Estimating Causal Responsibility for Explaining Autonomous Behavior

Saaduddin Mahmud[1⋆], Samer B. Nashed[1⋆], Claudia V. Goldman[2], and Shlomo Zilberstein[1]

[1] University of Massachusetts Amherst, Massachusetts, USA
{smahmud,snashed,shlomo}@cs.umass.edu
[2] General Motors Research, Tel Aviv, Israel
claudia.goldman@gm.com

Abstract. There has been growing interest in causal explanations of stochastic, sequential decision-making systems. Structural causal models and causal reasoning offer several theoretical benefits when exact inference can be applied. Furthermore, users overwhelmingly prefer the resulting causal explanations over other state-of-the-art systems. In this work, we focus on one such method, MeanRESP, and its approximate versions that drastically reduce compute load and assign a responsibility score to each variable, which helps identify smaller sets of causes to be used as explanations. However, this method, and its approximate versions in particular, lack deeper theoretical analysis and broader empirical tests. To address these shortcomings, we provide three primary contributions. First, we offer several theoretical insights on the sample complexity and error rate of approximate MeanRESP. Second, we discuss several automated metrics for comparing explanations generated from approximate methods to those generated via exact methods. While we recognize the significance of user studies as the gold standard for evaluating explanations, our aim is to leverage the proposed metrics to systematically compare explanation-generation methods along important quantitative dimensions. Finally, we provide a more detailed discussion of MeanRESP and how its output under different definitions of responsibility compares to existing widely adopted methods that use Shapley values.

Keywords: Causal Inference · Explainable AI · MDPs.

1 Introduction

Researchers from many fields have shown that developing trust in AI systems is required for their timely adoption and proficient use [18, 35, 39]. It is also widely accepted that autonomous agents with the ability to explain their decisions increase user trust [7, 13, 22]. However, there are many challenges in generating explanations. Consider, for example, an agent managing load on a power grid by setting electricity prices and engaging other physical resources within the grid.

⋆ Authors contributed equally.

Generating suitable explanations of such a system is hard due to the complexity of planning, which may involve large state spaces, stochastic actions, imperfect observations, and complicated objectives. Furthermore, useful explanations must somehow reduce the internal reasoning process to a form understandable by a user who likely does not know all of the algorithmic details. One significant class of autonomous decision-making models for which there is a desire to generate explanations is the Markov decision process (MDP) and its derivatives.

In our previous work [27], we developed a framework, based on *structural causal models* (SCMs) [11], for applying causal analysis to sequential decision-making agents. This framework creates an SCM representing the computation needed to derive a policy for an MDP and applies causal inference to identify variables that cause certain agent behavior. Explanations are then generated using these variables, for example by completing natural language templates. This framework is both theoretically sound, based on formalisms from the causality literature, as well as flexible, allowing multiple semantically different types of explanans.

This method, known as MeanRESP, has many different approximate versions and is compatible with several definitions of responsibility [8]. The theoretical characteristics of approximate MeanRESP, as well as its performance compared to the exact version, are yet to be explored in detail. Since in practice, the approximate versions are the most likely to be deployed, we see this as a critical gap in our current understanding of how to explain MDP agent behaviors. Moreover, as MeanRESP may produce many causes related to a decision, it is often necessary to reduce the size of this set to make the explanations more concise and therefore easier to understand. MeanRESP supports this type of 'top k' analysis natively, but little work has been done on understanding how to compare different outputs, either against each other or against the output from the exact version of the algorithm. To this end, we also propose several metrics which may be used to compare MeanRESP outputs at different levels of approximation. These metrics capture diverse types of differences and underscore the difficulty of devising a single metric for evaluating objects as complicated, nuanced, and context-dependent as explanations.

Our results include theoretical analyses regarding the correctness and sampling error rates for causal and responsibility determination for approximate MeanRESP, discuss several potential metrics for comparing explanations, empirical analyses of sampling error convergence rates and explanation dissimilarity between different versions of MeanRESP and Shapley-value based methods. Overall, these results establish several key facts about approximate MeanRESP as well as open the door for a variety of avenues of continued research.

2 Related Work

While this paper focuses specifically on deepening analysis related to one particular algorithm for automatic explanation generation, MeanRESP, the body of work on explainable machine learning (XML) — a focus area that aims to

explain the decisions of black-box machine learning algorithms [15, 19, 26] — and explainable planning (XAIP) — a focus area that aims to explain the outputs of planning algorithms or modify planning algorithms so that they produce plans that are inherently more explainable — is large and growing rapidly. In this section, we aim to provide some context to the existing literature to highlight the importance of MeanRESP.

