Autonomous Generation of Robust and Focused Explanations for Robot Policies

Oliver Struckmeier, Mattia Racca and Ville Kyrki

***Abstract*— Transparency of robot behaviors increases efficiency and quality of interactions with humans. To increase transparency of robot policies, we propose a method for generating robust and focused explanations that express why a robot chose a particular action. The proposed method examines the policy based on the state space in which an action was chosen and describes it in natural language. The method can generate focused explanations by leaving out irrelevant state dimensions, and avoid explanations that are sensitive to small perturbations or have ambiguous natural language concepts. Furthermore, the method is agnostic to the policy representation and only requires the policy to be evaluated at different samples of the state space. We conducted a user study with 18 participants to investigate the usability of the proposed method compared to a comprehensive method that generates explanations using all dimensions. We observed how focused explanations helped the subjects more reliably detect the irrelevant dimensions of the explained system and how preferences regarding explanation styles and their expected characteristics greatly differ among the participants.**

I. Introduction

When interacting with an artificial agent, humans tend to apply their concepts of social interaction and communication to it. Attributing familiar properties to an agent such as a robot may make it more trustworthy, explainable and predictable [1]. The assumptions humans make about the intentions, functions and purpose of robots are called the *mental model* [2], [3]. Having a precise mental model allows the user to predict the behavior of the robot, increasing the quality of the interaction as well as the trust that the user puts into the robot [4]–[6].

To support the formation of precise mental models, robots need to provide information about their purposes, operation states, and capabilities. Verbal explanations of policies have been shown to be an effective strategy to provide this transparency, affecting trust [6]–[9] and performance in collaborative scenarios [10], [11]. However, explaining complex policies can be a tedious and difficult job even for an expert annotator.

Explanations can be generated *automatically* by leveraging information about a particular type of decision model, such as a Markov decision process (MDP), either directly [8] or by learning such a model from annotated source code [12]. In contrast, we address model-agnostic explanations where the underlying system is treated as a black-box and explanations are generated by sampling the policy function [13], [14].

*This work was supported by the Strategic Research Council at Academy of Finland, decision 314180. The authors are with the School of Electrical Engineering, Aalto University, Finland (name.surname@aalto.fi).

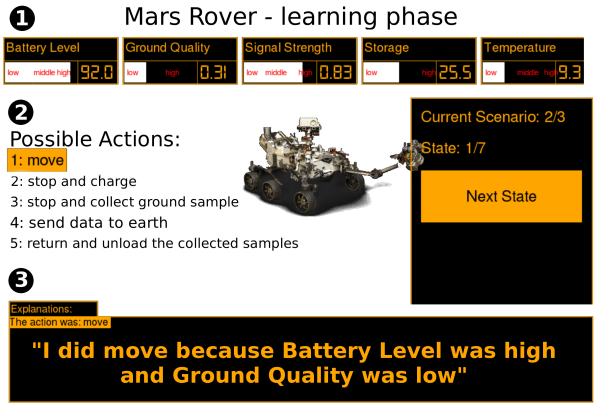


Fig. 1: Explanation scenario: a mars rover presents to the user (1) its current state through bar indicators with overlaid descriptors, (2) the action selected by its policy and (3) a focused explanation for the state-action pair.

However, existing model-agnostic approaches have been shown to suffer from lack of stability such that small changes in the policy can lead to different explanations [15].

To robustly explain policies, we propose a method that generates focused model-agnostic explanations for the actions of an intelligent system. Inspired by how human experts explain an unknown system to other humans [7], the generated explanations focus on most important variables that led to the explained action. Furthermore, the method avoids explanations that are not stable, i.e. ones which are not capable of consistently describing the state with given concepts, and those that are not relevant such that they do not not represent causal relationships.

To investigate the impact of focused explanations on user understanding of policies, we conducted a user study comparing focused and comprehensive explanations in a simulated space rover scenario. Results indicate that the proposed focused explanations allow users to better determine dimensions of the state space that do not play a role in the robot's policy and how usefulness of explanations varies among users, based on personal learning preferences.

II. Related Work

We next discuss the nature of explanations and how their structure can affect task performance and human-robot interaction quality. We then review research on autonomous explanation generation, highlighting the main challenges faced by these systems.

A. Nature of explanations

Explanations play an important role to make robots *interpretable*, defined as the degree to which a human can understand the cause of a decision [16]. Explanations should be structured such that they support the mutual understanding of mixed human-robot teams [10], [11] while providing a satisfying user experience [9], [17].

When inquiring about an unknown intelligent system, people's questions address two main issues [18]: understanding (i) *what* action the system took, and (ii) *why* did it take that action, that is, what policy is the system following. In particular, *why* explanations describing the policy have been found to lead to better understanding and higher trust [9]. In this work, we address both issues for policies with continuous state space and a discrete action set. Explaining the action in this case is simply achieved by verbalizing the action. In contrast, providing *why* explanations requires the explanation generation to be able to express the state of the system (input of the policy function) in natural language, which is not trivial for continuous action spaces.

Kulesza *et al.* [17] argue that *soundness* (truthfulness) and *completeness* (coverage) are important dimensions of an explanation. They showed with a user study how users value complete and detailed explanations and that oversimplified explanations can negatively affect the users' mental models. In contrast, Elizalde *et al.* [7] found evidence that humans tend to explain a system's policy by concentrating on the most important variables, and that, with such explanations, their subjects showed a better understanding of the system. These seemingly conflicting findings appear to stem from the fact that explanations that are both complete and sound can become too long and convoluted for complex systems, countering their benefits. In this work, we sacrifice completeness and study if *local explanations*, i.e. explanations that concentrate only on the action taken in a particular state, are able to decrease the complexity of explanations by focusing on the locally most relevant variables.

