Measuring Interpretability of Neural Policies of Robots with Disentangled Representation

Anonymous Author(s)
Affiliation
Address
email

Abstract: The advancement of robots, particularly those functioning in complex human-centric environments, relies on control solutions that are driven by machine learning. Understanding how learning-based controllers make decisions is crucial since robots are mostly safety-critical systems. This urges a formal and quantitative understanding of the explanatory factors in the interpretability of robot learning. In this paper, we aim to study interpretability of compact neural policies through the lens of disentangled representation. We leverage decision trees to obtain *factors of variation* [1] for disentanglement in robot learning; these encapsulate skills, behaviors, or strategies toward solving tasks. To assess how well networks uncover the underlying task dynamics, we introduce interpretability metrics that measure disentanglement of learned neural dynamics from a concentration of decisions, mutual information and modularity perspective. We showcase the effectiveness of the connection between interpretability and disentanglement consistently across extensive experimental analysis.

Keywords: Interpretability, Disentangled Representation, Neural Policy

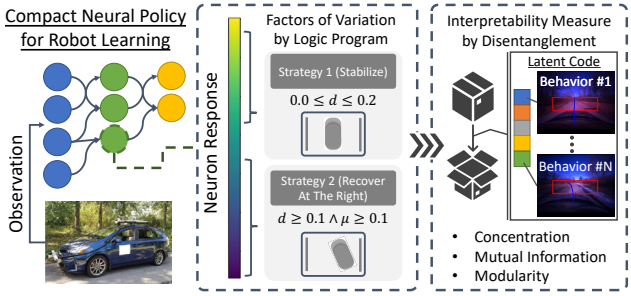
1 Introduction

Interpretability of learning-based robot control is important for safety-critical applications, because it provides human comprehension of how the system processes inputs and determines the most suitable actions to take. In general, achieving interpretability is difficult for learning-based robot control. The robot learning models make decisions without being explicitly programmed to perform the task and are often very large, thus it is impossible to synthesize and explain their processes. This lack of transparency, often referred to as the ”black box” problem, makes it hard to interpret the workings of learning-based robot control systems. Understanding why a particular decision was made or predicting how the system will behave in future scenarios remains a challenge, yet critical for physical deployments.

Through the lens of representation learning, we assume that neural networks capture a set of processes that exist in the data distribution; for robots, they manifest learned skills, behaviors, or strategies, which are critical to understand the decision-making of a policy. However, while these *factors of variation* [1] (e.g., color or shape representations) are actively studied in unsupervised learning for disentangled representation, in robot learning, they are less well-defined and pose unique challenges due to the intertwined correspondence of neural activities with emergent behaviors unknown a priori. In the present study, we aim to (i) provide a useful definition of factors of variation for policy learning, and (ii) explore how to uncover dynamics and factors of variation quantitatively as a measure of interpretability in compact neural networks for closed-loop end-to-end control applications. In this space, an entanglement correspondence to multiple neurons responsible for an emergent behavior can obstruct the interpretation of neuron response even with a small number of neurons [2, 3, 4, 5]. To this end, the disentanglement of learned representations [6, 7, 8] in compact neural networks is essential for deriving explanations and interpretations for neural policies.

Submitted to the 7th Conference on Robot Learning (CoRL 2023). Do not distribute.

We posit that each neuron should learn an abstraction (factor of variation) related to a specific strategy required for solving a subcomponent of a task. For example, in locomotion, one neuron may capture periodic gait, where the numerical value of the neuron response may be aligned with different phases of the gait cycle; another neuron may account for recovery from slipping. Besides, directly observing the neuron response along with sensory information provided as input to the policy can be extremely inefficient and tedious for identifying behaviors and interpreting decision-making.



In this work, our objective is to formulate an abstraction that represents the decision-making of a parametric policy to quantify the interpretability of learned behaviors, specifically through the lens of disentangled representation. To this end, we make the following key contributions:

- Provide a practical definition of factor of variation for robot learning by programmatically extracting decision trees from neural policies, in the form of logic program grounded by world states.
- Introduce a novel set of quantitative metrics of interpretability to assess how well policies uncover task structures and their factors of variation by measuring the disentanglement of learned neural dynamics from a concentration of decisions, mutual information, and modularity perspective.
- Experiment in a series of end-to-end policy learning tasks that (a) showcase the effectiveness of leveraging disentanglement to measure interpretability, (b) demonstrate policy behaviors extracted from neural responses, (c) unveil interpretable models through the lens of disentanglement.

2 Related Work

Compact neural networks. Compact neural networks are ideal in resource constraints situations such as RL and behavioral cloning [9, 10] and by nature easier to interpret due to a smaller number of neurons [2, 4]. A typical way to generate compact networks is by pruning [11] or compressing neural networks in end-to-end training [12]. Regularization has also been used to generate compact neural networks [13]. Compact representation of features can be learned using discriminative masking, yielding compact neural networks [14]. Neural Ordinary Differential Equations have also been used for learning compact network [15, 4]. In this work, we formally study the interpretability of compact neural policies through the lens of disentangled representation.