In XAIP literature, one common method for explaining complex planners is via policy summarization, where either A) the original reasoning problem is made simpler and then the solution is explained exactly, or B) the original problem is not reduced, but the solution (e.g., policy) is simplified post-hoc and the simplified policy is explained. For example, Pouget et al. [30] identify key state-action pairs via spectrum-based fault localization, and Russell et al. [31] use decision trees to approximate a given policy and analyze the decision nodes to determine which state factors are most influential for immediate reward. Panigutti et al. [29] used similar methods to explain classifiers. These methods are appealing in that they parallel our intuitions about simplification in a number of other settings, such as analogizing during an explanation [9], science communication [32], and even other AI tools, like automated text simplification [28, 34] or summarization [1]. However, these methods are driven primarily by heuristics and may be difficult to generalize to the many different forms of planners and models.

Research on explanations of stochastic planners specifically, such as MDPs, is relatively sparse. However, there are several notable existing efforts. Most present heuristics that are specifically designed for MDPs, such as generating counterfactual states and then identifying important state factors by analyzing how the value function changes given perturbations to different state factors [10]. Wang et al. [38] try to explain policies of partially observable MDPs by communicating the relative likelihoods of different events or levels of belief. However, research clearly indicates that humans are not good at using this kind of numerical information [23].

A more common heuristic approach is to analyze (and produce explanations that reference) the reward function. Khan et al. [16] first presented a technique to explain policies for factored MDPs by analyzing the expected occupancy frequency of states with extreme reward values. Later, Sukkerd et al. [36] proposed explaining factored MDPs by annotating them with "quality attributes" (QAs) related to independent, measurable cost functions. Explanations describe the QA objectives, expected consequences QA values given a policy, and how those values contribute to the policy's expected cost. The system also explains whether the policy achieves the best possible QA values simultaneously, or if there are competing objectives that required reconciliation and proposes counterfactual alternatives. Thus, it explains entire policies, not individual actions, using custom graphics and natural language templates, the latter of which has become the de facto standard for automatic explanations. Instead of looking at how the policy is affected by the reward function overall, Juozapaitis et al. [14] analyze how extreme reward values impact action selection in decomposed-reward RL agents, and Bertram and Peng [4] look at reward sources in deterministic MDPs.

While these approaches are computationally cheap and easy to implement, they have limited scope in the explanations they provide, and do not have many theoretical advantages, if any. Thus, recently, some research has investigated the application of causal modeling and causal analysis to the automatic generation of explanations for planners, including MDPs. One particularly compelling framework for doing so, which we study in this paper, is a method called MeanRESP [27]. MeanRESP is based on a responsibility attribution method called RESP, introduced in [3] to explain classification outcomes, which has its roots in prior work on formal definitions of causality and responsibility [11, 12, 8]. In this paper, we examine several choices related to the definitions of responsibility for use within MeanRESP.

The most similar work to this paper is other research that has proposed using SCMs for explaining MDPs and their variants in both planning and learning scenarios. Madumal et al. [21] use SCMs to encode the influence of particular actions available to the agent in a model-free, reinforcement learning, where it requires several strong assumptions including the prior availability of a graph representing causal direction between variables, discrete actions, and the existence of sink states.

Finally, our approach to estimating causal responsibility can be viewed as a form of feature attribution, which is a common approach in explainable Machine Learning (XML) for feature ranking [25], most often via Shapley values and their approximations [20, 33]. In this paper, we conduct a quantitative comparison between Shapley values and different versions of MeanRESP. Specifically, we analyze the approximation error between a prominent Shapley value-based feature attribution method [37] and various versions of approximate MeanRESP, considering the number of samples. Additionally, we assess the dissimilarity between explanations generated by these two attribution methods. The purpose of this comparison is to investigate whether there exists a significant disparity in the content of the explanations produced by these methods, potentially motivating future research on the relative advantages of each method.

3 Background

Here, we review some concepts and notations relevant to the three main ideas this paper builds upon: structural causal models (SCMs), our working definition of cause, and Markov decision processes (MDPs).

3.1 Structural Causal Models, Actual Causes, and Responsibility

SCMs model scenarios $\mathcal{S} = \langle U, V, \mathcal{M} \rangle$, which break causality or attribution problems down into three components:

1. A set of variables U, known as the context, which are required to define the scenario, but which should not be identified as causal. These variables are considered fixed for a given scenario. The choice of which variables belong

in the context is a design choice, and the main function of the context is to bound the size of the total problem.
2. A set of variables V, known as the endogenous variables, which we may want to identify as causal or highlight in an explanation. All variables in a scenario must be in $U \cup V$.
3. A set of equations, $\mathcal{M}$, which model how variables in V are calculated as functions of variables in U or other variables in V.

Nashed et al. [27] define several SCM representations of an MDP with different choices of the context U. For the purpose of analysis, throughout the rest of this paper, we will consider one of the most natural of those choices and describe its mathematical definition and interpretation in the following subsection. We now review our working definition of cause from [12].