B. Automatically generating explanations

Automatic generation of explanations typically requires addressing three choices: which explanation template to adopt, which explanation to use if multiple are available, and how to verbally describe individually states and actions.

The general structure of explanation templates proposed in the literature [8], [12] describes verbally one or more dimensions of state as well as the corresponding action, following the general prototype "The system performs ACTION when DIMENSION has VALUE (and/or DIMENSION2 has VALUE2 and/or...)". The approaches differ in that the explanation can refer to only current context ("The system performed ACTION because...") [8] or describe all contexts where a particular action is chosen ("ACTION is performed when...") [12]. The former approach leads to concise and easier to understand explanations while the latter one provides more complete ones. In line with the concept of local explanations, we follow the former approach.

When multiple explanations are available, the explainer has to select one based on some criteria. When the policy bases its decisions on a multidimensional state space, this selection can be about which and how many dimensions to include in the explanation. Including only the dimensions that matter can both decrease the complexity of explanations and make them more general. One approach is to determine a minimal subset of semantic logical predicates that defines the set of states resulting in a particular action [12]. An alternative is to leverage a particular characteristic of the decision model by turning a MDP policy into a factored MDP to identify the variable that has greatest effect on utility [8].

Ribeiro *et al.* [13] advocate for model-agnostic approaches that do not assume any particular structure of the policy. To achieve that, Hayes [12] proposed to transform the policy into a MDP using expert annotations in source code. In contrast, we propose to use sampling of the policy function in order to determine the relevance of particular dimensions, which eliminates the need for annotations. This can be seen as analogous to the LIME method [13], which produces non-verbal explanations for image and text classification. A recent study [15] investigated automatic explanation methods for classification, including LIME [13], regarding their robustness. Their results suggest that sampling may provide unstable explanations such that small changes in the state can lead to different explanations. This is one of the primary issues addressed by our contribution.

States or sets of states need to be described verbally in order to include them in explanations. When the state space is limited and discrete, it may be possible to use it directly [8]. With larger and continuous state spaces, predicates can be created for semantically meaningful subsets of states [12]. We also follow this approach, but instead of using binary classifiers as in [12], we use probabilistic classifiers trained on expert annotations. This allows us to address the problem of labeling ambiguity.

To address these challenges, we propose a method capable of autonomously generating *model-agnostic* and *robust* explanations for a robot's policy based on the *most relevant dimensions* of the state space. The proposed method quantifies how relevant each dimension of the state space is for the explanation and avoids explanations where (i) the verbal concept used to describe the state is not locally stable or (ii) the policy is not locally stable.

III. Explanation Generation

We are interested in explaining a policy defined on a continuous space with discrete actions. Formally, let $\mathbf{X}$, a close subset of $\mathbb{R}^N$, be the state space and $A = \{a_1, a_2, \ldots, a_M\}$ the discrete set of actions. We refer to the dimension i of a state vector $\mathbf{x} \in \mathbf{X}$ as x_i. A stochastic policy is a function $\pi(\mathbf{x}) : \mathbb{R}^N \to \mathbb{S}^M$, with $\pi_j(\mathbf{x})$ being the probability of taking action a_j in $\mathbf{x}$ and $\mathbb{S}^M$ being the probability simplex. For a stochastic policy, we denote the most probable action as $a^\star(\mathbf{x}) = \operatorname{argmax}_j \pi_j(\mathbf{x})$. We assume the numeric representation of the state space (i.e. the meaning of its dimensions) to be comprehensible by the user. However, in

order to ease understanding and reduce cognitive load, the explanations will use a natural language labels to describe the value of each dimension.

A. Comprehensive explanation

We can explain an action by describing the current state-action pair $\langle \boldsymbol{x}, a \rangle$ with an explanation structured as

"I did $\boxed{\textit{action } a}$ *because* $\boxed{\textit{dimension } d_1}$ *was* $\boxed{\gamma_1^\star}$ *and ... and* $\boxed{\textit{dimension } d_N}$ *was* $\boxed{\gamma_N^\star}$*"*,

where a and d_i are natural language labels of the action taken and dimension of state space, and γ_i are natural language descriptors, like *"high"* or *"low"*, that describe parts of each dimension. This formulation assumes that each dimension can be described independently. We call *comprehensive explanations* those that includes all dimensions of the state space.

To determine the natural language descriptor for dimension i, we use membership functions $\epsilon_{i,j}(x_i) \in [0, 1]$ that determine how well (part of) a dimension i can be described by natural language concept j. A membership function $\epsilon_{i,j}(x_i)$ is analogous to a fuzzy membership function [19] and its output can also be interpreted as an unnormalized probability. The membership functions can be encoded manually or learned from state space samples labeled by a domain expert.

