Explaining Robot Decisions Using Contrasting Examples

Ian C. Rankin[1], Seth McCammon[2], Thane Somers[3], Geoffrey A. Hollinger[1]

***Abstract*— In this paper, we propose the Contrastive, Feature-based eXplainability (CoFeX) method for explaining robotic decisions that provides: (1) contrasting examples, which illustrate the decision in terms of an alternative example, (2) post-hoc explanations that are generated after the decision making algorithm is run, which allows a wide range of decision making algorithms to be used, and (3) a feature-importance method that produces explanations that provide more detail than just trade-offs between objectives. Explainability of robotic actions is important for users to understand, trust, and manage robotic systems. While there are a few existing robotic explainability methods, they are either noncontrastive, rely on simple models that limit the complexity of the decision making algorithms, or use high-level trade-offs for the causal reasons which do not provide detailed explanations. To allow our explanation method to work with a wide range of decision making algorithms, we use a shared language of semantic features for communication between humans and robots. We then select the best causal reason that sets the decision making algorithm's chosen example apart from other considered examples. Finally, we select a contrasting example that best illustrates the causal reason.**

I. INTRODUCTION

In-situ data collection is important for scientific sampling applications that are time-consuming, dangerous, or hard to access. While there are different algorithms that can help scientists optimize robot paths [1–7], there is a gap between the state-of-the-art and state-of-practice in decision making algorithms. One reason these algorithms may not be used in practice is because users prefer using systems that act in an expected manner they can trust [8]. We propose using explanations to describe the robot's actions. State-of-the-art eXplainable Artificial Intelligence (XAI) focuses on providing explanations that provide additional information and transparency about the structure and function of the decision-maker [9–11], while research from the social sciences suggests a more effective way to provide an explanation is through a contrastive example [12]. Our goal is to help users understand the robot's decisions.

While several explainability methods for robotics have appeared in the literature, they either:

- Use noncontrastive explanations [13,14], i.e., explanations that directly try to explain a decision without relating it to an alternative example.
- Rely on interpretable models, such as behavior trees [14,15], that limit the decision making algorithm to simple models that a human can understand throughout.
- Provide the explanations in terms of trade-offs between high-level objectives [16–18], or only show contrastive examples, which do not provide as detailed causal reasons that CoFeX enables [19,20].

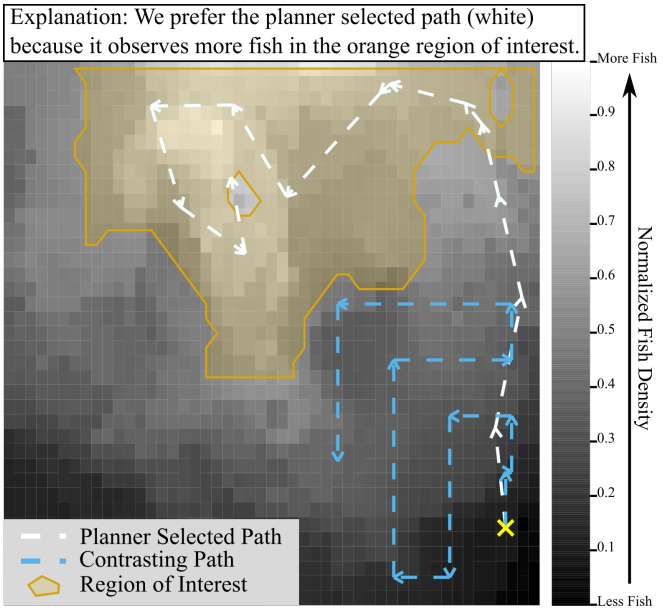

Fig. 1: A contrastive explanation which shows the causal reason selected by CoFeX, highlighted in orange, and the illustrative contrasting path (blue). The causal reason is selected from a set of semantic features that best shows what differentiates the selected path apart from other considered paths. The contrasting path is then selected to best illustrate the chosen causal reason.