Definition 1. *Let $X \subseteq V$ be a subset of the endogenous variables, and let x be a specific assignment of values for those variables. Given an event ϕ, defined as a logical expression, for instance $\phi = (\neg a \wedge b)$, a* weak cause *of ϕ satisfies the following conditions:*

1. *Given the context $U = u$ and $X = x$, ϕ holds.*
2. *Some $W \subseteq (V \setminus X)$ and some $\bar{x}$ and w exist such that:*
   A) using these values produces $\neg\phi$.
   B) for all $W' \subseteq W$, $Z \subseteq V \setminus (X \cup W)$, where
   $w' = w|W'$ and $z = Z$ given $U = u$, ϕ holds when $X = x$.

Here, condition 2B) is saying that given context $U = u$, $X = x$ alone is sufficient to cause ϕ, independent of some other variables W. This and similar definitions of cause are often called "but-for" definitions. There is a related, slightly older definition due to [11] in which condition 2B) is replaced with the following, simpler statement: for all $Z \subseteq V \setminus (X \cup W)$, where $w = W$ and $z = Z$ given $U = u$, ϕ holds when $X = x$.

Actual causes are defined as minimal weak causes. That is, an actual cause is a weak cause C_W for which no set $C'_W \subset C_W$ is also a weak cause. Note that in this paper, we only consider $|C_W| = 1$, and therefore the above definition also defines actual causes. Table 1 provides a reference for the common related notation used throughout the paper.

3.2 Markov Decision Processes

A Markov decision process (MDP) is a model for reasoning in fully observable, stochastic environments [2], defined as a tuple $\langle S, A, T, R, d\rangle$. S is a finite set of states, where $s \in S$ may be expressed in terms of a set of *state factors*, $\langle f_1, f_2, \ldots, f_N\rangle$, such that s indexes a unique assignment of values to the factors f; A is a finite set of actions; $T : S \times A \times S \to [0, 1]$ represents the probability of reaching a state $s' \in S$ after performing an action $a \in A$ in a state $s \in S$; $R : S \times A \times S \to \mathbb{R}$ represents the expected immediate reward of reaching a state $s' \in S$ after performing an action $a \in A$ in a state $s \in S$; and $d : S \to [0, 1]$

Notation	Meaning
X	A set of decision variables, $X = \{X_1, X_2, X_3\}$
x	An assignment of values to the set X, $\{X_1 = x_1, X_2 = x_2, X_3 = x_3\}$
$\mathcal{P}(X)$	Power set of X
$\mathcal{D}(X_1)$	Domain of the joint assignments of all $x \in X$
$x' \leftarrow x\|X'$	x' is the restriction of x to X', e.g., if $X' = \{X_1\}$ and $x = \{X_1 = x_1, X_2 = x_2, X_3 = x_3\}$, then $x' = \{X_1 = x_1\}$
$x \leftarrow [x\langle x']$	Replace values of x with values from x', e.g., if $x = \{X_1 = x_1, X_2 = x_2\}$ and $x' = \{X_1 = b\}$, then $x = \{X_1 = b, X_2 = x_2\}$

Table 1: Important notations, summarized from [11].

represents the probability of starting in a state $s \in S$. A solution to an MDP is a policy $\pi : S \to A$ indicating that an action $\pi(s) \in A$ should be performed in a state $s \in S$. A policy π induces a value function $V^\pi : S \to \mathbb{R}$ representing the expected discounted cumulative reward $V^\pi(s) \in \mathbb{R}$ for each state $s \in S$ given a discount factor $0 \leq \gamma < 1$. An optimal policy π^* maximizes the expected discounted cumulative reward for every state $s \in S$ by satisfying the Bellman optimality equation $V^*(s) = \max_{a \in A} \sum_{s' \in S} T(s, a, s')[R(s, a, s') + \gamma V^*(s')]$.

One of the most natural ways to represent an MDP as an SCM is to let U consist of all variables related to the reward function R, transition function T, start distribution d, and discount factor γ. Then, V can be defined as $F \cup \mathit{\Pi}$, where F is the set of variables representing state factors, $F = \{f_1, f_2, \ldots, f_N\}$, and $\mathit{\Pi}$ is the set of variables representing the optimal policy, $\mathit{\Pi} = \{\pi_{s_1 a_1}, \pi_{s_1 a_2}, \ldots, \pi_{s_{|S|} a_{|A|}}\}$. Here, π_{sa} is a variable that is true when action a may be taken in state s. Thus, an obvious choice for an event ϕ is a subset of $\mathit{\Pi}$ and their assignment. For example, if action a is taken in state s instead of a', we have

$$\phi = \langle [\pi(s) = a], [\pi(s) = a'] \rangle = \langle \textsc{True}, \textsc{False} \rangle.$$

Under this modeling setup, counterfactual settings to F do not result in new MDP policies as they would be variables from R or T to be used in V. Instead, this setup represents a fixed world model and a fixed model of agent capability, where counterfactual inputs represent different situations, or states, that the agent may encounter. Although MeanRESP may be applied to other components of the MDP using different definitions of U and V, we focus on this particular setup as it is computationally less demanding for empirical analysis. We should note that none of our theoretical analysis relies on this particular definition of U and V, or even that MeanRESP is used to analyze an MDP instead of a classifier.