Using the membership functions and the current state $\mathbf{x}$, the best matching descriptor for dimension i can be defined as

$$\gamma_i^\star(x_i) = \underset{j}{\operatorname{argmax}}\, \epsilon_{i,j}(x_i). \tag{1}$$

B. Extracting relevant dimensions

While comprehensive explanations are capable of verbalizing the current state of the robot, explanations for a high dimensional state space would be long and thus difficult to understand. They would also be overly specific, including dimensions that do not affect action selection (lack of *relevance*). Moreover, in some parts of the state space the action selection can be unstable (lack of *policy stability*). Similarly, the chosen descriptor may be locally unstable (lack of *state describability*). Finally, we want to avoid to use descriptors that describe different actions (lack of *consistency*).

To address these issues, we propose measures for each of the four factors: local measures to address stability and describability, and global measures for relevance and consistency.

1) Local measures: Local measures are used to quantify, locally around the state $\boldsymbol{x}$, the influence of the policy $\pi(\boldsymbol{x})$ on the action selection and how well the descriptors of the given dimension i describe the state. Locality in this context means that the behavior of action selection and descriptors is analyzed in the neighborhood of $\boldsymbol{x}$. We sample V uniformly distributed states s in a hyper-sphere around $\boldsymbol{x}$ with radius r,

$$S_L = \{\boldsymbol{s} | \|\boldsymbol{s} - \boldsymbol{x}\| \leq r\}. \tag{2}$$

For each sample $\boldsymbol{s} \in S_L$, the descriptors $\gamma_i^\star(\boldsymbol{s})$ are determined and the most likely selected action $a^\star(\boldsymbol{s})$ is identified by evaluating π.

a) Stability: To quantify if $a^\star(\boldsymbol{x})$ is the most probable action in the neighborhood of $\boldsymbol{x}$, we calculate its frequency in the sample set as

$$P(a|S_L, \pi, \boldsymbol{x}) \approx \mathcal{S}_i = \frac{\sum_{\boldsymbol{s} \in S_L | a^\star(\boldsymbol{s}) = a^\star(\boldsymbol{x})} \pi_{a\star}(s)}{V}. \tag{3}$$

If small changes in state lead to a different $a^\star$, the stability is low.

b) Describability: To quantify if the strongest descriptor $\gamma_i^\star$ is consistent in the neighborhood of $\boldsymbol{x}$, we calculate its frequency taking into account the soft membership as

$$P(\gamma_i^\star | S_L, \pi, \mathcal{C}_i, \boldsymbol{x}) \approx \mathcal{D}_i = \frac{\sum_{\boldsymbol{s} \in S_L | \gamma_i^\star(\boldsymbol{s}) = \gamma_i^\star(\boldsymbol{x})} \max_j \epsilon_{i,j}(\boldsymbol{s})}{V}. \tag{4}$$

If small changes in state lead to a different strongest descriptor, the describability is low.

2) Global measures: Global measures quantify properties over the entire state space. To evaluate them, we generate W dimension-specific, evenly spaced samples S along each dimension i by varying x_i and fixing the other dimensions $j \neq i$. The dimension specific global samples S_i are computed as

$$S_i = \{\boldsymbol{s} | \min x_i \leq s_i \leq \max x_i, s_j = x_j \forall j \neq i\}. \tag{5}$$

As for the local samples, $a^\star$ and $\gamma_i^\star$ are determined for each sample.

a) Consistency: To quantify if the policy is consistent with the descriptor (states described by the same descriptor lead to the same action), we calculate the frequency of the chosen action in the weighted support of the descriptor

$$P(\gamma_i^\star | S_i, \pi, \boldsymbol{x}) \approx \mathcal{C}_i = \frac{\sum_{\boldsymbol{s} \in S_i | \gamma_i^\star(\boldsymbol{s}) = \gamma_i^\star(\boldsymbol{x}) \wedge a^\star(\boldsymbol{s}) = a^\star(\boldsymbol{x})} \max_j \epsilon_{i,j}(\boldsymbol{s})}{\sum_{\boldsymbol{s} \in S_i | \gamma_i^\star(\boldsymbol{s}) = \gamma_i^\star(\boldsymbol{x})} \max_j \epsilon_{i,j}(\boldsymbol{s})}. \tag{6}$$

If states described by the same descriptor lead to multiple different actions, consistency is low.

b) Relevance: To quantify if dimension i affects the choice of action, we calculate the entropy of actions over the dimension,

$$\mathcal{R}_i = H(a|S_i, \pi), \tag{7}$$

with the action probabilities approximated using samples as

$$P(a|S_i, \pi) \approx \frac{\sum_{\boldsymbol{s} \in S_i | a^\star(\boldsymbol{s}) = a} 1}{W}. \tag{8}$$

If all samples along a particular dimension lead to the same action, i.e. changes along the dimension do not impact the policy's choice, the relevance of that dimension is low.

While allowing to explain black-box models, sampling is however affected by the curse of dimensionality and, as the number of dimensions rises, more advance sampling schemes are required. In this work, the sampling parameters r, V and W were manually tuned, although heuristics could be devised based on e.g. the number and location of dimension descriptors.

C. Generating Explanations

To choose which dimensions to include in the explanation, we first determine if a dimension is suitable by comparing all dimension specific measures $\mathcal{S}_i, \mathcal{D}_i, \mathcal{C}_i$ and $\mathcal{R}_i$ to related thresholds t_S, t_D, t_C, t_R. If all four measures exceed their thresholds, the given dimension can be used in an explanation. We selected these thresholds based on the desired probability to satisfy the corresponding measure. The effects of the thresholds on the explanations are described in Sec. IV-A.

