A Surrogate Model Framework for Explainable Autonomous Behaviour

Konstantinos Gavriilidis
Heriot-Watt University
Edinburgh, United Kingdom
kg47@hw.ac.uk

Andrea Munafo
SeeByte Ltd.
Edinburgh, United Kingdom
andrea.munafo@seebyte.com

Wei Pang
Heriot-Watt University
Edinburgh, United Kingdom
W.Pang@hw.ac.uk

Helen Hastie
Heriot-Watt University
Edinburgh, United Kingdom
H.Hastie@hw.ac.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. Furthermore, different stakeholders with varying levels of expertise, will require different levels of information. 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.**

***Index Terms*—Explainable Agents, Human-In-The-Loop Application, Surrogate Model, Feature Contribution.**

I. Introduction

Robotic and autonomous systems are at the stage where we are seeing them being adopted more frequently in a variety of environments, such as ground vehicles for inspection of disaster sites, or underwater for pipeline inspection. It is important that humans are kept in the loop in these operations and are able to intervene as necessary. However, this comes with challenges, such as underwater vehicles having limited bandwidth to broadcast updates [1]. Transparency and the ability to explain actions and decisions are key factors for safety, accountability and adoption [2]. However, these are non-trivial to implement, given the complexity of autonomous systems and the 'blackbox' nature of neural-based models.

Platform-specific explanation interfaces normally require a basic understanding of an agent's behaviour space (B), possible states (S) and decision-making (D) to comprehend its capabilities and what could be described to operators. Furthermore, user studies are necessary to recognise which behaviours may be certainly valid and appropriate but might perhaps confuse the operator. This can lead to mission aborts and inaccurate mental models [3].

These user studies and explanation methods also need to adapt to the stakeholder. The IEEE Standard for Transparency of Autonomous Systems (P7001) [2], defines a number of stakeholder groups from the expert operator, to the general public and lawmakers. They all require different types of information and level of detail to be included in the explanation.

Furthermore when working with commercial entities, they are continuously developing their autonomous models, adding new behaviours or states as required for new use cases and customers. If we are to provide accurate, up-to-date explanations, the explanation module will also need continuous updating, requiring considerable time and effort.

Fig. 1: Unmanned Surface Vehicles (USVs) Heron (left) and Philos (right) used during the trials on Charles River in Boston.

Here, we propose a generic method using surrogate models for generating autonomy-agnostic explanations that can be used without a deep understanding of the underlying autonomy and can be easily updated. Specifically addressing the following research questions:

- **RQ1:** How robust are surrogate models in approximating a complex deterministic agent's policy for behaviour activation?
- **RQ2:** Can these surrogate models be used to effectively generate explanations?
- **RQ3:** How is the performance affected when going from simulated data to real trials with real vehicles tested in a realistic environment?

For the remainder of this paper, in Section II we mention previous work, which has impacted our approach and in Section III we describe our use case and the explanation types that our framework provides. In Section IV, we describe the functionality of components, such as the surrogate models or the feature contribution estimators in the pipeline architecture. Finally, in Section V, we report the performance of the

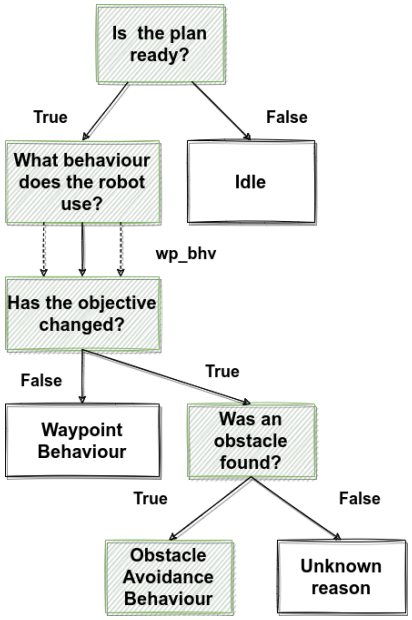


Fig. 2: Empirical decision tree for behaviour activation derived by a domain expert.

surrogate models in simulations and how it is affected during the trial or when new behaviours are incremented.

II. Related Work

Explanation Frameworks: Explainable agency has been introduced as a trait of robots to define what properties a transparent robot should have conceptually. Among these traits are the ability to explain (i) plan generation, (ii) executed actions and (iii) replanning in a user-friendly fashion [4]. Looking at previous work on explainable agency, the whole process can be broken down into three main parts, *explanation generation*, *explanation communication* and *explanation reception* [5]. For explanation communication, previous studies have looked into how robots could explain themselves as people do by including reasons for intentional and causes for unintentional behaviours [6]. Further studies investigate the desired verbosity of explanations in different scenarios [7], [8] and various types of explanations, which can be provided to a user from an explainable planning perspective [9].

