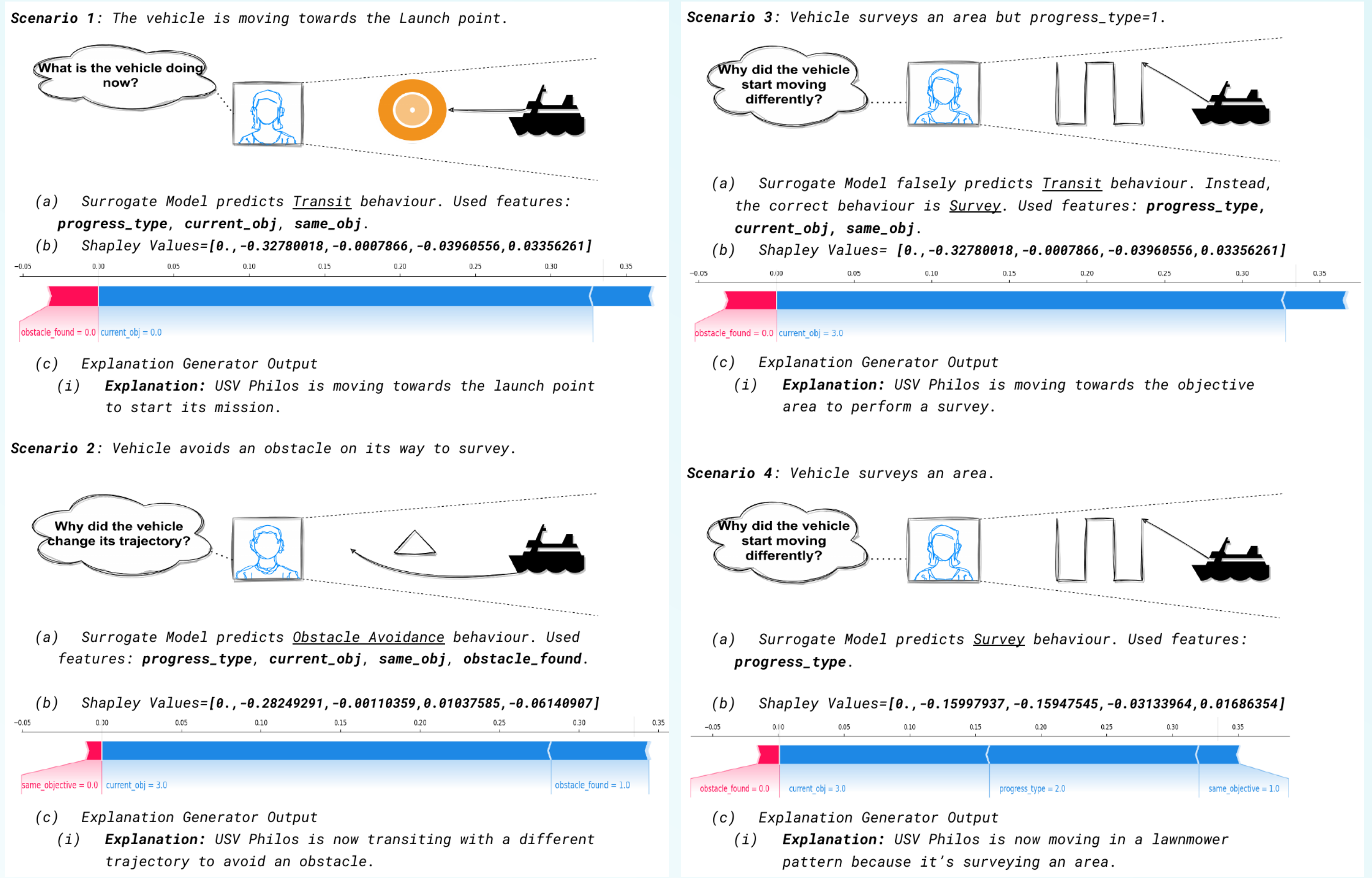
Results

RQ1: Model Selection Metrics



Models	Accuracy	Precision	Recall	F1-Score	Fit Time	Score Time
Decision Tree	**0.8981**	**0.9464**	0.8498	**0.8712**	25.0936	0.0025
CategoricalNB	0.8247	0.8701	**0.8616**	0.8379	**0.1127**	**0.0019**
KNN	0.6655	0.7806	0.8291	0.6953	4.8554	0.0721
SVM	0.8846	0.9163	0.8378	0.8535	14.9298	0.0547
Multilayer Perceptron (MLP)	**0.8987**	0.9459	0.8496	0.8707	147.8816	0.0075

RQ2: Comparison of intrinsic features and Shapley Values



Scenario 1: *The vehicle is moving towards the Launch point.*

(a) *Surrogate Model predicts Transit behaviour. Used features:* ***progress_type**, **current_obj**, **same_obj**.*

(b) *Shapley Values=**[0.,-0.32780018,-0.0007866,-0.03960556,0.03356261]***

(c) *Explanation Generator Output*

(i) ***Explanation:** USV Philos is moving towards the launch point to start its mission.*