For each dimension, we combine these measures into $\mathcal{Q}_i$, defined as their product $\mathcal{Q}_i = \mathcal{S}_i \cdot \mathcal{D}_i \cdot \mathcal{C}_i \cdot \mathcal{R}_i$.

An explanation is constructed by sorting the dimensions by their value of $\mathcal{Q}_i$. Let $\boldsymbol{q}$ be a vector of dimension indexes ranked, with $q(1)$ being the index of the dimension with the highest $\mathcal{Q}$. A focused explanation is constructed with the K best dimensions, based on this template

"I did $\boxed{\textit{action } a}$ *because* $\boxed{\textit{dimension } d_{q(1)}}$ *was* $\boxed{\gamma^*_{q(1)}}$ *and ... and* $\boxed{\textit{dimension } d_{q(K)}}$ *was* $\boxed{\gamma^*_{q(K)}}$*".*

The parameter K can be chosen based on the nature of the state space or based on the expertize of the end-user. We experimentally chose K, leaving its automatic selection for future work.

IV. Demonstration

We will next illustrate the explanation quality measures in a synthetic example with a 2-D state space, shown in Fig. 2. The policy to be explained has two actions, with decision boundary illustrated in red (a deterministic policy is used for clarity). The first four columns of plots show the measures $\mathcal{S}, \mathcal{D}, \mathcal{C}$ and $\mathcal{R}$ while the fifth shows the $\mathcal{Q}$ measure. The upper row illustrates the measures for dimension D_0, represented on the x axis of the plot. The lower row for dimension D_1, represented on the y axis. D_0 is described using three descriptors (namely, *low*, *medium* and *high*) while D_1 has two (*slow* and *fast*). Descriptor boundaries (equal membership contours) are illustrated in blue (only for the $\mathcal{D}$, $\mathcal{C}$ and $\mathcal{Q}$ measures). Three possible values (0.6, 0.7, 0.9) for thresholds t_S, t_D, t_C, t_R are shown.

A. Individual measures

1) Stability: Fig. 2(a) and (f) show the stability measure $\mathcal{S}$. The stability measure is low close to the decision boundary. Therefore, thresholding this measure prevents explaining where small changes in state lead to different actions, with the choice of the threshold determining the size of the unexplained area. In the case of stochastic policies, the stability measure avoids explanations in situations with high action uncertainty, since the samples would be distributed across several actions resulting in low stability.

2) Describability: Fig. 2(b) and (g) show the describability measure $\mathcal{D}$. It behaves similarly to the stability in region of the state space close the descriptor boundaries. The measures for the two dimensions differ, since the descriptors are specific for each dimension. Thresholding on the describability prevents explaining where small changes in the state would lead to a different descriptor, with the threshold value again affecting the size of the unexplained area. Furthermore, measure $\mathcal{D}$ also captures the inherent uncertainty of the concept descriptors. For example, if the descriptors are learned from expert annotations, areas of state space far from any training samples will have small membership for all descriptors; this means that there are no words to describe the current state and an explanation will not be generated.

3) Consistency: Fig. 2(c) and (h) show the consistency measure $\mathcal{C}$. Thresholding the consistency prevents explanations where the descriptor used to explain describes a region of state space that leads to different actions. As an example, consider the descriptors for the vertical dimension in Fig. 2(h). The descriptor $\gamma_{1,1}$ ("D_1 is slow") describes areas that lead to the same action, showing high consistency. On the other hand, descriptor $\gamma_{1,2}$ ("D_1 is fast") leads to two different actions in some areas of the state space; this is captured by the low values of $\mathcal{C}$.

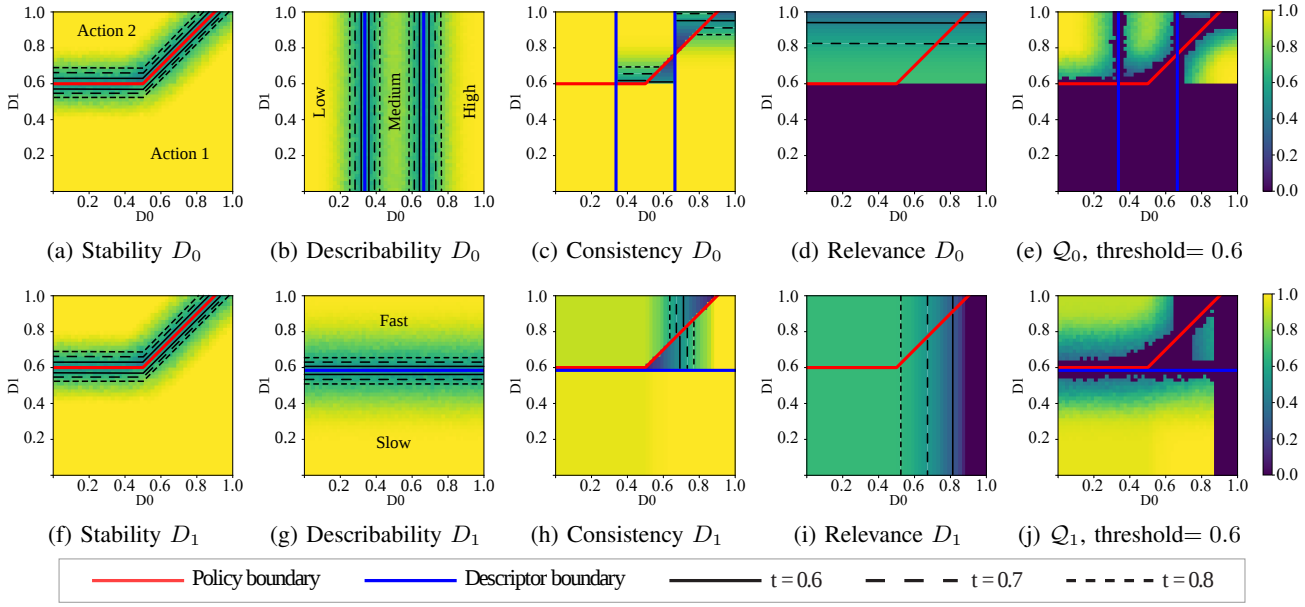
4) Relevance: Fig. 2(d) and (i) show the relevance measure $\mathcal{R}$. Thresholding the relevance prevents explanations without causality, i.e. the action choice is not affected by the dimension. For example consider Fig. 2(d) for dimension D_0. In the lower part of the plot (i.e. for low values of D_1), the action does not change by varying dimension D_0, resulting in zero relevance. On the other hand, for high values of D_1, moving along D_0 results in different actions, as captured by higher values of $\mathcal{R}$.