Interpretable neural networks. An interpretable neural network could be constructed from a physically comprehensible perspective [16, 17]. Self-explaining neural networks have been proposed in [18] to achieve robust interpretability of the model, in which explicitness, faithfulness, and stability are the main considerations. Knowledge representation is used to obtain interpretable Convolutional Neural Network [19]. Another active line of research is to study disentangled explanatory factors in learned representation [8]. A better representation should contain information in a compact and interpretable structure [1, 20]. Unlike prior works that study disentanglement based on *factors of variations* in the data such as object types, there is no notion of ground-truth factors in policy learning and thus we propose to use tree-based representation to construct a pseudo-ground-truth factor set that captures emergent behaviors of neural policies for interpretability analysis.

Interpretability in policy learning. Explainable AI has been recently extended to policy learning like reinforcement learning [21]. A common approach is to extract policies via imitation learning to interpretable models like decision tree [22]. Another line of work directly embeds the decision tree framework into the learning-based model to strike a balance between expressiveness and interpretability [23, 24, 25]. Explanation of policy behaviors can also be obtained by searching for abstract state with value function [26] or feature importance [27]. In this work, as opposed to presenting a method to interpret neural policies, we aim to offer a new perspective of disentangled representation to measure interpretability in policy learning.

Algorithm 1 Extract Abstraction via Decision Tree

Data Trajectories rollout from a compact neural policy $\mathcal{D}_{dt} = \{(o_0, s_0, a_0, z_0, \dots)_j\}_{j=1}^N$
Result Interpreters of neuron response $\{f_{\mathcal{S}}^i\}_{i\in\mathcal{I}}$
for $i \in \mathcal{I}$ **do**
  Train a decision tree T_{θ^i} from state $\{s_t\}$ to neural response $\{z_t^i\}$.
  Collect dataset $\mathcal{D}_{dp}$ with neuron response $\{z_t^i\}$ and decision paths $\{\mathcal{P}_{s_t}^i\}$.
  Train neuron response classifier $q_{\phi^i} : \mathbb{R} \to \{\mathcal{P}\}$ with $\mathcal{D}_{dp}$.
  Obtain decision path parser $r^i : \{\mathcal{P}\} \to \mathcal{L}$ by tracing out $\{\mathcal{P}_k^i\}_{k=1}^{K^i}$ in T_{θ^i}.
  Construct the mapping $f_{\mathcal{S}}^i = r^i \circ q_{\phi^i}$.
end for

3 Method

In this section, we describe how to obtain factor of variation by predicting logic programs from neuron responses that reflect the learned behavior of the policies (Section 3.1), followed by a set of quantitative measures of interpretability in the lens of disentanglement (Section 3.2).

3.1 Extracting Abstraction via Decision Tree

Our goal is to formulate a logic program that represents the decision-making of a parametric policy to serve as an abstraction of learned behaviors. First, we describe a decision process as a tuple $\{\mathcal{O}, \mathcal{S}, \mathcal{A}, P_a, h\}$, where at a time instance t, $o_t \in \mathcal{O}$ is the observation, $s_t \in \mathcal{S}$ is the state, $a_t \in \mathcal{A}$ is the action, $P_a : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the (Markovian) transition probability from current state s_t to next state s_{t+1} under action a_t, and $h : \mathcal{S} \to \mathcal{O}$ is the observation model. We define a neural policy as $\pi : \mathcal{O} \to \mathcal{A}$ and the response of neuron $i \in \mathcal{I}$ as $\{z_t^i \in \mathbb{R}\}_{i\in\mathcal{I}}$, where $\mathcal{I}$ refers to a set of neurons to be interpreted. For each neuron i, we aim to construct a mapping that infers a logic program from neuron response, $f_{\mathcal{S}}^i : \mathbb{R} \to \mathcal{L}$, where $\mathcal{L}$ is a set of logic programs grounded on environment states $\mathcal{S}$. Note that $f_{\mathcal{S}}^i$ does not take the state as an input as underlying states may be inaccessible during robot deployment. The overall procedure is summarized in Algorithm 1.

From states to neuron responses. Decision trees are non-parametric supervised learning algorithms for classification and regression. Throughout training, they develop a set of decision rules based on thresholding one or a subset of input dimensions. The relation across rules is described by a tree structure with the root node as the starting point of the decision-making process and the leaf nodes as the predictions for classification or regression. The property of decision trees to convert either categorical or continuous-valued data to a set of propositions is a natural fit for state-grounded logic programs. Given an off-the-shelf trained neural policy π, we collect a set of rollout trajectories $\mathcal{D}_{\text{dt}} = \{\tau_j\}_{j=1}^N$, where $\tau_j = (o_0, s_0, a_0, z_0, o_1, s_1, a_1, z_1 \dots)$. We first train a decision tree T_{θ^i} to predict the ith neuron response from states,

$$\theta^{i*} = \arg\min_{\theta^i} \sum_{(s_t, z_t^i)\in\mathcal{D}_{\text{dt}}} L_{\text{dt}}(\hat{z}_t^i, z_t^i), \quad \text{where } \hat{z}_t^i = T_{\theta^i}(s_t) \tag{1}$$

where L_{dt} represents the underlying classification or regression criteria. The decision tree T_{θ^i} describes relations between the neuron responses and the relevant states as logical expressions. During inference, starting from the root node, relevant state dimensions will be checked by the decision rule in the current node and directed to the relevant lower layer, finally arriving at one of the leaf nodes and providing information to regress the neuron response. Each inference traces out a route from the root node to a leaf node. This route is called a decision path. A decision path consists of a sequence of decision rules defined by nodes visited by the path, which combine to form a logic program,

$$\bigwedge_{n\in\mathcal{P}_{s_t}^i, j=g(n)} (s_t^j \le c_n) \longleftrightarrow \text{Behavior extracted from } \hat{z}_t^i \text{ via } T_{\theta^i} \tag{2}$$

where $\wedge$ is the logical AND, $\mathcal{P}_{s_t}^i$ is the decision path of the tree T_{θ^i} that takes s_t as inputs, g gives the state dimension used in the decision rule of node n (assume each node uses one feature for notation simplicity), and c_n is the threshold at node n.