Explainable Artificial Intelligence: The right of users to receive explanations about a robot's behaviour is supported by government regulations and the recommended direction is the development of transparent robotics [2], e.g. through the use of interpretable models [10]. Machine learning models differ in terms of simulatability, decomposability and algorithmic transparency [11]. In the case of opaque models, explanation methods should be applied to them to disambiguate their functionality. Explanation methods fall into two categories depending on the way they are applicable to a black-box model, namely *Model-specific* and *Model-agnostic* [12]. Surrogate models can belong to either category, depending on their intended use and are useful for deriving the causality behind any prediction. LIME [13] accomplishes this by locally approximating the model around a given prediction. SHAP [14] generates Shapley values, which indicate the contribution of each feature to the difference between the initial belief of a model and its actual prediction. Another option to highlight causal relationships is through counterfactuals, where several feature contributions can help the user understand the relationship between feature values and the corresponding predictions [15]. However, surrogate models are not always consistent and, as a result, robustness has been introduced as a metric that represents how stable explanations are when inputs are slightly modified [16].

Explainable Robotics: In Sakai and Nagai [17], the relationship between XAI and explanations for transparent robotics is defined. Existing work has examined the use of both algorithmically transparent models and the combination of opaque models with posthoc explanation methods to explain robotic failures [18]. The decision-making of reinforcement learning agents has also been explained either with the use of surrogate models [19], [20] or the generation of Shapley values to explain robot grasping [21]. Focusing more on explanation communication and reception, the studies in Thielstrom et al. [22] and Robb et al. [23] present videos of robotic failures along with explanations to users. Meanwhile, in Das et al. [24], natural language explanations are generated with a Neural Translation Network to improve human assistance in fault recovery. Each approach performed a corresponding user study to evaluate how explanations affect the mental model of users.

Continuing our effort in [25], we have developed a framework that retrieves data from deterministic agents and, with model selection, finds the optimal classifier for behaviour prediction. Depending on the transparency of the intermediate model, we capture the causality behind behaviours either by directly analysing the model or with the application of a posthoc explainer.

Knowledge Representation and Verbalisation: Knowledge representations have always played an important role in the unification or the completion of knowledge for autonomous agents. In Li et al. [26], an ontology is used to tackle the issue of information heterogeneity and to facilitate collaboration between underwater robots. Additionally, in Gavriilidis et al. [27] an ontology is used to relate sensor readings to hardware errors and make a new plan, while a ROS listener retrieves and verbalises these outcomes with a surface realiser. Furthermore, Suh et al. [28] use a multilayered ontology to complement the perception of a household robot for object recognition and assist with its localisation. At the same time, knowledge representations play an important role in Natural Language Generation. In [29], a Neural Language Model efficiently transforms Wikipedia infoboxes into biography summaries, while in [30] a fine-tuned T5 model generates sentences just by connecting plain utterances from concept sets. On the other hand, Ghosal [31] collected a dialogue reasoning dataset, where additional context is incorporated into utterances to teach a T5 model to make more intuitive transitions in dialogue. This type of data-driven natural language generation

is out of scope for the work described here, but is clearly an area for future use, particularly with the advent of more advanced large language models such as GPT-4 [32].

III. Use Case and Explanation Types

The use case examined here is a hybrid autonomy, that combines a ROS-based deterministic agent with a reactive agent, prioritising behaviours through multi-objective optimization for the maritime domain. Specifically, our scenarios focus on *Unmanned Surface Vehicles* (USV) and *Autonomous Underwater Vehicles* (AUV), as illustrated in Figure 1. This work was done in collaboration with industry partner SeeByte Ltd, who have developed an autonomous agent for driving such vehicles for a variety of maritime applications, such as inspection. To have an initial understanding of the autonomous agent and how it selects a behaviour, we interviewed in-depth a domain expert from the company. From this interview, we derived an abstract definition of behaviour decision-making in a tree format. This *empirical decision tree* is illustrated in Figure 2.