Using insights from the social sciences, we propose the CoFeX method, which removes these prior limitations. We use contrastive explanations to illustrate causal reasons, i.e., 'Why A instead of B' rather than 'Why A'. We select contrasting explanations because previous social science research has shown they are easier for humans to understand [12]. We also choose post-hoc explanations, ones where the explanations are generated after the decision making algorithm has made a selection, to allow complex decision making algorithms to be used. In this work, we define a set of semantic features in the environment as the 'words' in our shared communication between the human and robot. Finally, we provide a method to select a single most important causal reason to show to the user which is informed by previous feature-importance methods [21]. This aligns with research that shows humans typically accept a single best cause that differentiates the event from other events, instead of many causes in the full causal attribution [12].

We present three novel main contributions in this paper:

This research was funded in part by NSF grants IIS-1845227.
[1]Collaborative Robotics and Intelligent Systems (CoRIS) Institute, Oregon State University, Corvallis OR, United States; {rankini, geoff.hollinger}@oregonstate.edu. [2]Woods Hole Oceanographic Institution, Woods Hole, MA, United States; smccammon@whoi.edu. [3]Transnetyx Inc., Cordova, TN, United States; tsomers@transnetyx.com.

1) A post-hoc (explanations are generated after the decision making algorithm is run) explainability framework for robotic decision making that is agnostic to the particular algorithms being used.
2) A method to select the most likely casual reason for a decision given the relative importance of features.
3) A method to select a contrastive example that best illustrates the causal reason for the decision.

We show preliminary validation of our method by running a user study, $N = 50$, that shows we improve the user's understandability of the robot decision making on a scientific data collection task.

II. PROBLEM DEFINITION

The problem we seek to address is providing explanations for robotic decision making algorithms that are both highly informative to a human user and not reliant on the interpretability of the internal decision making representation. We define a *Contrastive Explanation* as an explanation that illustrates the causal reasons as explaining A instead of B, rather than just A [12]. *Post-hoc* explanations are defined as being generated after the decision making process. Post-hoc explanations contrast with Interpretable Models that can be inspected at every point. *Interpretable Models* [22] are decision making processes that are understandable to humans at all layers of their computation, but require specialized and simplified decision-making models, such as Behavior Trees [14]. *Feature-based Explanations* are ones where features are used to describe the causal reason for the decisions. These methods use local reasons that describe the causal justification in terms of features rather than high-level objectives.

Using Miller's [12] recommendations, explanations should be given in the form of counterfactual cases. Consequently, we use contrastive examples from alternative paths the decision-making algorithm considered, $p \in P_{considered}$, as the mechanism for explanation. Explanations also need a causal reason. We define these causal reasons as a set of semantic features, $f \in F$. We employ these semantic features as the shared language between the robot and user. For example, in the problem domain tested in the user study, we utilize the information collected in regions of interest as the semantic features. Miller [12] notes that humans typically prefer only the single most informative cause, rather than the full list of causes for an explanation. Thus, we simplify our definition of explanations to the combination of a single cause and single illustrative contrastive example,

$$e = (f, p) \in E = (F \times P_{\text{considered}}). \tag{1}$$

Since we are using a post-hoc method, the explanations are generated after the decision making process has selected its optimal decision $p^* \in P_{\text{considered}}$. To find the most informative cause, f_e, we utilize the idea of selecting for anomalies that differentiate p^* from the rest of considered possibilities $P_{\text{considered}}$ [12]. In order to calculate differences in the features, we require some measure of the relative importance of features. We define the importance as $\varphi_{p,f}$, for each example $p \in P_{\text{considered}}$, and for each semantic feature, $f \in F$. Next, we define the anomalous feature as the feature importance value with the largest possible value difference from the median importance of features:

$$f_e = \underset{f \in F}{\operatorname{argmax}} \left(\varphi_{p^*,f} - \underset{p}{\operatorname{median}}\, \varphi_{p,f} \right). \tag{2}$$

Now, we define the contrasting example as the one that best illustrates the cause. We define this illustrative example as the one most similar to p^* in the feature importance value feature space, while being as different as possible in the selected feature,

$$p_e = \underset{p \in P_{\text{considered}}}{\operatorname{argmax}} \; (\varphi_{p^*,f_e} - \varphi_{p,f_e}) - \text{similarity}(p^*, p). \tag{3}$$

Once the causal feature and illustrative paths are found, they are combined to form $e_{p^*} = (f_e, p_e)$, our mathematical formulation for an explanation of the optimal decision p^*.