Scenario 2: *Vehicle avoids an obstacle on its way to survey.*

(a) *Surrogate Model predicts Obstacle Avoidance behaviour. Used features:* ***progress_type**, **current_obj**, **same_obj**, **obstacle_found**.*

(b) *Shapley Values=**[0.,-0.28249291,-0.00110359,0.01037585,-0.06140907]***

(c) *Explanation Generator Output*

(i) ***Explanation:** USV Philos is now transiting with a different trajectory to avoid an obstacle.*

Scenario 3: *Vehicle surveys an area but progress_type=1.*

(a) *Surrogate Model falsely predicts Transit behaviour. Instead, the correct behaviour is Survey. Used features:* ***progress_type**, **current_obj**, **same_obj**.*

(b) *Shapley Values= **[0.,-0.32780018,-0.0007866,-0.03960556,0.03356261]***

(c) *Explanation Generator Output*

(i) ***Explanation:** USV Philos is moving towards the objective area to perform a survey.*

Scenario 4: *Vehicle surveys an area.*

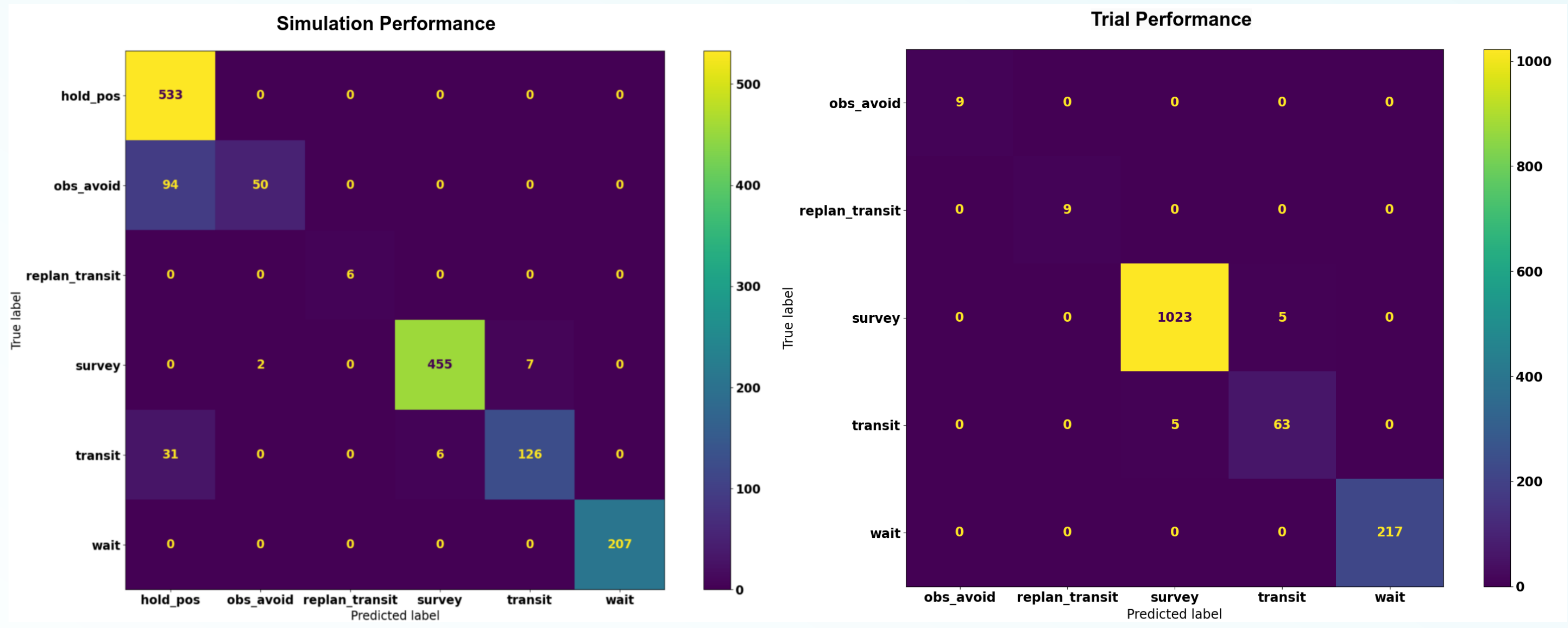
(a) *Surrogate Model predicts Survey behaviour. Used features:* ***progress_type**.*

(b) *Shapley Values=**[0.,-0.15997937,-0.15947545,-0.03133964,0.01686354]***

(c) *Explanation Generator Output*

(i) ***Explanation:** USV Philos is now moving in a lawnmower pattern because it's surveying an area.*