We then investigated if this behaviour tree covers aspects of a mission in simulation ahead of the real trial. The simulation scenario involves a restricted coastal area on the River Charles in Boston, where two vehicles collaborate to complete a mission. The versatile USV Heron performs each objective according to the uploaded plan, while USV Philos inspects the area to detect obstacles and notifies the other vehicle if something out of the plan is found. Each mission contains a launch and recovery objective and a number of survey or target objectives, where the vehicle needs to hold its position for a default amount of time. Additionally, the obstacles are either static (with locations specific to each mission but there for the duration), or dynamic (i.e. appearing during the mission). In terms of capabilities, both vehicles support 6 behaviours B = *(wait, transit, survey, hold_position, replanned_transit, avoid_obstacle)* and they both run two autonomy models simultaneously (one for each vehicle) in a master-slave architecture.

The following explanation types were derived in consultation with the expert and captured in the empirical decision tree in Figure 2 and are listed here.

E.1 Behaviour Causality describes how a robot selects its current behaviour or action. Especially for operators, it can be difficult to comprehend how a robot closely observes objects around it, updates its world model and acts according to its goals. The utterances of this explanation usually entail the name of the behaviour, its use and the cause of activation. ***Answers question:*** *Why did you do that?*

E.2 Replanning Clarification complements the previous category and covers cases where unexpected outcomes arise and force the autonomy to alter its plan. Some indicative examples are obstacle avoidance and platform integrity, where for safety reasons the robot has to make a stop in a new location. ***Answers question:*** *Why do I need to replan at this point?*

E.3 Counterfactual Explanation allows the operator to ask the autonomy how it would react if its internal state changed in a specific way. With this functionality, the user can learn about alternative outcomes at any given point and to better comprehend the underlying logic of the autonomous agent. ***Answers question:*** *What if?*

IV. Method

The overall pipeline architecture that goes from the autonomous vehicle to the explanation interface is illustrated in Figure 3. Its aim is to act as a wrapper application that does not disrupt the existing autonomy but clearly conveys the approximated policy to users. For the ROSListening component, we found the relevant ROS topics that provide the vehicle states needed to predict exhibited behaviours (per the behaviour definition in Figure 2). We then created a listener with two uses in mind: (i) data collection in simulation and (ii) online behaviour prediction during plan execution. Using the acquired data, we trained a number of classifier models as surrogate autonomy models. These models predict which of the 6 behaviours the vehicle is exhibiting. We investigated a number of classifiers with varying transparency and compared the accuracy of the models.

If the highest accuracy model is transparent, we would directly extract the feature contribution for each behaviour prediction. Otherwise, a post-hoc explanation method would be applied to the opaque model to derive feature contribution. Here, we examine both of these options. Finally, a knowledge representation that contains this information is generated and fed into a rule-based natural language explanation generator that conveys the same content in a user-friendly format. These components are described in detail below and represented in Algorithm 1.

A. The Data

We collected a Behaviour Causality Dataset from 10 simulations of missions. For each simulation, we monitored eight ROS topics with a listener module corresponding to five vehicle states where S = {*ready_plan, current_objective, progress_type, same_objective, obstacle_found*} along with the corresponding behaviour. Each mission lasted 22.5 minutes on average and resulted in a dataset of 5056 data instances with 5 categorical features and a target value.

B. Surrogate Model Training and Selection

After data collection, we made a comparison of various models to decide on the most suitable option for Behaviour Classification. Specifically, we tested three algorithmically explainable models, which are robust for classification with categorical features (K-Nearest Neighbours (KNN) [33], Categorical Naive Bayes (CategoricalNB) [34] and Decision Tree [35]) and we also included two more complex models (Support Vector Machine (SVM) [36], Multilayer Preceptron (MLP) [37]) to check if there is a significant performance difference.

A total of 5 categorical features were given as input to each model (ready_plan, current_obj, progress_type, same_obj, obstacle_found) to predict the current behaviour of the vessel