From neuron responses to decision paths. So far, we recover a correspondence between the neuron response and the state-grounded program; however, this is not sufficient for deployment since the decision tree T_{θ^i} requires as input the ground-truth state and not the observable data to the policy (like o_t, z_t). To address this, we find an inverse of T_{θ^i}. Based on the inference process, there is a one-to-one correspondence between $\hat{z}_t^i$ and $\mathcal{P}_{s_t}^i$, thus z_t^i and $\mathcal{P}_{s_t}^i$. We collect another dataset $\mathcal{D}_{\text{dp}}$ based on the inference of T_{θ^i} and train a classifier q_{ϕ^i} to predict decision paths from neuron responses,

$$\phi^{i*} = \arg\min_{\phi^i} \sum_{(z_t^i, \mathcal{P}_{s_t}^i) \in \mathcal{D}_{\text{dp}}} L_{\text{dp}}(q_{\phi^i}(z_t^i), \mathcal{P}_{s_t}^i) \tag{3}$$

where L_{dp} is a classification criterion. While $\mathcal{P}_{s_t}^i$ is state-dependent, there exists a finite set of decision paths $\{\mathcal{P}_k^i\}_{k=1}^{K^i}$ given the generating decision tree. We define the mapping from the decision tree to the logic program as $r : \{\mathcal{P}\} \to \mathcal{L}$, which can be obtained by tracing out the path as described above. Overall, the desired mapping is readily constructed as $f_{\mathcal{S}}^i = r^i \circ q_{\phi^i}$.

3.2 Quantitative Measures of Interpretability

Programmatically extracting decision trees for constructing a mapping from the neuron response to a logic program offers a representation that facilitates the interpretability of compact neural policies. Furthermore, building on the computational aspect of our approach, we can quantify the interpretability of a policy with respect to several metrics through the lens of disentanglement.

A. Neuron-Response Variance. Given the finite set of decision paths $\{\mathcal{P}_k^i\}_{k=1}^{K^i}$ associated with a tree T_{θ^i} at the ith neuron, we first compute the average normalized variance of the neuron response corresponding to each decision path,

$$\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \frac{1}{K^i} \sum_{k=1}^{K^i} \text{Var}_{\substack{(s_t, z_t^i) \in \mathcal{D}_{\text{dt}} \\ t \in \{u | \mathcal{P}_{s_u}^i = \mathcal{P}_k^i\}}} \left[\frac{z_t^i}{Z^i} \right] \tag{4}$$

where Z^i is a normalization factor that depends on the range of response of the ith neuron. The set $\{u | \mathcal{P}_{s_u}^i = \mathcal{P}_k^i\}$ contains timesteps whose extracted behavior corresponds to $\mathcal{P}_k^i$. This metric captures the concentration of the neuron response that corresponds to the same strategy represented by the logic program defined by T_{θ^i}. In practice, we discretize all neuron responses to N bins, compute the index of bins to which a value belongs, divide the index by N and compute their variance.

B. Mutual Information Gap. Inspired by [8], we integrate the notion of mutual information in our framework to extend disentanglement measures for unsupervised learning to policy learning. Specifically, while previous literature assumes known ground-truth factors for disentanglement such as object types, viewing angles, etc., there is no straightforward equivalence in neural policies since the emergent behaviors or strategies are unknown a priori. To this end, we propose to leverage the decision path sets to construct pseudo-ground-truth factors $\mathcal{M}_{dp} = \bigcup_{i \in \mathcal{I}} \{\mathcal{P}_k^i\}_{k=1}^{K^i} = \{\mathcal{P}_k\}_{k=1}^{K}$. Note that there may be a correlation across decision paths, i.e., $P(\mathcal{P}_i, \mathcal{P}_j) \neq P(\mathcal{P}_i)P(\mathcal{P}_j)$ for $i \neq j$. For example, one decision path corresponding to a logic program of the robot moving forward at high speed has a correlation to another decision path for moving forward at low speed. This may occur because a neuron of a policy can learn arbitrary behaviors. However, this leads to a non-orthogonal ground-truth factor set and can be undesirable since high correlations of a neuron to multiple ground-truth factors (e.g., $I[z^i; \mathcal{P}_i]$ and $I[z^i; \mathcal{P}_j]$ are large) can result from not only entanglement of the neuron but also the correlation between factors (e.g., $I[\mathcal{P}_i; \mathcal{P}_j]$ is large). Hence, this urges the need to calibrate mutual information for computing disentanglement measures. We start by adapting the Mutual Information Gap (MIG) [20] to our framework:

$$\frac{1}{K} \sum_{k=1}^{K} \frac{1}{H[\mathcal{P}_k]} \Big(I[z^{i^*}; \mathcal{P}_k] - \max_{j \neq i^*} I[z^j; \mathcal{P}_k] - I[z^j; \mathcal{P}_k; \mathcal{P}_{k^j}] \Big) \tag{5}$$

where H is entropy, I is interaction information that can take an arbitrary number of variables (with 2 being mutual information), $i^* = \arg\max_i I[z^i; \mathcal{P}_k]$, and $k^j = \arg\max_l I[z^j; \mathcal{P}_l]$. Intuitively, this measures the normalized difference between the highest and the second-highest mutual

Table 1: Quantitative results of classical control.