B. Generated explanations

Fig. 2(e) and (j) show measure $\mathcal{Q}$ after thresholding the individual measures ($t = 0.6$). In regions where none of the measures exceeds the threshold, $\mathcal{Q}$ is zeroed, excluding the particular dimension from the explanation.

Fig. 3 illustrates the resulting explanations over the state space. In the green areas only D_0 is used, in the light blue areas D_1. In the dark blue areas both dimensions are used to explain, while no explanation that can be produced in the yellow area. As an example, area marked with ① is explained by "I performed ACTION2 because D_0 was MEDIUM and D_1 was FAST.", demonstrating that both dimensions matter and that the explanation is locally stable. In contrast, the area marked with ② is explained by "I performed ACTION1 because D_1 was SLOW.", showing that changes along D_0 would not affected the action for this state. Similarly, the area marked with ③ is explained by "I performed ACTION1 because D_0 was HIGH.", omitting the irrelevant D_1.

A few observations can be made. First, no explanations are generated close to the decision boundary, illustrating that the proposed method addresses the robustness problem presented in [15]. Second, a dimension is not used in explanation if its verbal description is not locally stable, addressing the robustness of the natural language description of the state. Third, our method does not explain in states where changing each dimension one by one would not change the action. This can be seen in the lower right part of the state space, where both dimensions are deemed irrelevant.



(a) Stability D_0 (b) Describability D_0 (c) Consistency D_0 (d) Relevance D_0 (e) $\mathcal{Q}_0$, threshold= 0.6

(f) Stability D_1 (g) Describability D_1 (h) Consistency D_1 (i) Relevance D_1 (j) $\mathcal{Q}_1$, threshold= 0.6

Fig. 2: Individual measures $\mathcal{S}_i, \mathcal{D}_i, \mathcal{C}_i$ and $\mathcal{R}_i$ (a)–(d), (f)–(i) with three values of thresholds t, and their combination $\mathcal{Q}_i$ (e), (j). Best viewed in color.

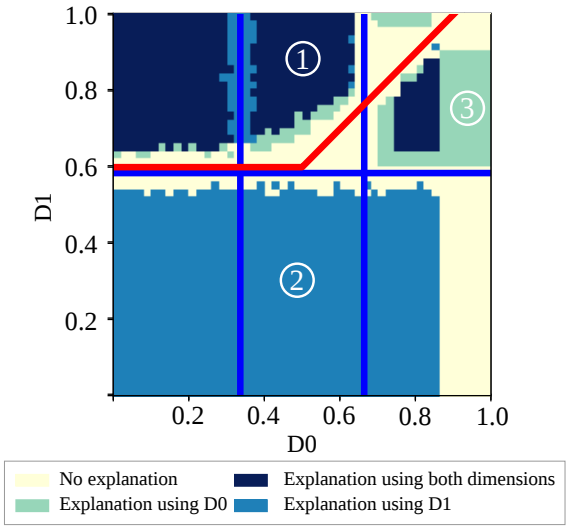


Fig. 3: Explanation generation: dimensions used to explain, for each region of the state space. Best viewed in color.

V. User Study

To explore the usability of our method, we conducted a user study comparing two explanation generation methods with a mock-up of a rover operation setting.

A. Experiment design

The participants interacted through a graphical interface with two simulated rovers, each having its own sensors (different state space $\mathbf{X}$), different actions A and policy π. We compared the proposed focused explanations (**F**) using the $K = 2$ most relevant dimensions (Sec. III-C) with comprehensive explanations (**C**) that verbalize all dimensions of the state space in random order (Sec. III-A).

The participants were tasked to learn about the rovers' policies by observing a set of state-action pairs and the corresponding explanations. Each participant was then asked to replicate the policies, by taking control of the robot's action selection. We explored the effects of different explanation methods on the subjects' understanding of the rovers' policies as well as the participants' preferences regarding the nature of the generated explanations.

Participants: Eighteen participants (age M = 26.6, SD = 5.5, female 44%) were recruited at a university campus. Participants received a movie ticket as compensation.