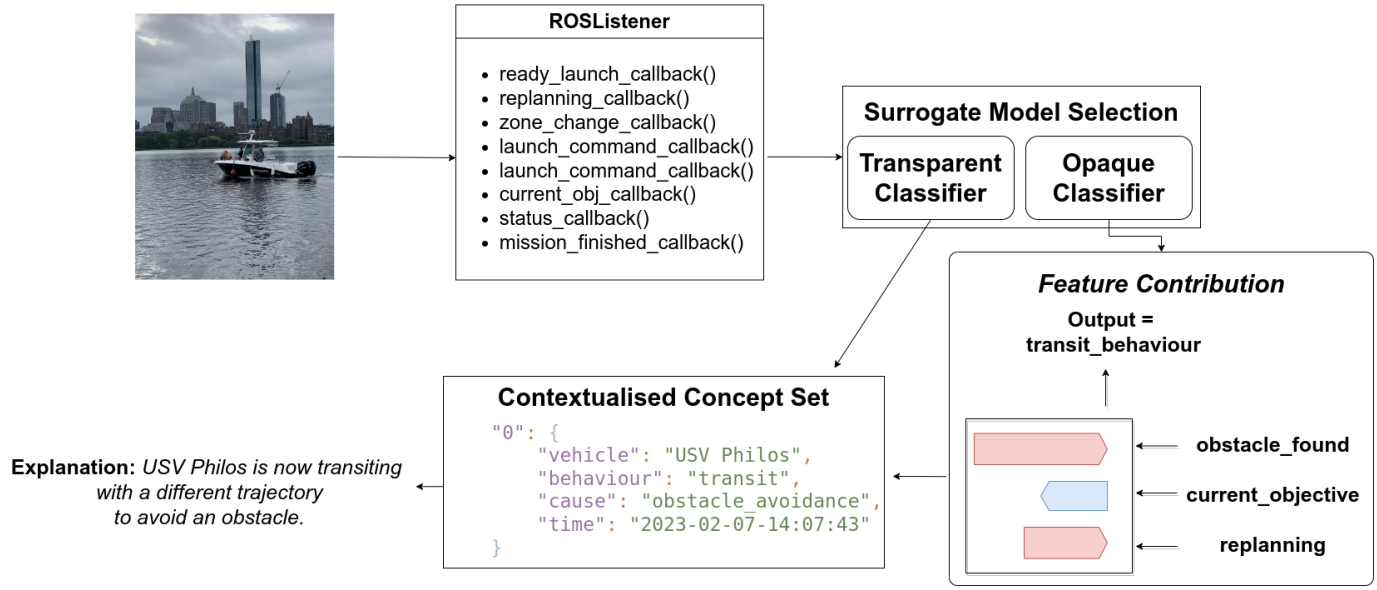


Fig. 3: Illustration of the proposed pipeline architecture, where a Surrogate Model approximates agent policy and Feature Contribution is estimated to detect behaviour causality. The output of this framework is an explanation with content which originates from a Contextualised Concept Set.

Algorithm 1: Explanation Framework Workflow

Input: $\mathcal{UP}$: User Preference.
Input: $\mathcal{SD}$: State Dataset.
Output: $\mathcal{IKR}$: Intermediate Knowledge Representation.
Output: $\mathcal{SV}$: State Verbalisation.

begin
  $\mathcal{UC} \leftarrow$ SelectUseCase($\mathcal{UP}$)
  $D \leftarrow$ SplitAndEncode($\mathcal{SD}$,$\mathcal{UP}$)
  $\mathcal{M} \leftarrow$ TrainModel($\mathcal{D}$,$\mathcal{UP}$)
  $\mathcal{E} \leftarrow$ ConfigureExplainer($\mathcal{M}$,$\mathcal{UP}$)
  $s \leftarrow$ StartExecution($\mathcal{UC}$)
  while ***not*** $s \leftarrow \emptyset$ **do**
    $p \leftarrow M$.Predict(s)
    $e \leftarrow E$.ExplainPrediction(p)
    $r \leftarrow$.GenerateRepresentation(p, e)
    $v \leftarrow$.VerbaliseRepresentation(r)
    if $\mathcal{UC} \leftarrow$ *isFinished* **then**
      $i \leftarrow \emptyset$
    else
      $s \leftarrow$ GetNextState()

(wait, transit, survey, hold_position, replanned_transit, obstacle_avoidance). Nested cross-validation was used to select the best combination of hyperparameters for each model and to retrieve unbiased metrics indicative of each model's performance [38].

From the results in Table I, for the transparent models, it is clear that the Decision Tree and Categorical Naive Bayes algorithms outperform KNN. With regards the more opaque algorithms, SVM achieved similar performance to MLP but with the latter performing slightly better at correctly classifying behaviours. As for model training and evaluation, the time needed for transparent models to do both was much shorter than for Neural Networks. Thus going forward, we decided to use the decision tree (*max_depth = 8, max_leaf_nodes = 15*) given that it has similar accuracy to more complex models. Furthermore, its high transparency means that it can be used to verify the validity of the surrogate framework for more opaque models, in case these are used in future use cases and datasets.