Network Architecture	Disentanglement Variance ↓	MI-Gap ↑	Modularity ↑	Explanation Size ↓ Vertical	Horizontal	Cognitive Chunks ↓
FCs	0.0242	0.3008	0.9412	5.00	1.91	1.65
GRU	0.0329	0.2764	0.9096	4.90	1.96	1.65
LSTM	**0.0216**	0.2303	0.9355	4.75	2.02	1.90
ODE-RNN	0.0287	0.3062	0.9376	4.90	1.93	1.80
CfC	0.0272	0.2892	0.9067	4.70	1.82	1.50
NCP	0.0240	**0.3653**	**0.9551**	**3.45**	**1.51**	**1.30**

Table 2: Alignment between disentanglement and explanation quality in classical control.

Re-signed Rank Correlation ↑	Explanation Size Vertical	Horizontal	Cognitive Chunks
Variance	-0.146	0.002	0.040
MI-Gap	**0.427**	**0.505**	**0.449**
Modularity	-0.114	0.156	0.032

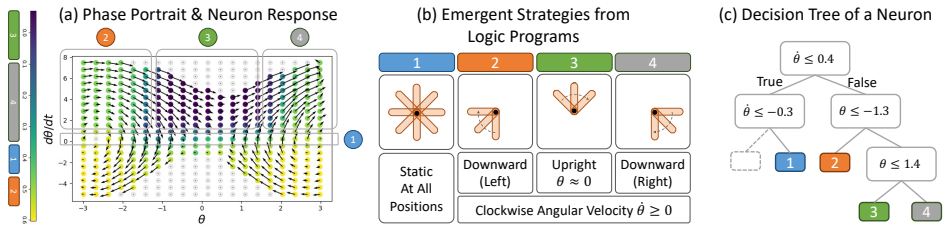


Figure 1: In classical control (Pendulum): (a) Phase portrait with empirically measured closed-loop dynamics and neuron response. Each arrow and colored dot are the results averaged around the binned state space. (b) Emergent strategies from logic programs. (c) Decision tree extracted for command neuron 3 in NCP.

information of each decision path with individual neuron activation. The last term $I[z^j; \mathcal{P}_k; \mathcal{P}_{k^j}]$ is for calibration and captures the inherent correlation between z^j and $\mathcal{P}_k$ resulted from potentially nonzero $I[\mathcal{P}_k; \mathcal{P}_{k^j}]$ with $\mathcal{P}_{k^j}$ being a proxy random variable of z^j in the ground-truth factor set. We show how to compute the term $I[z^j; \mathcal{P}_k] - I[z^j; \mathcal{P}_k; \mathcal{P}_{k^j}]$ in practice in Appendix Section C.

C. Modularity. We compute modularity scores from [28] with the same calibration term,

$$\frac{1}{\mathcal{I}} \sum_{i \in \mathcal{I}} 1 - \frac{\sum_{k \neq k^*} (I[z^i; \mathcal{P}_k] - I[z^i; \mathcal{P}_k; \mathcal{P}_{k^*}])^2}{(K-1) I[z^i; \mathcal{P}_{k^*}]^2}, \tag{6}$$

where $k^* = \arg\max_l \; I[z^i; \mathcal{P}_l]$. In a modular representation, each neuron depends on at most a single factor of variation. Suppose for each neuron i has the best "match" with a decision path (ground-truth factor) k^*, non-modularity of that neuron is computed as the normalized variance of mutual information between its neuron response and all non-matched decision paths $\{\mathcal{P}_k\}_{k \neq k^*}$. In practice, we discretize neuron responses into N bins to compute discrete mutual information.

4 Experiments

We conduct a series of experiments in various policy-learning tasks to answer the following: (i) *How effective is disentanglement to measure the interpretability of policies?* (ii) *What can we extract from neural responses?* (iii) *What architecture is more interpretable through the lens of disentanglement?*

4.1 Setup

Network architecture. We construct compact neural networks for each end-to-end learning to control task. For all tasks, our networks are constructed by the following priors: (i) Each baseline network is supplied with a perception backbone (e.g., a convolutional neural network) (ii) We construct policies based on different compact architectures that take in feature vectors from the perception backbone and output control with comparable cell counts (instead of actual network size in memory as we assess interpretability metrics down to cell-level). The perception backbone is followed by a neural controller designed by compact feed-forward and recurrent network architectures including fully-connected network (**FCs**), gated recurrent units (**GRU**) [29], and long-short term memory (**LSTM**) [30]. Additionally, we include advanced continuous-time baselines designed by ordinary differential equations such as **ODE-RNN** [31], closed-form continuous-time neural models (**CfCs**) [32], and neural circuit policies (**NCPs**) [4]. We interpret the dynamics of the neurons in the last layer before the output in FCs, the command-neuron layer of NCPs, and the recurrent state of the rest. We then extract logic programs and measure interpretability with the proposed metrics.