Conditions and Protocol: Each subject interacted with two simulated rovers (a Mars and a Moon rover) through the GUI[1] shown in Fig. 1. The order of explanation methods was counterbalanced such that each method was tested with each rover. The state space of each rover consisted of 5 continuous dimensions. Each dimension had its own label, together with up to three descriptors and their related classifiers. Each rover had 5 distinct actions, with policies encoded as decision trees. The state space, actions and policies were different for each rover. The dimensions for the Mars rover were *Battery Level, Ground Quality, Signal Strength, Storage and Temperature* and for the Moon rover *Radiation Level, Dustiness, Velocity, Elevation, Brightness*. The rovers' policies used only 4 of the 5 state dimensions (making one dimension irrelevant for the action choice), allowing us to observe if the different explanation methods helped the subjects discern between relevant and irrelevant dimensions.

First, the participant was introduced to the GUI and to

[1]Code available at `github.com/Oleffa/FocusedPolicyExplanation`

the experiment's structure. For each rover, each participant had two distinct phases: a *learning phase* and a *testing phase*. During the learning phase, the subjects observed a set of 25 state-action pairs. For each pair, an explanation was generated using either method **F** or **C**. In the testing phase, the participants acted on behalf of the policy and used their mental model of π to select the correct action for 11 states.

During the learning phase, the GUI presented participants with information about the current state, the selected action and the generated explanation. The current state was presented with (a) numerical values and (b) visually with bar indicators, overlaid with the descriptors of the given dimension. We included those state representation in addition to the explanations as basic mean to make the system transparent.

During the testing phase, the GUI provided no explanations but queried the participants for an action given a state. Participants had the option to answer *"I don't know"*. We logged the subjects' action choices and compared it against the policies acting as ground truth.

After each learning phase, participants selected which of the state dimensions were irrelevant for the policy ("*Which of the parameters mattered the LEAST for the rover to choose an action?*"). Furthermore, participants indicated what learning medium they used the most ("*What did you use the most to learn? the explanations or the bar indicators/numbers?*"). Additionally, the participants filled a questionnaire with the following 1-7 Likert scale (*1 Completely Disagree - 7 Completely Agree*) statements:

1) *I understand the behavior of the rover* [Perceived Understanding]
2) *Learning the rover's behavior was easy* [Ease of Learning]

Statements came with an optional *Why? Please explain* question. After the testing phase, the participants filled another questionnaire with the following 1-7 Likert scale (*1 Completely Disagree - 7 Completely Agree*) statement:

1) *I performed well in the test* [Perceived Performance],

followed by a *Why? Please explain* open question. The scores are shown in Fig. 4. Finally, we collected participants preferences for each explanation method, with open feedback about their concept of ideal explanation (*"If you think about a robot that can explain its actions, what would you like the explanations to be like?"*).

B. Results and Discussion

1) Policy Understanding: To assess the subjects' understanding of the rovers' policies with different explanation methods, we operationalized their understanding **U** as the percentage of correct state-action pairs provided during the testing phase. Following from [9], our hypothesis was that subjects would have a better understanding with focused explanations **F** compared to comprehensive explanations **C**. We however observed a **U** of 50.5% for method **F** and 49.0% for method **C**, with no statistically significant differences (Wilcoxon Signed-rank test, Z=0.47, p>0.05). We then

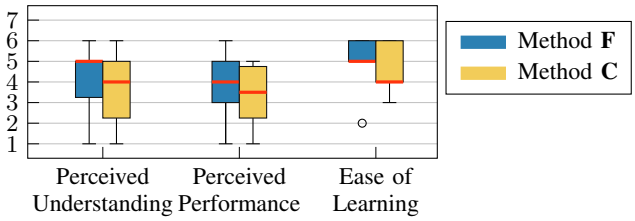


Fig. 4: Questionnaire scores (1-7 Likert scale) for each explanation method.

looked how often the subjects replied *"I don't know"* during the test for both methods. While subjects replied incorrectly on average to half of the test questions, the number of *"I don't know"* was extremely low, with only 5 such answers out of 198 for method **F** and 12 for method **C**. This may suggest how the explanations bolstered the subjects' confidence, inflating their self assessed understanding of the policies.

We therefore analyzed the scores of the three Likert statements, with the results summarized in Fig. 4. The subjects rated their perceived understanding of the policy after the learning phase with a median of 5 for method **F** and 4 for method **C**. The score slightly decreased after the test phase, with the participants self assessing their test performance with a median of 4 for method **F** and 3.5 for method **C**. In the free-form feedback, the subjects mainly commented about their ability (or difficulties) in learning the state-action pairs dictated by the rovers' policies.

Regarding the ease of learning score, both explanation method received above average scores, with method **F** having a median of 5 and method **C** of 4. Looking at the subjects' comments regarding the easiness of learning with method **F**, we can see how shorter explanations are consistently considered an easier way of learning. Furthermore, 6 subjects realized how these explanations were not only shorter but focused on the important aspects of the action selection, despite no information was given them regarding the different methods. While the questionnaire scores are slightly better for method **F**, the differences on the questionnaire scores are not statistically significant (Wilcoxon Signed-rank test, p>0.05 for all scores), indicating that the differences between explanation methods are likely small if any exist.