Figure 4 provides a confusion matrix of the predictions of the decision tree per behaviour. For **transit**, **hold_pos** and **survey** behaviours, there are some false classifications due

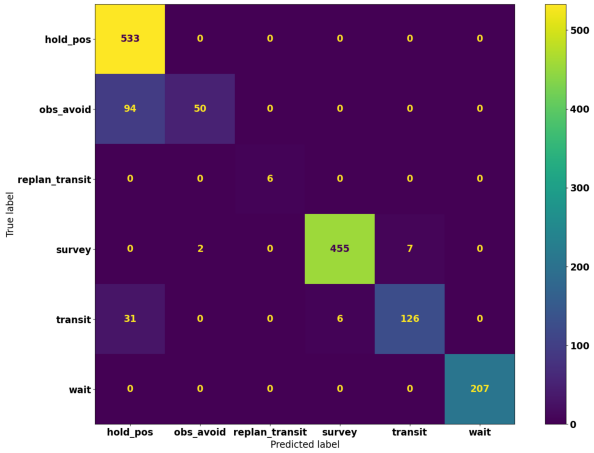


Fig. 4: Confusion matrix indicating classification performance per behaviour with a Decision Tree during simulations.

to some inconsistency between the progress_type feature and the corresponding behaviour, which indicates that an internal autonomy state could be missing. As for false classifications between **hold_pos**, **survey** and **obs_avoid** behaviours, we noticed that even though replanning is triggered and an obstacle is found, the vehicle finds a way to perform its objective, however, the explanation framework misses this fact probably because an internal autonomy state is missing once again.

C. Explanation Layer

Once a trained surrogate model is in place to predict the corresponding behaviour of a vehicle state, the feature contribution for the classification of the behaviour is used as a basis for the causal reasoning explanation. One way to do feature contribution is to examine the trained surrogate model itself [13]. This is feasible for transparent models, such as the decision tree chosen here, but not so for more complex models such as Neural Networks. Opaque models such as Neural Networks may be needed in future applications as the complexity of the autonomy increases and the datasets grow in size. For these more complex models, an alternative is to use *Shapley Values* [14], which has been shown to be a reliable and descriptive approach. Here, we follow this latter method as a proof of concept. Each model has initially a prior belief about what the expected value will be and a Shapley value describes how a specific feature creates the difference between the expected and actual values ($E(x) - f(x)$).

D. Knowledge Representation and Explanation Generation

The final two components of the pipeline use, as input, the prediction of the surrogate model along with the Shapley values estimated by the feature contribution estimator. Based on the importance of each feature towards a prediction, behaviour causality is inferred and this knowledge is represented with contextualised concept sets. Contextualisation is incorporated with the use of key-value pairs, as opposed to simple triplets to indicate the role of each value. The end result is a knowledge base with $(vessel, behaviour, causality, time)$ sets, which describe the sequence of behaviours exhibited by the robot in JSON format. An example of a contextualised concept set can be found in Figure 3, where the current behaviour (Transit) and its trigger (Obstacle) can be distinguished. This entry indicates that the current transit behaviour has a modified trajectory that goes around the obstacle to avoid collision. With regards to natural language generation, for each new entry in the Knowledge Base, the key-value pairs are passed to a *Surface Realiser*, which produces an explanation that has been syntactically checked with SimpleNLG [39].

V. Results and Discussion

With regards **RQ1** and **RQ2**, as discussed above and in Table I, the surrogate models have accuracies for behaviour prediction of around 90%. This could further be further improved by training on more data both in simulation and with real vehicles.

Even with this accuracy, we observed accurate explanations as give illustrated in Figure 5. In this figure, we present four continuous scenarios from a single mission along with the explanations which were generated. *Scenario 1* involves a vessel moving to the launch point to retrieve the relative positions of the objective areas and begins working on each objective. In this case, the surrogate model correctly predicts the current behaviour by using the **progress_type**, **current_obj** and **same_obj** features. To validate the results, we also calculated the Shapley values, which highlight the **current_obj** as the main contributor. In *Scenario 2*, while the vehicle is moving from the launch point to the survey area, it encounters an obstacle and avoids it by changing its trajectory. Behaviour prediction was also successful in this case, with the model utilising **progress_type**, **current_obj**, **same_obj** and **obstacle_found** features. In *Scenario 3*, a false explanation is generated, due to the value of **progress_type** even though the survey has already started. Here, the features used by the surrogate model are, **progress_type**, **current_obj** and **same_obj**, while SHAP only attributes this prediction to the current objective. As for *Scenario 4*, where the vessel performs a survey, the surrogate model can immediately detect the new behaviour thanks to its unique **progress_type**, while SHAP adds to the causality both **current_obj** and **same_obj**, which seem reasonable causes in this case. An informal evaluation has been conducted but a formal subject evaluation is future work.