III. METHODOLOGY

We break our CoFeX framework into a set of discrete steps, as shown in Figure 2. First, a small set of examples considered by the decision making algorithm are selected, $P_{\text{considered}}$. While our method works for any arbitrary decision making algorithm, we assume the algorithm can output multiple alternative examples. Next, we extract semantic features from every example in $P_{\text{considered}}$. Then, we learn a random forest of regression trees using the semantic features as input with the utility score of the example as the output. The trees are fed into the TreeExplainer algorithm [21] to extract the relative contribution of each feature using the Shapley values [23]. Next, we select f_e as the semantic feature for the causal explanation using Eq. 2. After f_e is selected, we pick p_e as the example that best illustrates the chosen causal feature as defined in Eq. 3. Finally, text and visual explanations are generated using templates of explanations given the type of feature selected. We use regions of interest as shown in Figure 1 as the main semantic features in our examples.

A. Shapley Values from TreeExplainer

The decision making algorithm generates a set of alternative examples, $P_{\text{considered}}$, and has semantic features extracted for all of the examples. We want to find the relative importance of each semantic feature to the utility score of the example. Shapley values are utilized, which are the average relative marginal contribution of the feature to the resulting utility, as a measure for this relative importance. However, calculating the Shapley values directly by enumeration of coalitions is infeasible for most domains. Instead, we approximate the Shapley values using the TreeExplainer algorithm developed in Lundberg et al. [21].

Before the TreeExplainer algorithm can be used, a decision tree or ensemble of decision trees must be constructed. Let each example $p \in P_{\text{considered}}$ have a utility $v(p)$. We then construct a regression decision tree that takes the semantic features of the paths as input and outputs $v(p)$ for each path $p \in P_{\text{considered}}$. This method allows explanations to cover a wide variety of semantic feature types, such as real-valued,

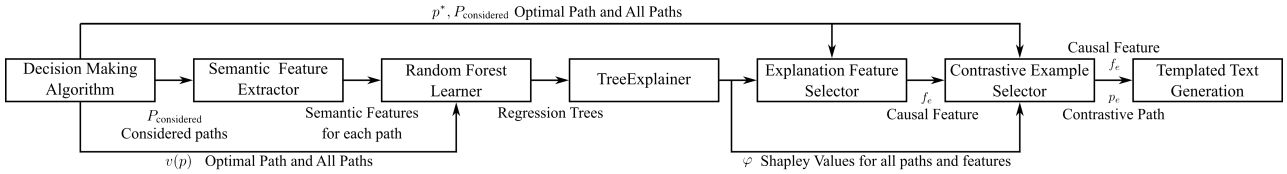

Fig. 2: Block diagram overview of our CoFeX method. After the decision making algorithm generates a set of examples to consider, we extract semantic features as our shared language between human and robot. Next, we generate a model tree of the decision making process and use the TreeExplainer algorithm to find approximate Shapley values of all paths and features. We then determine the semantic feature that best describes the causal reason the selected example was chosen for. Finally, we find a contrastive example that best illustrates the causal feature and generate the templated text explanations.

integer, and categorical features. Once the decision tree has been learned, we pass it to the TreeExplainer algorithm.

B. Explanation Feature and Constrastive Example Selection

Once the Shapley values for all examples $p \in P_{\text{considered}}$, and all features $f \in F$ has been found by the TreeExplainer algorithm, we want to find the best feature to use as the causal in the explanation. We directly solve Eq. 2 by iterating through all features to find the one that maximizes the equation. We use the found feature f_e as the causal reason in the explanation and pass it to the contrastive example selection. One causal reason, f_e, is shown as the region of interest (ROI) in Figure 1.

To select the contrastive example that illustrates the causal reason, we solve Eq. 3 by maximizing the difference in importance of the given feature f_e while minimizing the overall difference between the optimal decision p^* and the illustrative example p_e. We define the similarity between the two paths as the L2-norm in the feature space between the two examples. We normalize the similarity and feature space and directly solve Eq. 3 by iteration. An example of a selected illustrated path is the contrastive path shown in Figure 1.