4 MeanRESP

Chockler and Halpern [8] defined the *responsibility* (RESP), of an actual cause X' with contingency set W as $\frac{1}{1+|W|}$. Based on that, we define the MeanRESP

score, ρ, of an actual cause X', to be the expected number of different ways X' satisfies the definition of actual cause weighted by a responsibility share. Hence, the MeanRESP score equates the strength of the causal effect with the number of different scenarios under which X' can be considered a cause for the event.

There are several plausible versions of MeanRESP, all of which detect sets of variables that satisfy the definition of actual cause given above. To facilitate understanding throughout the rest of the paper, we now provide a novel, high-level description of a generalized version of MeanRESP and its relation to different definitions of cause. Moreover, we would like the MeanRESP score to behave in a manner summarized by the following properties:

1. **Property 1:** A set of variables $X' \subseteq X$ that is not a cause of the event ϕ should have $\rho = 0$. A set of variables $X' \subseteq X$ that is a cause of the event ϕ should have $\rho > 0$.
2. **Property 2:** As the cause allows a set of witness variables, ρ should divide the causal responsibility among the cause and witness in a principled manner.
3. **Property 3:** A relatively higher value of ρ for a cause $X' \subseteq X$ should indicate the event ϕ is relatively more affected by the assignment $x|X'$ of X'.

Responsibility scores are important in practice since they allow both users and developers to differentiate between causes that are highly relevant to the given scenario and those which may have less explanatory value. Here, we present a generalized version of MeanRESP (Algorithm 1) that has all three of these properties. The algorithm considers witness sets of size up to $|W|_{max} = |X| - 1$ (Line 4). After fixing a witness $W = w$ (lines 6-7), it calculates the RESP score (line 9) using either RESP-UC (Algorithm 2) if we use weak cause Definition 1 or RESP-OC (Algorithm 3) if we use the original weak cause definition. In RESP-UC, if 2B holds from Definition 1 (lines 4-9), then we check for 2A (lines 10-12). Notice that RESP-UC will return a value greater than 0 whenever both 2A and 2B hold. This ensures property 1. Note that the condition in definition 1 always holds for a deterministic policy or classifier, and therefore is not explicitly checked. Additionally, in both RESP-UC and RESP-OC, accumulating the RESP scores in lines 13 and 10, respectively, provides property 2. Intuitively, the RESP score scales with the number of different ways X' satisfies the definition of actual cause weighted by a responsibility share. Hence, the RESP score equates the strength of the causal effect with the number of different scenarios under which X' can be considered a cause for the event. This gives MeanRESP property 3. We use the following notation to denote whether event ϕ occurred.

$$\phi(x_a) = \begin{cases} \text{True} & \text{if } \Pi(x) = \Pi(x_a) \\ \text{False} & \text{if } \Pi(x) \neq \Pi(x_a) \end{cases} \tag{1}$$

Overall, there are several design choices one can make regarding how exactly to compute the mean RESP scores, generating a family of closely related algorithms. First, either RESP-UC or RESP-OC may be used, depending on the desired definition of cause. Not only will this affect the resultant RESP scores,

Algorithm 1 MeanRESP

Input: All Causal Variables X, Variable of Interest X', Inference Model Π, Variable assignment x, Responsibility function $RESP$
Output: Mean Responsibility Scores ρ.
$MEANRESP \leftarrow 0$
for all $\beta = 0...|W|_{max}$ **do**
  $\sigma, T \leftarrow 0$
  for all $W \in \mathcal{P}(X \setminus X')$ such that $|W| = \beta$ **do**
    for all $w \in Dom(W)$ **do**
      $T \leftarrow T + 1$
      $\sigma \leftarrow \sigma + RESP(\Pi, X, X', x, d \sim Dom(X'), W, w)$
  $MEANRESP \leftarrow MEANRESP + \frac{\sigma}{T}$
return $\frac{MEANRESP}{|W|_{max}+1}$

Algorithm 2 RESP-UC

Input: Π, X, x, d, W, w
Output: score, σ.
$D_1, D_2 \leftarrow 1$
for all $W' \in \mathcal{P}(W)$ **do**
  $w' \leftarrow w|W'$
  $x_p \leftarrow [x\langle w']$
  if $\neg\phi(x_p)$ **then**
    $D_1 \leftarrow 0$
    break
$x_m \leftarrow [x\langle(d \cup w)]$
if $\phi(x_m)$ **then**
  $D_2 \leftarrow 0$
return $\frac{D_1}{1+|W|} D_2$

but most importantly, it will change what is identified as a cause; some sets of variables will have RESP scores of zero under one definition but not the other.

Second, the mean RESP score can be calculated in two ways. It may be tallied over only the witness sets of size β_{min}, where β_{min} is the smallest β for which there exists a satisfying witness set (as in [27]). Or, it may be tallied overall witness sets, regardless of β, as in Algorithm 1. Actual causes with at least some small witness sets will receive lower RESP scores under the latter design.