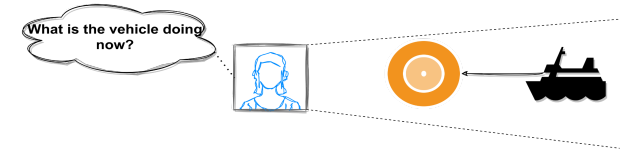
A. Going from Simulation to Real Trial

For **RQ3**, a real trial took place with two Unmanned Surface Vehicles collaborating to complete a survey, while obstacles appeared in the dynamic environment. As a result, we tested the surrogate model on a separate trial test set consisting of 1331 instances with 5 features and corresponding behaviours. Looking at the overall performance, an accuracy of 99% was achieved showing the capability of the surrogate model to comprehend the autonomous behaviour activations. See Figure 6 for the confusion matrix for this test. The only errors observed are false classifications between the **survey** and **transit** behaviours, which we attribute to a potential missing vehicle state since the **progress_type** could not fully indicate the transition between these two behaviours. Finally, the **hold_position** behaviour is missing from Figure 6, because this objective was not used during the trial for practical reasons.

VI. Conclusion and Future Work

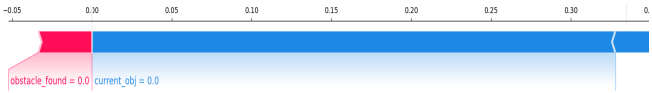
With this work, a framework for approximating behaviour activations and replanning of an autonomous agent with classification models has been introduced. Our approach is capable of discovering the causality of autonomous decisions with the estimation of feature contribution for each action prediction. The main advantage of this framework is the storage of information in generic knowledge representations such as concept sets which can be later leveraged to produce user-friendly modalities such as natural language explanations.

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



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

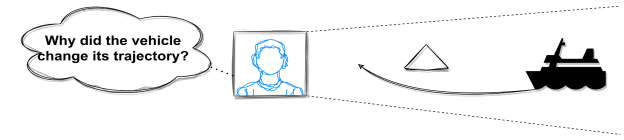
*(b) Shapley Values=**[0.,-0.32780018,-0.0007866,-0.03960556,0.03356261]***



(c) Explanation Generator Output

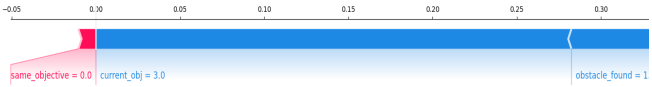
(i) ***Explanation:** USV Philos is moving towards the launch point to start its mission.*

***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.***

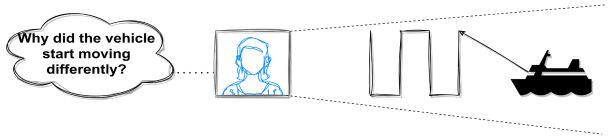
*(b) Shapley Values=**[0.,-0.28249291,-0.00110359,0.01037585,-0.06140907]***



(c) Explanation Generator Output

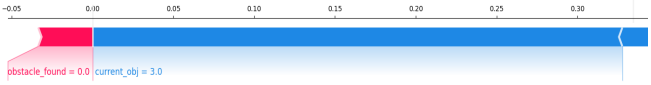
(i) ***Explanation:** USV Philos is now transiting with a different trajectory to avoid an obstacle.*

***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.***

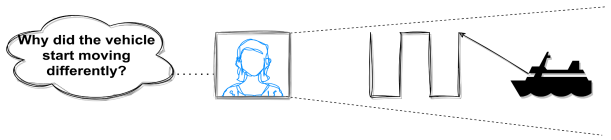
*(b) Shapley Values= **[0.,-0.32780018,-0.0007866,-0.03960556,0.03356261]***



(c) Explanation Generator Output

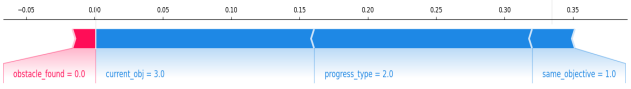
(i) ***Explanation:** USV Philos is moving towards the objective area to perform a survey.*

***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.*

Fig. 5: Four continuous events from a single mission along with their behaviour predictions and the corresponding explanations. Scenarios 1, 2 and 4 contain correctly predicted behaviours, while Scenario 3 demonstrates a false prediction that has been encountered.