C. Textual Explanation

Finally, once the full mathematical explanation $e_{p^*} = (f_e, p_e)$ has been found, we generate a text version of the explanation. We use a set of templates to generate the textual explanation, such as "We prefer the {better} path because it {reason}.", where the reason can be templated for different types of semantic features. In the case of the ROI features used in the user study one reason template is "observed more fish in the {feature}." We combine these text explanations with visual explanations of the selected semantic feature and illustrative path. This combination of visual and textual explanation of the selected causal feature and example are used as the full explanation to show to the user as seen in Figure 1.

IV. STUDY DESIGN

The goal of the user study is to show how using CoFeX affects the the user's understanding and confidence in their understanding of a robot planner's decisions. We want their confidence to be proportional to their actual understanding of the planner's decisions. This leads us to propose the following hypotheses:

- **H1:** CoFeX's explanations improve user's understanding of the robot planner's decisions.
- **H2:** CoFeX's explanations improve confidence of the user's understanding of the robot planner's decisions.

We also hypothesized that the contrastive example indirectly encapsulated most of the information gained from the textual explanation and feature. This led us to propose our final hypothesis:

- **H3:** Contrastive explanations without features or text yield similar results to CoFeX.

A. Problem Domain

We task the participants with selecting which of two informative path planning [24] paths with the same starting point is expected to observe more fish. The planner optimizes the path over real-world maps of approximate fish density in the Mid-Atlantic Coast [25]. We show the participants the two informative paths and the fish density map as well as any additional information provided by the method being tested. This task allows us to use a direct quantitative measures about user-understanding rather than qualitative or self-reported results.

B. Semantic Features

Once paths have been generated for the user study, we need a set of semantic features that are used as the shared language between the robot and the human user. In our study domain we used the information collected in ROIs and two path-based metrics as our semantic features.

The ROI semantic features are turned into a single number representing the expected number of fish observed in each ROI for each path. We use two types of ROIs, the first being a systematically generated grid over the environment. The second type of ROI is automatically generated hotspots, using the method described in [26]. We use two path metrics as semantic features in our study. The first metric keeps a count of the number of times a path crosses itself. The second metric is the general path exploration measured by taking the path integral of the distance the path is from itself.

C. Study Measures

For the study, we asked the participants to select which of two paths they think will observe more fish and to rate their confidence in their decision. This leads to the following study measures of accuracy, confidence, and reliable confidence.

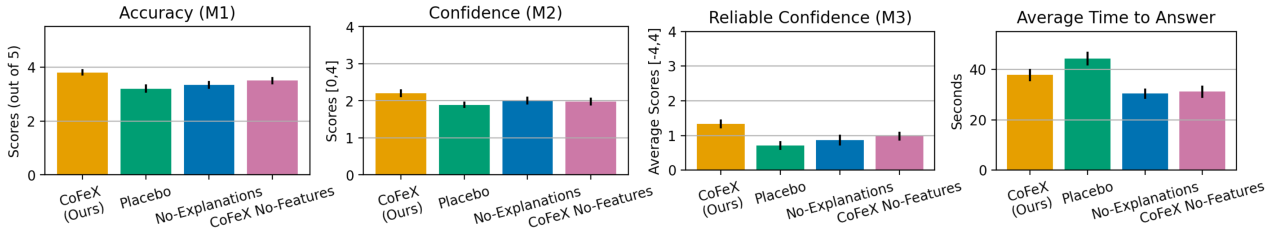

Fig. 3: From top-left to bottom-right: accuracy of the participant selecting the correct path, average confidence of participants decisions, average reliable confidence of participants, and average time participants took to answer the prompt. Our CoFeX method outperforms the baseline Placebo and No-Explanations for both accuracy and reliable confidence. In the confidence case, our CoFeX method performs only slightly better than the No-Explanations or No-Feature methods. In the time to answer the prompt case, both methods with text explanations (CoFeX, Placebo) take longer than the non-text versions (No-Explanations, CoFeX with No-Features). Additionally, the Placebo that provides poor information takes participants significantly longer to understand.