Third, as responsibility incrementally accrues with respect to an actual causal set, these increments can either be counted equally or can be normalized by the size of the domain of the actual cause. We refer to this as the option to perform domain normalization, and the theory behind it is that with a larger domain the chance that some assignment $X = \bar{x}$ will meet the conditions of Definition 1 increases, and thus the responsibility should correspondingly decrease.

None of these choices interfere with properties 1-3, but they may subtly alter relative responsibility assigned to different actual causes. As there is no clear reason based on first principles as to the correct choice, these decisions involve

Algorithm 3 RESP-OC

Input: Π, X, x, d, W, w
Output: score, σ.
$D_1, D_2 \leftarrow 1$
$x_p \leftarrow [x\langle w]$
if $\neg\phi(x_p)$ **then**
  $D_1 \leftarrow 0$
$x_m \leftarrow [x\langle(d \cup w)]$
if $\phi(x_m)$ **then**
  $D_2 \leftarrow 0$
return $\frac{D_1}{1+|W|} D_2$

Algorithm 4 Sampled MeanRESP

Input: All Causal Variable X, Variable of Interest X', Inference Model Π, Variable assignment x, Responsibility function $RESP$, Sample Size, T
Output: Mean Responsibility Scores ρ.
$\sigma \leftarrow 0$
for all $t = 0...T$ **do**
  $W \sim \mathcal{P}(X \setminus X')$
  $w \sim Dom(W)$
  $\sigma \leftarrow \sigma + RESP(\Pi, X, X', x, d \sim Dom(X'), W, w)$
return $\frac{\sigma}{T}$

tradeoffs. For example, short-circuiting after finding a single witness set of size β that satisfies Definition 1 will save compute time, but may give a slightly higher or lower ρ score depending on whether the variables of interest are important under many counterfactual scenarios or only a few.

4.1 Approximating MeanRESP

Algorithm 1 is an exact algorithm that iterates over all possible scenarios to count where X' satisfies the definition of cause. When the state space is very large, due to either continuous variables or large discrete domains, we can use essentially the same algorithm adapted to sample witness set assignments using Monte Carlo sampling. Algorithm 4 approximates exact MeanRESP, and reproduces the exact algorithm in the limit. Sampling may be constrained along several dimensions independently, depending on the most expensive features of the problem. Here, we present in detail a novel sample-based algorithm to calculate responsibility scores. We then discuss its connection to the popular Shapley value-based attribution score. In subsequent sections, we will theoretically and empirically analyze this algorithm.

The main difference is that instead of going through all possible scenarios (i.e. $W \in \mathcal{P}(X \setminus X'), w \in \mathcal{D}(W), d \in \mathcal{D}(X')$) we sample different scenarios uniformly. The expression being estimated can be written as the following equation for RESP-UC:

$$E_{W\sim\mathcal{P}(X\setminus X'),w\sim\mathcal{D}(W),d\sim\mathcal{D}(X')}[\frac{D1}{1+\beta}(\phi(x)-\phi(x_m))] \tag{2}$$

For RESP-OC it can be written as:

$$E_{W\sim\mathcal{P}(X\setminus X'),w\sim\mathcal{D}(W),d\sim\mathcal{D}(X')}[\frac{\phi(x_p)}{1+\beta}(\phi(x)-\phi(x_m))] \tag{3}$$

It can be verified that this expression is the same as the following:

$$E_{W\sim\mathcal{P}(X\setminus X'),w\sim\mathcal{D}(W),d\sim\mathcal{D}(X')}[\frac{\phi(x_p)}{1+\beta}(\phi(x_p)-\phi(x_m))] \tag{4}$$

This rewrite provides us with insight into the connection between Shapley value and RESP. In particular, the Monte Carlo approximation of the expected Shapely value can be written as:

$$E_{W\sim\mathcal{P}(X\setminus X'),w\sim\mathcal{D}(W),d\sim\mathcal{D}(X')}[(\phi(x_p)-\phi(x_m))] \tag{5}$$

Intuitively, from equations 4, and 5 MeanRESP can be thought of as distanced weighted Shapely Value. Here, $1+\beta$ captures the difference in the original input x and x_m. $\phi_(x_p)$ captures the difference in original output $\Pi(x)$ and $\Pi(x_p)$.

Finally, when the action space is not continuous the definition of ϕ might not work as well due to the floating point values. Therefore, we can consider a softer version of ϕ as below:

$$\phi_{soft}(x_a) = e^{-\beta(|\Pi(x_a))-\Pi(x))|} \tag{6}$$

Here, $\beta \to \infty : \phi_{soft}(x_a) \to \phi(x_a)$. However, note that for ϕ_{soft} equations 2 and 3 are no longer equivalent.

5 Theoretical Analysis

Proposition 1. MeanRESP *Score* $\rho > 0$ *using the* RESP-UC *function implies actual cause according to Definition 1.*

Proof Sketch: MeanRESP Score $\rho > 0$ iff there exists at least one contingency (W, w) for which all the in Definition 1 is satisfied.