Evaluation. To evaluate the effectiveness of measuring interpretability through the lens of disentanglement, we adopt the metrics proposed in [33], which studies human interpretability of decision sets [34] (a representation of explanation similar to that in this work). They show human response

Table 3: Quantitative results of locomotion.

Network Architecture	Disentanglement Variance ↓	MI-Gap ↑	Modularity ↑	Explanation Size ↓ Vertical	Horizontal	Cognitive Chunks ↓
FCs	0.0187	0.1823	0.9622	5.66	2.54	4.02
GRU	0.0259	0.1830	0.9713	5.78	2.52	3.94
LSTM	0.0108	0.1453	0.9600	5.62	2.52	3.92
ODE-RNN	0.0210	0.1880	0.9701	6.00	2.57	4.16
CfC	0.0234	0.1596	0.9628	5.94	2.58	4.20
NCP	**0.0107**	**0.2164**	**0.9791**	**3.94**	**2.08**	**2.72**

Table 4: Alignment between disentanglement and explanation quality in locomotion.

Re-signed Rank Correlation ↑	Explanation Size Vertical	Horizontal	Cognitive Chunks
Variance	**0.512**	0.456	0.443
MI-Gap	0.422	**0.504**	**0.481**
Modularity	0.170	0.180	0.173

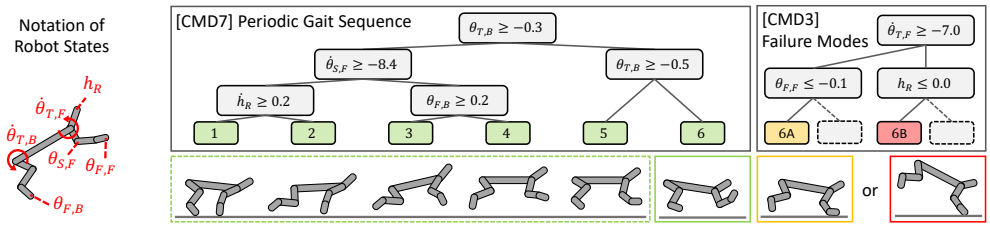


Figure 2: Qualitative neural activation along a gait sequence on HalfCheetah [35]. We focus on neurons 7 and 3 for illustrative purposes. Neuron 7 exhibits a periodic activation pattern that reacts to different phases of the gait cycle (left). Neuron 3 displays peak activity in situations with the potential to destabilize gait, such as early touchdown (left) and forward flipping (right). Our approach can therefore help to facilitate failure detection by monitoring responses of individual key neurons.

time and subjective satisfaction are highly correlated with *explanation size* and *cognitive chunks*. Explanation size consists of vertical size , the number of cases (the number of decision paths per neuron), and horizontal size, the complexity of each case (the length of each decision path). Cognitive chunks refer to the presentation of newly defined concepts, which we quantify as the introduction of new symbols in the logic program. Furthermore, we measure the alignment of disentanglement quantification and the above-mentioned explanation quality metrics. We compute *re-signed rank correlation* by re-signing Spearman's rank correlation coefficient to make larger values always refer to better alignment, e.g., given higher modularity being better while lower explanation size being better, better alignment corresponds to negative correlation and we thus negate the coefficient.

4.2 Classical Control

Environment and policy learning. We use the OpenAI Gym Classical Control Pendulum task [35]. The environment has simple yet nonlinear dynamics and allows for straightforward visualization of the entire state space. The environment states include θ (joint angle) and $\dot{\theta}$ (joint angular velocity). θ is in the range of $\pm\pi$ with $\theta = 0$ as the upright position. $\dot{\theta}$ is along the clockwise direction. The control is u (joint torque). The goal is to stabilize at the upright position ($\theta = \dot{\theta} = 0$) with limited control energy consumption ($u \downarrow$). We use Proximal Policy Optimization (PPO) [36] to train the policy with early stop by reaching episode reward -500 or a maximal number of training iterations. We run each model with 5 different random seeds and report average results.

Quantitative analysis. Table 1 shows that, among all models, NCP achieves the best performance in disentanglement and explanation quality (i.e., explanation size and cognitive chunks), suggesting that it is more interpretable from the perspective of both our work and [33]. Beyond alignment of the best performance, Table 2 indicates the consistency of overall ranking between disentanglement and explanation quality. We found that while variance and modularity are (partially) aligned in the best performance in Table 1, only the mutual information gap is correlated to explanation quality in the overall ranking. Another interesting finding is that CfCs have the lowest logic conflict. By empirically checking the decision trees, they construct non-trivial but highly-overlapping decision paths, thus leading to considerably fewer conflicts in logic programs across neurons.

Neuron responses and underlying behaviors. While all models learn reasonable strategies, as exemplified by focusing on the sign of θ and $\dot{\theta}$, we now dive deeper into understanding individual neural dynamics. To this end, we focus on NCPs as they provide a lower variance from the disentanglement perspective in their logic programs. We found different neurons roughly subdivide the state space into quadrants and focus on their respective subsets. In Figure 1, we show the interpretability analysis of command neuron 3 as an example. This neuron developed fine-grained strategies for different situations like swinging clockwise in the right or left downward positions, and stabilizing

Table 5: Quantitative results of visual servoing.