1) Accuracy (M1): Since the robot planner knows which of the two paths is expected to collect more fish, we use the participants' accuracy in selecting the higher-valued path as a direct quantitative measure.

2) Confidence (M2): Next, we use the average confidence of participants' decisions as a self-reported measure of their own confidence in their understanding of the robot planner given the information provided to them. We use a 5-point Likert scale to measure the confidence.

3) Reliable Confidence (M3): However, pure confidence does not account for appropriate trust [8,27]. A user may trust even when they do not know the correct solution. To account for appropriate trust, we use the reliable confidence measure proposed by [17]. This measure takes the confidence score and maps it to $[0, 4]$, if the participant has selected the correct path, we multiply the confidence by $+1$, otherwise we multiple it by -1. This maps the reliable confidence between $[-4, 4]$. We then take the average of these scores across the five examples shown for each study section.

4) Average Time To Answer (M4): Finally, we measured the average time it took participants to respond about which path they thought was better.

D. Comparison Methods

For our comparison methods, we chose a No-Explanation method, a Placebo method, and a No-Feature contrastive explanation method. No-explanations and Placebo are used as baselines, and the No-Features method is used to provide insights into whether features are useful to the explanations.

1) Explanations (CoFeX): The first method we examine in our study is our proposed CoFeX method described in Section III. We show two paths: the first is the best path, p^*, and the second path shown is the contrasting example p_e, with the order of the paths randomized. In order to make the problem non-trivial for the participant to solve, explanations are shown for both paths.

2) No-Explanations: We use No-Explanations as the first baseline to compare against. The comparison selects and shows two paths to the users, p^* and p_r, in a random order. The first one shown is the best path p^*. The second path is randomly selected from $p_r \in P_{\text{considered}}$ such that $v(p_r) \leq v(p_e)$, where p_e is the contrast selected by CoFeX.

3) Placebo: In the Placebo case, we once again show the best path in $P_{\text{considered}}$, p^*, and the randomly selected path, p_r described above. However, in this case we select semantic features to explain both p^* and p_r. These two semantic features are uniformly randomly selected from F.

4) No-Features Contrastive Explanation Without Text: We hypothesized that the contrasting path contained all of the relevant information without needing to directly show the feature or text, H3. To test this, we added a comparison method that performs all of the same steps discussed in Section III. However, this method does not show the selected feature f_e to the user or the related explanation text and only shows the paths p^* and p_e.

V. RESULTS

We show preliminary results for $N = 50$ convenience participants. These results for each of the study measures are shown in Figure 3 with mean and standard error of the mean shown. Since these are preliminary results, we do not perform statistical tests. The Accuracy (M1) results show that **H1**: Explanations improve user-understanding, is likely supported since our proposed method performs better than either the Placebo or No-explanations method. While the Confidence (M2) scores are inconclusive, Reliable Confidence (M3) is better for explanations than both the Placebo and No-explanations methods. This implies **H2**: CoFeX's explanations improve confidence of the user's understanding of the robot planner's decisions (M2, M3), is trending towards being only true for confidence that is weighted by if the user is correct, but does not raise overall confidence in their decision whether a path is better or not. For **H3**: contrastive explanations without features or text yield similar results to CoFeX, is not supported as CoFeX No-Features had fairly different performance than the full CoFeX method in all of the metrics. Finally, participants took longer to respond when given explanations versus without. This implies a higher cognitive load on users when given explanations than without. This could be a positive by allowing users to more actively consider the decision making process which could account for better performance in M1 and M3. Conversely, the additional load could also be an annoyance. Future work could analyze the cognitive load and consider when full explanations are required versus just showing the decision.

REFERENCES

[1] G. Best, O. M. Cliff, T. Patten, R. R. Mettu, and R. Fitch, "Dec-MCTS: Decentralized planning for multi-robot active perception," *The International Journal of Robotics Research*, vol. 38, no. 2-3, pp. 316–337, 2019.

[2] J. Binney and G. S. Sukhatme, "Branch and bound for informative path planning," in *Proc. IEEE International Conference on Robotics and Automation*, 2012, pp. 2147–2154.

[3] G. A. Hollinger and G. S. Sukhatme, "Sampling-based robotic information gathering algorithms," *The International Journal of Robotics Research*, vol. 33, no. 9, pp. 1271–1287, 2014.