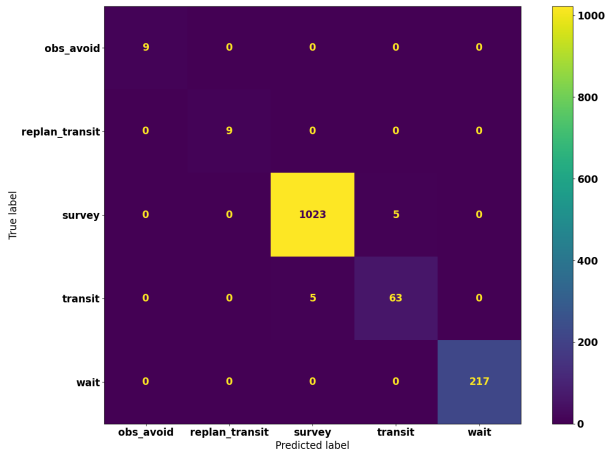


Fig. 6: Confusion matrix indicating classification performance per behaviour with a Decision Tree during the real trial.

Another advantage is that the framework is agnostic to the autonomy model, making it reusable in different domains. Moving forward, we plan on extending the functionality of this framework and investigating data-driven language explanations such as large language models in order to stochastically map knowledge representations to informative natural language explanations about robotic behaviour. Further evaluation of explanations is also required to examine the capacity of our approach to disambiguate robotic behaviours.

ACKNOWLEDGMENT

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.

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

TABLE I: Classification performance metrics across five different models derived with nested cross-validation on simulation data. Transparent models are at the top part of the table and opaque ones are at the bottom. Training and score times (in seconds) are also included to indicate the effort for acquiring a surrogate model in other similar use cases.

REFERENCES

[1] Paull, L., Saeedi, S., Seto, M. and Li, H., 2013. AUV navigation and localization: A review. IEEE Journal of oceanic engineering, 39(1), pp.131-149.

[2] Winfield, A.F., Booth, S., Dennis, L.A., Egawa, T., Hastie, H., Jacobs, N., Muttram, R.I., Olszewska, J.I., Rajabiyazdi, F., Theodorou, A. and Underwood, M.A., 2021. IEEE P7001: A proposed standard on transparency. Frontiers in Robotics and AI, 8, p.665729.

[3] Hastie, H., Liu, X. and Patron, P., 2017, November. Trust triggers for multimodal command and control interfaces. In Proceedings of the 19th ACM International Conference on Multimodal Interaction (pp. 261-268).

[4] Langley, P., Meadows, B., Sridharan, M. and Choi, D., 2017, February. Explainable agency for intelligent autonomous systems. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 31, No. 2, pp. 4762-4763).

[5] Anjomshoae, S., Najjar, A., Calvaresi, D. and Främling, K., 2019. Explainable agents and robots: Results from a systematic literature review. In 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019 (pp. 1078-1088). International Foundation for Autonomous Agents and Multiagent Systems.

[6] De Graaf, M.M. and Malle, B.F., 2017, October. How people explain action (and autonomous intelligent systems should too). In 2017 AAAI Fall Symposium Series.

[7] Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I. and Wong, W.K., 2013, September. Too much, too little, or just right? Ways explanations impact end users' mental models. In 2013 IEEE Symposium on visual languages and human centric computing (pp. 3-10). IEEE.

[8] Garcia, F.J.C., Robb, D.A., Liu, X., Laskov, A., Patron, P. and Hastie, H., 2018, November. Explainable autonomy: A study of explanation styles for building clear mental models. In Proceedings of the 11th International Conference on Natural Language Generation (pp. 99-108).

[9] Fox, M., Long, D. and Magazzeni, D., 2017. Explainable planning. arXiv preprint arXiv:1709.10256.

[10] Wachter, S., Mittelstadt, B. and Floridi, L., 2017. Transparent, explainable, and accountable AI for robotics. Science robotics, 2(6), p.eaan6080.

[11] Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R. and Chatila, R., 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information fusion, 58, pp.82-115.

[12] Adadi, A. and Berrada, M., 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). (pp.52138-52160). IEEE access.

[13] Ribeiro, M.T., Singh, S. and Guestrin, C., 2016, August. " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

[14] Lundberg, S.M. and Lee, S.I., 2017. A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

[15] Chou, Y.L., Moreira, C., Bruza, P., Ouyang, C. and Jorge, J., 2022. Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. Information Fusion, 81, pp.59-83.

[16] Alvarez-Melis, D. and Jaakkola, T.S., 2018. On the robustness of interpretability methods. arXiv preprint arXiv:1806.08049.

[17] Sakai, T. and Nagai, T., 2022. Explainable autonomous robots: A survey and perspective. Advanced Robotics, 36(5-6), pp.219-238.