Proposition 2. *The false positive rate of sampled MeanResp is 0.*

Proof Sketch: $\rho > 0$ iff all the constraint of the Definition 1 is satisfied at least once.

Proposition 3. *The false negative rate of sampled Mean Resp with n sample is at most* $(1-(\rho^*(|W|_{max}+1)))^n$.

Proof Sketch The probability of not classifying X' as a weak cause i.e. false negative rate using 1 sample will be at most $1 - (\rho^*(|W|_{max} + 1))$. If samples are drawn independently then the false negative rate using n samples is at most $(1 - (\rho^*(|W|_{max} + 1)))^n$.

Proposition 4. *$P(|\rho - \rho^*| \geq \sqrt{\epsilon\rho^*}) \leq 2e^{-\epsilon n/3k}; k = W_{max} + 1$, n = number of samples. In words, probability that the estimated ρ deviate by $\sqrt{\frac{\epsilon}{\rho*}}$ is at most $2e^{-\epsilon n/3k}$.*

Proof Sketch According to Chernoff Bound:

$$P(|\frac{\rho n}{k} - \frac{\rho^* n}{k}| \geq \delta\frac{\rho^* n}{k}) \leq 2e^{-\delta^2\rho^* n/3k} \tag{7}$$

Setting $\delta = \sqrt{\frac{\epsilon}{\rho*}}$:

$$P(|\frac{\rho n}{k} - \frac{\rho^* n}{k}| \geq \frac{\sqrt{\epsilon\rho*}n}{k}) \leq 2e^{-\epsilon n/3k} \tag{8}$$

This is same as:

$$P(|\rho - \rho^*| \geq \sqrt{\epsilon\rho*}) \leq 2e^{-\epsilon n/3k} \tag{9}$$

Proposition 5. MeanRESP *score is upper-bounded by Shapley Value.*

Proof Sketch Since in Equation 3, $0 \leq \frac{\phi(x_p)}{1+\beta} \leq 1$, MeanRESP score will always be upper-bounded by shapely value when we consider Equation 4.

6 Empirical Analysis

In this section, we will first discuss several metrics for comparing feature rankings generated using exact methods to those generated using different forms of approximation. Then, we will use some of these metrics to empirically evaluate the sampling error of the two proposed approximate MeanRESP methods (UC and OC) in conjunction with the Shapley value [37]. Finally, we will examine the disagreement in feature rankings among these three methods. This disagreement will help us evaluate whether MeanRESP differs significantly from the widely adopted Shapley value, potentially warranting a human-subject study.

6.1 Environment Details

To conduct experiments in this section, we used three open-source environments designed for sequential decision-making. The initial two environments were obtained from the OpenAI Gym library [6]: Blackjack and Taxi. The Blackjack environment consists of 704 states and 2 actions, with each state being represented by 3 features. As for the Taxi environment, it comprises 500 states and 6 actions, with each state being represented by 4 features. Additionally, we employed the

Ground Truth	**N = 2000**	**N = 1000**	**N = 500**	**N = 50**
1. Vehicle-2_X	1. Vehicle-2_X	1. Vehicle-1_Y	1. Vehicle-1_Y	1. Vehicle-1_Y
2. Vehicle-1_Y	2. Vehicle-1_Y	2. Vehicle-2_X	2. Vehicle-2_X	2. Vehicle-3_X
3. Vehicle-3_Y	3. Vehicle-3_Y	3. Vehicle-3_Y	3. Vehicle-3_Y	3. Vehicle-Ego_Y
4. Vehicle-Ego_Y	4. Vehicle-Ego_Y	4. Vehicle-3_X	4. Vehicle-3_X	4. Vehicle-2_Y
5. Vehicle-3_X	5. Vehicle-3_X	5. Vehicle-2_Y	5. Vehicle-2_Y	5. Vehicle-2_X

Table 2: Example of the top $k = 5$ features identified as causes by the exact (ground truth) MeanRESP and approximate MeanRESP-OC methods after different numbers of samples ($N = 2000, 1000, 500, 50$). We observe that: a) After 50 samples, MeanRESP identifies most of the causal variables (4 out of 5). b) By 500 samples, the first 3 rankings match exactly with the exact method. c) After 2000 samples, the ranking completely matches the exact method. While the score estimates fluctuate with additional samples, highly influential variables (ranks 1, 2, and 3) are relatively easy to identify. Other weakly influential variables appear frequently but may not always be ranked correctly. We observe this trend of influential variables stabilizing early throughout our experiment.

highway-fast-v0 environment from the Highway-env library [17] (referred to as "Highway" hereafter). This environment encompasses 20 features (we exclude the features indicating *presence* from our experiment) within its feature space, each with continuous values. To facilitate our analysis, we discretized the feature domain into 20 equidistant points. Consequently, the total number of states in this environment amounted to 20^{20}. In this environment, the agent can select from 5 different discrete actions. Note that due to such a large state space, it is computationally infeasible to estimate exact MeanRESP. For the empirical analysis, we used a very large sample size of 10^5 to emulate exact MeanRESP For the Blackjack and Taxi environments, we employed value iteration [5] to compute the optimal policy, while for Highway, we utilized Deep Q-learning [24] to approximate an optimal policy.