Network Architecture	Disentanglement Variance ↓	MI-Gap ↑	Modularity ↑	Explanation Size ↓ Vertical	Horizontal	Cognitive Chunks ↓
FCs	0.0124	0.1354	0.9704	4.88	2.36	2.88
GRU	0.0158	0.1614	0.9801	3.88	**1.94**	**1.75**
LSTM	0.0172	0.1950	**0.9851**	4.25	2.06	2.38
ODE-RNN	0.0151	0.1588	0.9766	5.25	2.24	2.75
CfC	0.0191	0.1391	0.9677	5.50	2.43	3.00
NCP	**0.0068**	**0.3902**	0.9770	**4.12**	2.00	1.88

Table 6: Alignment between disentanglement and explanation quality in visual servoing.

Re-signed Rank Correlation ↑	Explanation Size Vertical	Horizontal	Cognitive Chunks
Variance	0.371	0.314	0.314
MI-Gap	**0.657**	**0.771**	**0.771**
Modularity	0.257	0.314	0.314

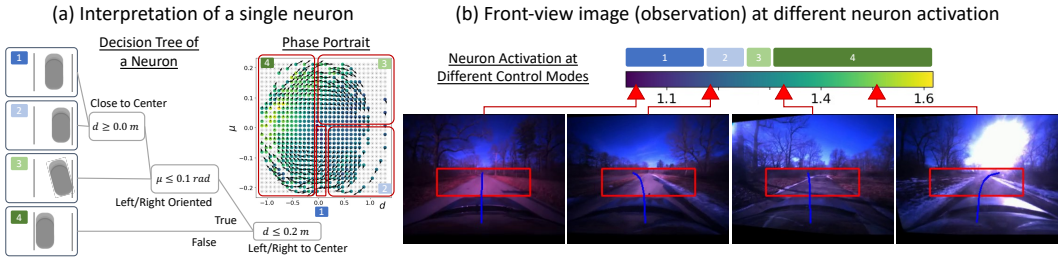


Figure 3: Explanation of neural policies for end-to-end visual servoing (Image-based Driving). (a) Phase portrait of local heading error μ and lateral deviation d with empirically measured mean neuron response and closed-loop dynamics. (b) Front-view images retrieved based on neuron response.

positive angular velocity around the upright position, as shown in Figure 1 (b)(c). We further provide phase portrait in Figure 1 (a). The arrows indicate empirically measured closed-loop dynamics (with control from the policy) and the color coding indicates average neuron response at a specific state from evaluation. The color of the neuron response (which can correspond to logic programs) and the arrows (which implicitly capture actions) highlight different fine-grained strategies in the phase portrait. Note that this finding is not limited to NCPs, and could be observed in other network architectures with similar functions as well.

4.3 Locomotion

Environment and policy learning. We consider a planar locomotion task based on OpenAI Gym's HalfCheetah environment [35]. The agent is rewarded for forward locomotion based on a simple base velocity reward. We optimize our policies with PPO until a maximum number of episodes has been reached. For each model, we run five trials with different random seeds and report average results. Here, our objective is to extend our interpretability framework to a higher-dimensional control task. Specifically, we investigate whether our approach is capable of extracting consistent single-neuron activation patterns that align with individual phases of a periodic gait cycle.

Quantitative analysis. In Table 3, we observe consistent results with classical control that NCP achieves the best performance in disentanglement and explanation quality. We further observe that the LSTM achieves a desirable low disentanglement variance comparable to NCPs. LSTMs and most networks compared to NCPs on the other hand show a lower Mutual Information Gap. This suggests that in these networks neuron responses are concentrated for different decision paths but not quite identifiable from a probabilistic perspective, as certain neuron activation cannot be uniquely mapped to a decision path. Besides, Table 4 shows consistent findings that the mutual information gap has the best ranking correlation with explanation quality. As opposed to classical control, the ranking correlation is overall much higher and we hypothesize that more complex tasks may yield better alignment in overall ranking between disentanglement and explanation quality.

Exploring Gait Pattern. As the state space is significantly larger than the Pendulum task, we complement our quantitative interpretability results with qualitative results that focus on two exemplary neurons, namely command neurons (CMDs) 3 and 7. Figure 2 provides the extracted decision trees for CMDs 7 (left) and 3 (right). We find that the former displays periodic activation patterns that align very well with individual phases of regular gait. In particular, it leverages position readings of the back thigh joint in conjunction with fore shin velocity to coarsely differentiate between stance and flight phase. More fine-grained coordination of lift-off and touchdown is handled by the leftmost and rightmost branches, respectively. In addition to periodic neuron activations following regular gait, we also observe more specialized decision trees that respond to potential safety-critical situ-

ations. For example, the decision tree of CMD 3 includes two branching options that align with variations of tripping due to premature touchdown during the flight phase corresponding to a forward trip (6A) and a forward flip (6B). Monitoring the respective activation ranges of CMD 3 can therefore serve as a proxy for failure detection.