We then looked at the ability of the participants to single out the irrelevant dimension in the rover's policy (see Fig. 5). For method **F**, 13 out of 18 subjects correctly identified the irrelevant dimension. With method **C** instead, only 2 subjects selected the correct irrelevant dimension. The rest of them either selected the wrong dimension (10) or listed additional dimensions together with the correct one (6). Although no effect is seen on the policy understanding **U**, the proposed method **F** helped the participants to single out the irrelevant dimensions more reliably. This result is in line with what theorized and observed in [8] where a policy with a five dimensional state space was explained with similarly structured focused explanations, helping the subjects having a better

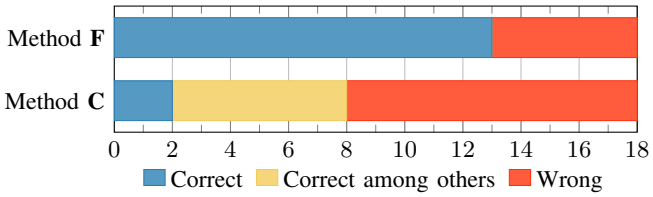


Fig. 5: Correctly identified irrelevant dimensions with focused explanations (method **F**) and comprehensive explanations (method **C**).

policy understanding. This advantage, already visible with these experiments' relatively low dimensional state space, is likely to be even more important with more complex policies, where method **C** would generate explanations cluttered with irrelevant dimensions.

2) Explanation Preferences: When asked about their preferred explanation style, 8 subjects preferred method **F**'s explanations and 8 method **C**'s explanations, with 2 subjects stating no preference. Subjects that preferred **F**'s explanations backed their choice with comments like *"Very long explanations are more difficult for me"*, *"With fewer variables to look at, it is easier to learn and then perform well in the test"* and *"It helps you see the priority of the parameters"*. Subjects that liked method **C**'s style of explanations instead commented *"It feels more human like with longer ones"*, *"I feel like I had more information and could therefore understand the choice a bit better"* and *"I prefer longer explanations but only if they are organized by importance"*. We see again how the subjects were able to describe the characteristics of each explanation method without being explicitly informed about them.

The surprising even split in the subjects' preferences between method **F** and **C** resembles the contrasting results found in the literature, with work advocating for explanation completeness [17] and other work showing the advantages of focused explanations over complete one [7]. In the light of the results, we believe the choice between explanations styles in our experiment to be influenced by personal traits of the users, like interest and curiosity towards the explained system and commitment towards the learning task. The results align with [18], suggesting that there is no perfect type of explanation that satisfies everyone's needs and at the same time improves the level of understanding. We therefore believe that further research is needed to investigate if and how explanation methods should adapt to the users.

In the free-form feedback at the end of the experiment, 4 subjects restated how their ideal explanation needs to be short, focused and clear. For another 3 subjects, the order by which the information is provided in an explanation is the most important factor, with the most relevant information being presented as early as possible. These explanation requirements are met by our method **F** thanks to the proposed quality measures that allow to rank the dimensions and threshold them to remove irrelevant dimensions.

Finally, 6 subjects stated that they expected extra information in addition to the proposed explanation template. While our explanations are capable of conveying the state-action pairing given by the policy, the subjects expected extra semantic information about the logic behind why certain actions occurred as a reaction to certain states. One participant described this concept by commenting *"I would prefer to explain the important things and say why. Not only state the important values but also state why that is important. Like* ***I do an emergency shutdown because radiation is high and brightness is low which poses risks for the rover's components****"*. If we take the explanation in Fig. 1 as example, the following information would need to be added: *"I did move because the Battery Level is high* ***and I therefore don't risk to run out of Battery*** *and the Ground Quality is low* ***and my goal is to collect high quality samples****"*. Such requirements pose serious challenges to all automatic explanations strategies. In [8], a step towards meeting these requirements was done by augmenting the explanations with extra information coming from an hand-coded knowledge base of relations between variables, components and procedures of the system. However, such detailed semantic information is rarely provided together with the policy by its designer, or might not be directly available (e.g. because encoded in the reward function in a Reinforcement Learning scenario). Therefore, generating explanations of this nature will require an augmentation of the policy itself or of the process employed to learn it, opening challenging research directions.

3) Explanations vs Bar Indicators: As described in Sec. V-A, the GUI included visual indicators (in the form of bar indicators) to illustrate the current rover state and action. The participants had therefore two information sources about the policy: the bar indicators and the explanations. When surveyed about their preferred medium, 9 subjects reported the explanations and 9 subjects selected the bar indicators.

As half of the subjects did not use explanations as their primary information source, the results presented earlier comparing the explanation methods are likely to be heavily shadowed by this effect. We therefore grouped the participants according to their preferred learning medium, denoting the group that preferred to learn from explanations **E** and the group that preferred bar indicators **I**. We then compared the groups with respect to policy understanding **U** and the questionnaire scores, with Table I summarizing the results.

While we still did not see significant differences between groups **I** and **E** regarding the policy understanding **U**, we observed higher scores of perceived understanding and ease of learning for group **E**. While there seems to be a trend towards method **F** across these measures, the preferred learning medium seems to have a larger influence, with explanations having higher scores regardless of the method used to generate them.

VI. Conclusions

In this paper, we presented a method to automatically generate explanations expressed in natural language for poli-

TABLE I: Policy understanding **U** and questionnaire's scores: descriptive statistics (median) with subjects grouped by learning medium preference (explanations **E** vs indicators **I**), for each explanation method (**F** and **C**) and combined. Test statistics and $p-$values of unpaired comparison (Mann-Whitney U test).