[4] D. Jones, G. A. Hollinger, M. J. Kuhlman, D. A. Sofge, and S. K. Gupta, "Stochastic optimization for autonomous vehicles with limited control authority," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 2395–2401.

[5] S. Manjanna, M. A. Hsieh, and G. Dudek, "Scalable multirobot planning for informed spatial sampling," *Autonomous Robots*, vol. 46, no. 7, pp. 817–829, 2022.

[6] S. McCammon, D. Jones, and G. A. Hollinger, "Topology-aware self-organizing maps for robotic information gathering," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, Nevada (Virtual)*, 2020, pp. 1717–1724.

[7] A. Singh, A. Krause, C. Guestrin, and W. J. Kaiser, "Efficient informative sensing using multiple robots," *Journal of Artificial Intelligence Research*, vol. 34, pp. 707–755, 2009.

[8] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human factors*, vol. 46, no. 1, pp. 50–80, 2004.

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

[10] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should i trust you?" explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[11] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller, "Explaining deep neural networks and beyond: A review of methods and applications," *Proceedings of the IEEE*, vol. 109, no. 3, pp. 247–278, 2021.

[12] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial intelligence*, vol. 267, pp. 1–38, 2019.

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

[14] Z. Han, D. Giger, J. Allspaw, M. S. Lee, H. Admoni, and H. A. Yanco, "Building the foundation of robot explanation generation using behavior trees," *Transactions on Human-Robot Interaction*, vol. 37, no. 4, 2020.

[15] Z. Han and H. A. Yanco, "Communicating missing causal information to explain a robot's past behavior," *ACM Transactions on Human-Robot Interaction*, 2022.

[16] R. Sukkerd, R. Simmons, and D. Garlan, "Toward explainable multi-objective probabilistic planning," in *Proc. IEEE/ACM 4th International Workshop on Software Engineering for Smart Cyber-Physical Systems (SEsCPS)*, 2018, pp. 19–25.

[17] ——, "Tradeoff-focused contrastive explanation for MDP planning," in *Proc. 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2020, pp. 1041–1048.

[18] R. Wohlrab, J. Cámara, D. Garlan, and B. Schmerl, "Explaining quality attribute tradeoffs in automated planning for self-adaptive systems," *Journal of Systems and Software*, p. 111538, 2022.

[19] M. S. Lee, H. Admoni, and R. Simmons, "Machine teaching for human inverse reinforcement learning," *Frontiers in Robotics and AI*, vol. 8, p. 188, 2021.

[20] ——, "Reasoning about counterfactuals to improve human inverse reinforcement learning," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 9140–9147.

[21] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable AI for trees," *Nature machine intelligence*, vol. 2, no. 1, pp. 56–67, 2020.

[22] S. Mohseni, N. Zarei, and E. D. Ragan, "A multidisciplinary survey and framework for design and evaluation of explainable AI systems," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 11, no. 3-4, pp. 1–45, 2021.

[23] L. S. Shapley, "Quota solutions of n-person games," *Contributions to the Theory of Games*, vol. 2, pp. 343–359, 1953.

[24] A. Singh, A. Krause, C. Guestrin, W. J. Kaiser, and M. A. Batalin, "Efficient planning of informative paths for multiple robots," in *Proc. International Joint Conference on Artificial Intelligence*, vol. 7, 2007, pp. 2204–2211.

[25] C. Curtice, J. Cleary, E. Shumchenia, and P. Halpin, "Marine-life data and analysis team (MDAT) technical report on the methods and development of marine-life data to support regional ocean planning and management." Prepared on behalf of the Marine-life Data and Analysis Team (MDAT), Tech. Rep., 2019.

[26] S. McCammon and G. A. Hollinger, "Topological hotspot identification for informative path planning with a marine robot," in *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 4865–4872.

[27] H. Vasconcelos, M. Jörke, M. Grunde-McLaughlin, T. Gerstenberg, M. Bernstein, and R. Krishna, "Explanations can reduce over-reliance on AI systems during decision-making," *arXiv preprint arXiv:2212.06823*, 2022.