4.4 End-to-end Visual Servoing

Environment and policy learning. We consider vision-based end-to-end autonomous driving where the neural policy learns steering commands for lane-following. The model takes front-view RGB images of the vehicle as input, and outputs control commands for the steering wheel and speed. We use the high-fidelity data-driven simulator *VISTA* [37] as our environment. We adopt a training strategy called *guided policy learning* that leverages *VISTA* to augment a real-world dataset with diverse synthetic data for robust policy learning. The training dataset contains roughly 200k image-and-control pairs and mean squared error is used as the training objective. For evaluation, we initialize the vehicle at a random position throughout the entire track and evaluate the policy for 100 frames (roughly 10s) for 100 episodes. The performance is estimated as the ratio of the length of the path traversed without a crash and the total path length. Note that apart from classical control and locomotion, other than task definition, this task has two additional major distinctions: (1) policy learning based on supervised learning as opposed to reinforcement learning (2) policies take in data (images) different from states on which logic program is grounded.

Quantitative analysis. In Table 5, we have consistent findings with classical control and locomotion that NCP achieves the best performance in both disentanglement and explanation quality (more precisely, comparable with the best in modularity, horizontal explanation size, and cognitive chunks). In Table 6, the mutual information gap achieves consistently the best alignment in the overall ranking. Also, the overall numerical values of ranking correlation are much higher than classical control and comparable with or higher than locomotion. This suggests again that more complex tasks yield better alignment between disentanglement and explanation quality.

Maneuver strategies from visual inputs. In Figure 3, we show extracted behaviors for a neuron in the NCP driving policy. While the state space of driving is higher dimensional, we focus on local heading error μ and lateral deviation from the lane center d in the lane following task. We compute the statistics and plot neuron response and closed-loop dynamics in the d-μ phase portrait. This specific neuron develops more fine-grained control for situations when the vehicle is on the right of the lane center, as shown in Figure 3(a), with images retrieved from neuron response in Figure 3(b).

5 Discussion and Limitation

We summarize all consistent findings to answer the questions asked at the beginning of Section 4. First, disentanglement is highly indicative of explanation quality in the best performance across all tasks, suggesting that, among all compact neural policies, the ones with more disentangled representation are more interpretable for humans [33]. We further showed these results are robust across hyperparameters in Appendix Section F. In addition, compared to neuron response variance and modularity, the mutual information gap consistently has the best alignment in the overall ranking with explanation quality. Besides, there are certain network architectures (NCPs) that exhibit superior performance in disentanglement and explanation quality consistently across experiments. Another interesting finding is that more complex tasks yield better alignment between disentanglement and explanation quality (by comparing the numerical values of all entries between Table 2, 4, and 6). Finally, by digging into qualitative results, we showed that learned behaviors of neural policies, e.g., gait patterns or maneuver strategies, can be extracted from neuron responses. **Limitation.** The proposed framework involves extracting factor of variation relevant to strategies and task structures; however, the empirical implementation only considers abstraction (i.e., logic program) in a single time step. Extensions to temporal reasoning could include the introduction of temporal logic [38] or using decision trees with temporal capability [25, 39]. Furthermore, the abstraction is grounded on a set of world states pre-determined by human; however, these states may not be sufficiently expressive to capture the learned behavior of the policy. This requires estimating the information carried between observation and grounding symbols or methods to extract the latter from the former.

References

[1] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[2] M. Lechner, R. Hasani, M. Zimmer, T. A. Henzinger, and R. Grosu. Designing worm-inspired neural networks for interpretable robotic control. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 87–94. IEEE, 2019.

[3] R. Hasani, M. Lechner, A. Amini, D. Rus, and R. Grosu. Liquid time-constant networks. *arXiv preprint arXiv:2006.04439*, 2020.

[4] M. Lechner, R. Hasani, A. Amini, T. A. Henzinger, D. Rus, and R. Grosu. Neural circuit policies enabling auditable autonomy. *Nature Machine Intelligence*, 2(10):642–652, 2020.

[5] C. Vorbach, R. Hasani, A. Amini, M. Lechner, and D. Rus. Causal navigation by continuous-time neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

[6] J. Schmidhuber. Learning factorial codes by predictability minimization. *Neural computation*, 4(6):863–879, 1992.

[7] J. Peters, D. Janzing, and B. Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[8] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019.

[9] S. Emmons, B. Eysenbach, I. Kostrikov, and S. Levine. Rvs: What is essential for offline RL via supervised learning? In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=S874XAIpkR-.

[10] C. R. Banbury, V. J. Reddi, M. Lam, W. Fu, A. Fazel, J. Holleman, X. Huang, R. Hurtado, D. Kanter, A. Lokhmotov, et al. Benchmarking tinyml systems: Challenges and direction. *arXiv preprint arXiv:2003.04821*, 2020.

[11] C. Baykal, L. Liebenwein, I. Gilitschenski, D. Feldman, and D. Rus. Sensitivity-informed provable pruning of neural networks. *SIAM Journal on Mathematics of Data Science*, 4(1): 26–45, 2022. doi:10.1137/20M1383239. URL https://doi.org/10.1137/20M1383239.

[12] C. Hawkins, X. Liu, and Z. Zhang. Towards compact neural networks via end-to-end training: A bayesian tensor approach with automatic rank determination. *SIAM Journal on Mathematics of Data Science*, 4(1):46–71, 2022. doi:10.1137/21M1391444. URL https://doi.org/10.1137/21M1391444.

[13] S. Oymak. Learning compact neural networks with regularization. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3966–3975. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/oymak18a.html.

[14] J. Bu, A. Daw, M. Maruf, and A. Karpatne. Learning compact representations of neural networks using discriminative masking (dam). *Advances in Neural Information Processing Systems*, 34, 2021.