6.2 Metrics

In this subsection, we will discuss several existing metrics used to compare feature rankings generated by exact methods with those generated using various forms of approximation. Some of these metrics rely solely on the contents of the explanations, others rely only on the relative rankings of feature sets, and some require ranking scores.

To evaluate the effectiveness of these metrics, we employed the Highway environment. In Table 2, we present snapshots of the top k most responsible variables for a given action outcome, illustrating examples of both the exact MeanRESP-OC method and various stages during the sampling process within the approximate MeanRESP-OC approach. The behavior of these metrics throughout the sampling process is summarized in Figures 1.

Ranking Only We can calculate the ranking of the features by sorting the ρ-values in descending order. Using Kendall's τ and Spearman's ρ, we can calculate a rank correlation coefficient to compare the ranks of features. This coefficient should increase towards 1 as we increase the number of samples used to estimate. Finally, one simple approach is to check if the two rankings are identical.

Responsibility Having access to the raw responsibility scores provides an opportunity for additional nuance in our metrics. Here, we present several options.

First, let us treat each feature in X as a 2D point. The x-value will be the true ρ-value, as given by the exact method, ρ^* for that set. The y-value will be the estimated ρ-value. As the number of samples increases, the slope of the least squares fit line on these 2D points should approach 1.

Second, we can use Pearson's r correlation factor to calculate the correlation between ρ^* and ρ. As the number of samples increases the correlation factor should reach 1.

Third, we take the top k features from both exact and approximate methods, sum the ρ^* values associated with the exact results, and call this ρ^*_k. Then, sum the ρ^*-values for approximate results and call this ρ_k^{approx}. The fraction $\frac{\rho_k^{approx}}{\rho_k^*}$ should approach 1 as number of samples increases.

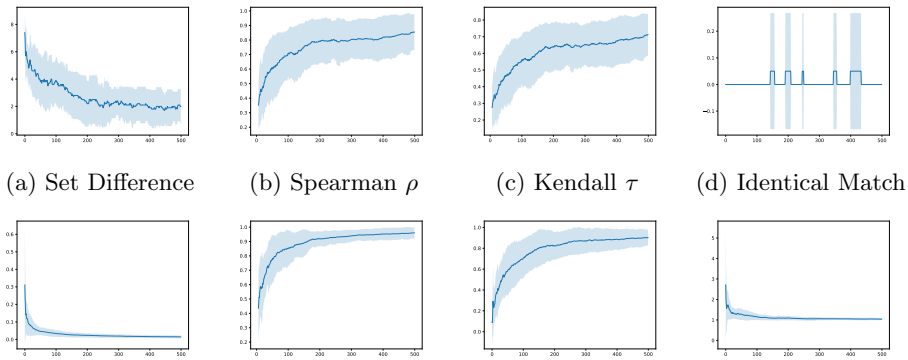
Finally, we can calculate the Euclidean distance between the vectors representing ρ^*- and ρ-values for every potential feature set. As the number of samples increases the distance should reach 0.

Feature Set Contents The previous metrics concern the relative importance assigned to different causes identified by MeanRESP or a similar algorithm. Here, we consider the presence or absence of information represented within the causal sets, since in many cases, the user will ultimately see only the causes and not their relative importance. If we create a set C^*_k that is the union of the top k features from the exact algorithm, and similarly define a set C_k that is the union of the top k features from the approximate algorithm, we then have a basis to understand what has been erroneously included or omitted. In this case, we propose simply finding the number of insertions and deletions required to make the sets identical or the edit distance. This number should approach 0 as the number of samples increases.

Discussion In the context of explanation generation, we argue that feature-set content is a more interpretable metric as it tells us exactly how many different factors will be communicated to the user. For sample error estimation, while all the metrics under responsibility are good candidates, we found no clear winner. We opted for using Euclidean distance for our experiments.

6.3 Sampling Error

In Figure 2 we show estimation error versus the number of samples used to estimate the score. We use the following estimation error:



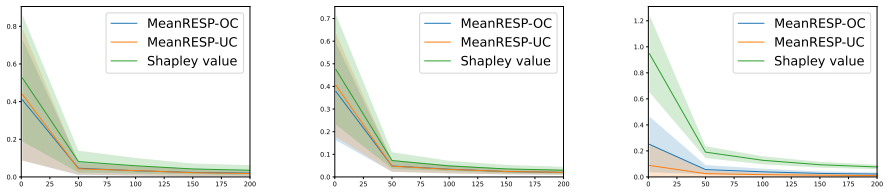
(a) Set Difference (b) Spearman ρ (c) Kendall τ (d) Identical Match