RQ3: Behaviour Predictions

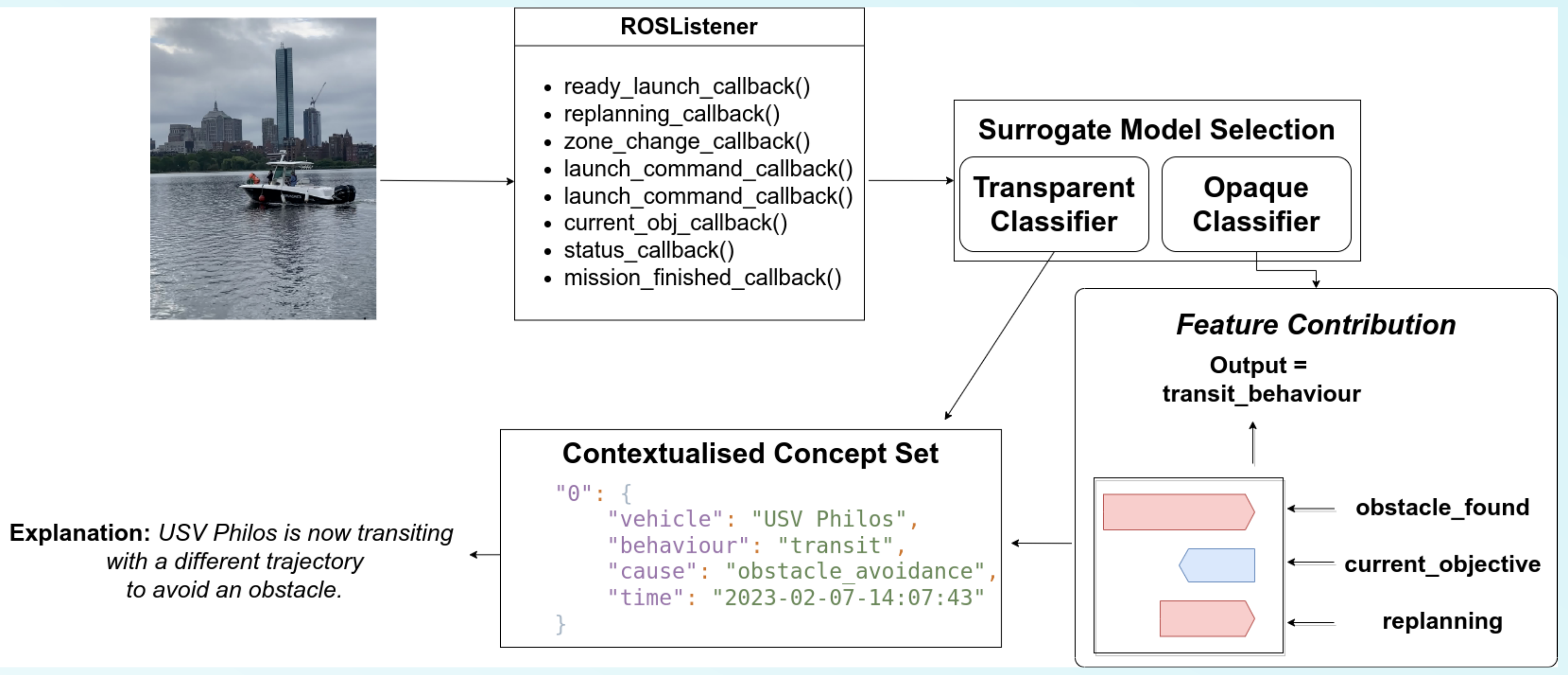


For **RQ1**, we demonstrate the classification performance across multiple models and select the best for our use case (Decision Tree). For **RQ2**, we made a comparison of intrinsic features within our surrogate model and corresponding Shapley Values to compare each estimated causality. Finally, for **RQ3** we present classification performance for both simulated and real missions.

Methodology

The proposed framework consists of the following steps:

1) Extraction of vehicle states for surrogate model training and real-time explanation generation.
2) Model Selection with Nested Cross-validation to choose optimal Surrogate Model.
   1) Use of intrinsic features for transparent models or feature contribution estimation for opaque models.
3) Representation of exhibited behaviours with Contextualised Concept Sets.
4) Use of concept sets to generate Natural Language Explanations.



ROSListener

- ready_launch_callback()
- replanning_callback()
- zone_change_callback()
- launch_command_callback()
- launch_command_callback()
- current_obj_callback()
- status_callback()
- mission_finished_callback()

Surrogate Model Selection

Transparent Classifier | **Opaque Classifier**

Feature Contribution

Output = transit_behaviour

obstacle_found

current_objective

replanning

Contextualised Concept Set

```
"0": {
    "vehicle": "USV Philos",
    "behaviour": "transit",
    "cause": "obstacle_avoidance",
    "time": "2023-02-07-14:07:43"
}
```

Explanation: *USV Philos is now transiting with a different trajectory to avoid an obstacle.*

Conclusion & Future Work

1. A domain-agnostic framework for approximating behaviour activations and replanning of an autonomous agent with classification models has been introduced.
2. Our approach is capable of discovering the causality of autonomous decisions and storing that information with Contextualised Concept Sets.
3. Moving forward, we plan on using these representations to investigate data-driven language explanations such as large language models.
4. Further evaluation of explanations is also required to examine the capacity of our approach to disambiguate robotic behaviours.