[15] M. Torkamani, P. Wallis, S. Shankar, and A. Rooshenas. Learning compact neural networks using ordinary differential equations as activation functions. *arXiv preprint arXiv:1905.07685*, 2019.

[16] B. A. Toms, E. A. Barnes, and I. Ebert-Uphoff. Physically interpretable neural networks for the geosciences: Applications to earth system variability. *Journal of Advances in Modeling Earth Systems*, 12(9):e2019MS002002, 2020.

[17] R. Hasani, A. Amini, M. Lechner, F. Naser, R. Grosu, and D. Rus. Response characterization for auditing cell dynamics in long short-term memory networks. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.

[18] D. Alvarez Melis and T. Jaakkola. Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, 31, 2018.

[19] Q. Zhang, Y. N. Wu, and S.-C. Zhu. Interpretable convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8827–8836, 2018.

[20] R. T. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31, 2018.

[21] A. Heuillet, F. Couthouis, and N. Díaz-Rodríguez. Explainability in deep reinforcement learning. *Knowledge-Based Systems*, 214:106685, 2021.

[22] O. Bastani, Y. Pu, and A. Solar-Lezama. Verifiable reinforcement learning via policy extraction. *Advances in neural information processing systems*, 31, 2018.

[23] A. Silva, M. Gombolay, T. Killian, I. Jimenez, and S.-H. Son. Optimization methods for interpretable differentiable decision trees applied to reinforcement learning. In *International conference on artificial intelligence and statistics*, pages 1855–1865. PMLR, 2020.

[24] Z. Ding, P. Hernandez-Leal, G. W. Ding, C. Li, and R. Huang. Cdt: Cascading decision trees for explainable reinforcement learning. *arXiv preprint arXiv:2011.07553*, 2020.

[25] A. Pace, A. J. Chan, and M. van der Schaar. Poetree: Interpretable policy learning with adaptive decision trees. *arXiv preprint arXiv:2203.08057*, 2022.

[26] D. Amir and O. Amir. Highlights: Summarizing agent behavior to people. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1168–1176, 2018.

[27] N. Topin and M. Veloso. Generation of policy-level explanations for reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2514–2521, 2019.

[28] K. Ridgeway and M. C. Mozer. Learning deep disentangled embeddings with the f-statistic loss. *Advances in neural information processing systems*, 31, 2018.

[29] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

[30] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[31] Y. Rubanova, R. T. Chen, and D. K. Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. *Advances in neural information processing systems*, 32, 2019.

[32] R. Hasani, M. Lechner, A. Amini, L. Liebenwein, M. Tschaikowski, G. Teschl, and D. Rus. Closed-form continuous-depth models. *arXiv preprint arXiv:2106.13898*, 2021.

[33] I. Lage, E. Chen, J. He, M. Narayanan, B. Kim, S. Gershman, and F. Doshi-Velez. An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1902.00006*, 2019.

[34] H. Lakkaraju, S. H. Bach, and J. Leskovec. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1675–1684, 2016.

[35] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

[36] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[37] A. Amini, T.-H. Wang, I. Gilitschenski, W. Schwarting, Z. Liu, S. Han, S. Karaman, and D. Rus. Vista 2.0: An open, data-driven simulator for multimodal sensing and policy learning for autonomous vehicles. In *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022.

[38] A. Camacho and S. A. McIlraith. Learning interpretable models expressed in linear temporal logic. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 29, pages 621–630, 2019.

[39] L. Console, C. Picardi, and D. T. Duprè. Temporal decision trees: Model-based diagnosis of dynamic systems on-board. *Journal of artificial intelligence research*, 19:469–512, 2003.

[40] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

[41] J. Frankle and M. Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2018.

[42] L. Liebenwein, R. Hasani, A. Amini, and D. Rus. Sparse flows: Pruning continuous-depth models. *Advances in Neural Information Processing Systems*, 34, 2021.

[43] L. Liebenwein, C. Baykal, B. Carter, D. Gifford, and D. Rus. Lost in pruning: The effects of pruning neural networks beyond test accuracy. *Proceedings of Machine Learning and Systems*, 3:93–138, 2021.

[44] L. Li, T. J. Walsh, and M. L. Littman. Towards a unified theory of state abstraction for mdps. In *AI&M*, 2006.

[45] T. D. Kulkarni, K. Narasimhan, A. Saeedi, and J. Tenenbaum. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. *Advances in neural information processing systems*, 29, 2016.

[46] E. S. Spelke and K. D. Kinzler. Core knowledge. *Developmental science*, 10(1):89–96, 2007.

[47] T. Dean and R. Givan. Model minimization in markov decision processes. In *AAAI/IAAI*, pages 106–111, 1997.

[48] E. van der Pol, T. Kipf, F. A. Oliehoek, and M. Welling. Plannable approximations to mdp homomorphisms: Equivariance under actions. *arXiv preprint arXiv:2002.11963*, 2020.

[49] I. Higgins, D. Amos, D. Pfau, S. Racaniere, L. Matthey, D. Rezende, and A. Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

[50] H. Caselles-Dupré, M. Garcia Ortiz, and D. Filliat. Symmetry-based disentangled representation learning requires interaction with environments. *Advances in Neural Information Processing Systems*, 32, 2019.