[18] Alvanpour, A., Das, S.K., Robinson, C.K., Nasraoui, O. and Popa, D., 2020, August. Robot failure mode prediction with explainable machine learning. In 2020 IEEE 16th International Conference on Automation Science and Engineering (CASE) (pp. 61-66). IEEE.

[19] Gjærum, V.B., Strümke, I., Løver, J., Miller, T. and Lekkas, A.M., 2023. Model tree methods for explaining deep reinforcement learning agents in real-time robotic applications. Neurocomputing, 515, pp.133-144.

[20] Sieusahai, A. and Guzdial, M., 2021, October. Explaining deep reinforcement learning agents in the atari domain through a surrogate model. In Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (Vol. 17, No. 1, pp. 82-90).

[21] Remman, S.B. and Lekkas, A.M., 2021, June. Robotic lever manipulation using hindsight experience replay and shapley additive explanations. In 2021 European Control Conference (ECC) (pp. 586-593). IEEE.

[22] Thielstrom, R., Roque, A., Chita-Tegmark, M. and Scheutz, M., 2020, November. Generating explanations of action failures in a cognitive robotic architecture. In 2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (pp. 67-72).

[23] David A. Robb, Xingkun Liu, and Helen Hastie. 2023. Explanation Styles for Trustworthy Autonomous Systems: Extended Abstract. In Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023, IFAAMAS.

[24] Das, D., Banerjee, S. and Chernova, S., 2020. Explainable AI for System Failures: Generating Explanations that Improve Human Assistance in Fault Recovery. arXiv preprint arXiv:2011.09407.

[25] Gavriilidis, K., Munafo, A., Hastie, H., Cesar, C., DeFilippo, M. and Benjamin, M.R., 2022. Towards Explaining Autonomy with Verbalised Decision Tree States. arXiv preprint arXiv:2209.13985.

[26] Li, X., Bilbao, S., Martín-Wanton, T., Bastos, J. and Rodriguez, J., 2017. SWARMs ontology: A common information model for the cooperation of underwater robots. Sensors, 17(3), p.569.

[27] Gavriilidis, K., Carreno, Y., Munafo, A., Pang, W., Petrick, R.P. and Hastie, H., 2021, August. Plan Verbalisation for Robots Acting in Dynamic Environments. In Proceedings of ICAPS 2021 Workshop on Knowledge Engineering for Planning and Scheduling.

[28] Suh, I.H., Lim, G.H., Hwang, W., Suh, H., Choi, J.H. and Park, Y.T., 2007, October. Ontology-based multi-layered robot knowledge framework (OMRKF) for robot intelligence. In 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems (pp. 429-436). IEEE.

[29] Lebret, R., Grangier, D. and Auli, M., 2016. Neural text generation from structured data with application to the biography domain. arXiv preprint arXiv:1603.07771.

[30] Lin, B.Y., Zhou, W., Shen, M., Zhou, P., Bhagavatula, C., Choi, Y. and Ren, X., 2019. CommonGen: A constrained text generation challenge for generative commonsense reasoning. arXiv preprint arXiv:1911.03705.

[31] Ghosal, D., Shen, S., Majumder, N., Mihalcea, R. and Poria, S., 2022. CICERO: A Dataset for Contextualized Commonsense Inference in Dialogues. arXiv preprint arXiv:2203.13926.

[32] OpenAI, 2023. "GPT-4 Technical Report.". arXiv preprint arXiv:2303.08774.

[33] Larose, D.T. and Larose, C.D., 2014. k-nearest neighbour algorithm.

[34] Zhang, H., 2004. The optimality of naive Bayes. Aa, 1(2), p.3.

[35] Vega, F.A., Matías, J.M., Andrade, M.L., Reigosa, M.J. and Covelo, E.F., 2009. Classification and regression trees (CARTs) for modelling the sorption and retention of heavy metals by soil. Journal of Hazardous Materials, 167(1-3), pp.615-624.
[36] Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J. and Scholkopf, B., 1998. Support vector machines. IEEE Intelligent Systems and their applications, 13(4), pp.18-28.
[37] Hinton, G.E., 1990. Connectionist learning procedures. In Machine learning (pp. 555-610). Morgan Kaufmann.
[38] Cawley, G.C. and Talbot, N.L., 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. The Journal of Machine Learning Research, 11, pp.2079-2107.
[39] Gatt, A. and Reiter, E., 2009, March. SimpleNLG: A realisation engine for practical applications. In Proceedings of the 12th European workshop on natural language generation (ENLG 2009) (pp. 90-93).