	F, **E**	**C**, **E**	Combined **E**	**F**, **I**	**C**, **I**	Combined **I**	Combined **E** vs Combined **I**
Policy Understanding **U**	50.5%	55.5%	53.0%	49.5%	43.5%	45.6%	U=130.5, p>0.05
Perceived Understanding	5.0	5.0	5.0	4.0	3.0	3.5	**U=75.0, p<0.01****
Perceived Performance	5.0	4.0	4.0	3.0	3.0	3.0	U=121.5, p>0.05
Ease of Learning	5.0	4.0	5.0	4.0	2.0	3.5	**U=58.5, p<0.01****

cies with discrete actions and continuous state spaces. The method evaluates explanations based on four quality measures, designed to encourage robustness and compactness. Moreover, the method does not make assumptions on the policy's nature, requiring only that the policy function can be evaluated at samples to compute the quality measures.

To evaluate the generated focused explanations, we conducted a user study, exploring how explanations influence the user's understanding of a robot's policy as well as their experience during the interaction. The results show that focused explanations help the subjects to better detect which dimensions of a system are relevant for the action selection when compared to comprehensive explanations. However, personal learning styles seemed to play a big role in the subjects' ability to benefit from different kinds of explanations and we believe further research is needed to investigate how verbal explanations fare against others means of communication such as the visual indicators used in our user study. Furthermore, we believe that the choice of explanation method affects the user's level of understanding and thus the user's goal should affect the choice. For example, supervising a robot's operation or being able to debug its operation would require different levels of understanding and might thus benefit from different explanation methods.

While our study concentrated on the amount of information included by comparing focused and comprehensive explanations, the proposed method can also detect when an explanation is robust and abstain from explaining when the proposed measures are low. This ability to not explain in certain circumstances brings two challenges. First, ways to express this incapability to explain need to be designed: *non-explanations* could take advantage of the presented measures and adjust the conveyed message to help the user understand the reason why no explanation was generated (e.g. *I did not explain because I cannot describe the current situation with my vocabulary*). Second, we need to investigate the effects of *non-explanations* on the user. While the robot's ability to avoid explaining could help the user form a truthful mental model of the policy, this ability is likely to influence the users' trust towards the robot and their perceptions of the robot's capabilities and utility.

REFERENCES

[1] J. Fink, "Anthropomorphism and human likeness in the design of robots and human-robot interaction," in *International Conference on Social Robotics*, pp. 199–208, Springer, 2012.

[2] J. de Greeff and T. Belpaeme, "Why robots should be social: Enhancing machine learning through social human-robot interaction," *PLoS one*, vol. 10, no. 9, 2015.

[3] J. B. Lyons, "Being transparent about transparency," in *AAAI Spring Symposium*, 2013.

[4] S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza, "Power to the people: The role of humans in interactive machine learning," *AI Magazine*, vol. 35, no. 4, pp. 105–120, 2014.

[5] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. De Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction," *Human Factors*, vol. 53, no. 5, 2011.

[6] A. Theodorou, R. H. Wortham, and J. J. Bryson, "Why is my robot behaving like that? designing transparency for real time inspection of autonomous robots," in *AISB Workshop on Principles of Robotics*, University of Bath, 2016.

[7] F. Elizalde, E. Sucar, M. Luque, J. Diez, and A. Reyes, "Policy explanation in factored markov decision processes," in *Proceedings of the 4th European Workshop on Probabilistic Graphical Models (PGM 2008)*, pp. 97–104, 2008.

[8] F. Elizalde, E. Sucar, J. Noguez, and A. Reyes, "Generating explanations based on markov decision processes," in *Mexican International Conference on Artificial Intelligence*, pp. 51–62, Springer, 2009.

[9] B. Y. Lim, A. K. Dey, and D. Avrahami, "Why and why not explanations improve the intelligibility of context-aware intelligent systems," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2119–2128, ACM, 2009.

[10] N. Wang, D. V. Pynadath, and S. G. Hill, "The impact of pomdp-generated explanations on trust and performance in human-robot teams," in *Proceedings of the 2016 international conference on autonomous agents & multiagent systems*, pp. 997–1005, International Foundation for Autonomous Agents and Multiagent Systems, 2016.

[11] A. St Clair and M. Mataric, "How robot verbal feedback can improve team performance in human-robot task collaborations," in *Proceedings of the tenth annual acm/ieee international conference on human-robot interaction*, pp. 213–220, ACM, 2015.

[12] B. Hayes and J. A. Shah, "Improving robot controller transparency through autonomous policy explanation," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '17, (New York, NY, USA), pp. 303–312, ACM, 2017.

[13] M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-agnostic interpretability of machine learning," in *ICML Workshop on Human Interpretability in Machine Learning*, 2016.

[14] M. T. Ribeiro, S. Singh, and C. Guestrin, "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, ACM, 2016.

[15] D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," in *ICML Workshop on Human Interpretability in Machine Learning (WHI 2018)*, 2018.

[16] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, 2018.

[17] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, and W.-K. Wong, "Too much, too little, or just right? ways explanations impact end users' mental models," in *Visual Languages and Human-Centric Computing (VL/HCC), 2013 IEEE Symposium on*, pp. 3–10, IEEE, 2013.

[18] B. Y. Lim and A. K. Dey, "Assessing demand for intelligibility in context-aware applications," in *Proceedings of the 11th international conference on Ubiquitous computing*, pp. 195–204, ACM, 2009.

[19] L. A. Zadeh, "Fuzzy sets," *Information and control*, vol. 8, no. 3, pp. 338–353, 1965.