(e) Euclidean (f) Pearson r (g) Least Squares (h) Top K ratio

Fig. 1: Traces of all 8 metrics over time as they compare exact and approximate responsibility estimates. The solid blue line represents the mean (50 runs using 50 different states total), and the blue-shaded regions represent one standard deviation. Clearly, some metrics are more sensitive than others. Moreover, some shifts appear to be detected universally, for example, near 200 samples, while at other points some metrics respond to updated estimates while others do not. Notably, different is the boolean metric in (d) that checks whether the top $k = 5$ items in the set are identically ranked. The trace also shows us when the first time results become identical to the exact methods. Due to the Monte Carlo sampling, we see some oscillation. In addition, (a) is a version of edit distance, measuring how many insertions or deletions need to be made before the sets are identical. Here, both absolute and relative responsibility scores are irrelevant; only inclusion somewhere in the top k is captured.

$$\frac{1}{|S|}\sum_{s\in S}\sqrt{\sum_{i\in[1,|s|]}(\rho_i^*(s)-\rho_i(s))^2} \tag{10}$$

Here, $\rho_i^*(s)$ and $\rho_i(s)$ are the ground truth value and approximated value respectively for the $i-$th feature of state s. Note that this is equivalent to the average Euclidean distance. We use the average of 30 different evaluations of Equation 10 to create Figure 2. In both environments, we see MeanRESP-UC and MeanRESP-OC perform similarly. However, for the same amount of samples, we see 10%-70% more error in the estimation for the Shapley Value compared to both MeanRESP.



(a) Blackjack (b) Taxi (c) Highway

Fig. 2: Convergence rates for different attribution methods. Horizontal axes represent the number of samples taken, and vertical axes represent the absolute error in attribution value with respect to the exact solution. Color gradients represent one standard deviation.

6.4 Explanation Dissimilarity

Environment	OC vs. UC	OC vs. Shapley	UC vs. Shapley
BlackJack	0.05	0.15	0.13
Taxi	0.05	0.11	0.14
Highway	1.8	2.5	2.4

Table 3: The average Feature-set Difference among attribution methods. The numbers appear surprisingly high, considering the relatively small number of features in the problems and the relative similarity of the methods.

In this subsection, we show the ranking disagreement among MeanRESP-OC, MeanRESP-UC, and Shapley in Table 3. The ranking is created by sorting the features based on their attribution score in descending order and then selecting the top 33% of the features (for Taxi, the top feature, for Blackjack,

the top 2 features, and for Highway, the top 7 features). We then calculated pairwise explanation dissimilarity using the Feature-set Difference metric discussed previously. We generated explanations for 100 sampled states in each environment and reported the average Feature-set Difference. In all cases, we see that the disagreement between MeanRESP-OC and MeanRESP-UC is smaller than Shapley. Also, MeanRESP-OC is more similar to Shapley than MeanRESP-UC, in two out of three environments. These results suggest that there is a significant amount of difference in the explanations created by these methods, especially in larger environments. This motivates a potential future human subject study of explanation preference.

7 Conclusion

In summary, this study provides a comprehensive examination of MeanRESP, a framework for causal analysis of MDPs using structural causal models. The theoretical and empirical analyses shed light on crucial properties of approximate MeanRESP, including the convergence of error rates. Additionally, we introduce various metrics that contribute to a deeper understanding of the ranking generated by approximate MeanRESP. Moving forward, future research will involve conducting user preference studies to empirically evaluate the effectiveness of these methods.

Acknowledgments

This work was supported in part by the National Science Foundation grant number IIS-1954782.

References

1. Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E.D., Gutierrez, J.B., Kochut, K.: Text summarization techniques: A brief survey. arXiv preprint arXiv:1707.02268 (2017)
2. Bellman, R.: On the theory of dynamic programming. National Academy of Sciences of the United States of America **38**(8), 716 (1952)
3. Bertossi, L., Li, J., Schleich, M., Suciu, D., Vagena, Z.: Causality-based explanation of classification outcomes. arXiv preprint arXiv:2003.06868 (2020)
4. Bertram, J., Wei, P.: Explainable deterministic MDPs. arXiv preprint arXiv:1806.03492 (2018)
5. Bertsekas, D.P.: Dynamic programming and optimal control (1995)
6. Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., Zaremba, W.: OpenAI Gym. ArXiv **abs/1606.01540** (2016)
7. Chen, J.Y., Lakhmani, S.G., Stowers, K., Selkowitz, A.R., Wright, J.L., Barnes, M.: Situation awareness-based agent transparency and human-autonomy teaming effectiveness. Theoretical Issues in Ergonomics Science **19**(3), 259–282 (2018)
8. Chockler, H., Halpern, J.Y.: Responsibility and blame: A structural-model approach. Journal of Artificial Intelligence Research **22**, 93–115 (2004)

9. David Wong, E.: Understanding the generative capacity of analogies as a tool for explanation. Journal of research in science teaching **30**(10), 1259–1272 (1993)
10. Elizalde, F., Sucar, E., Noguez, J., Reyes, A.: Generating explanations based on Markov decision processes. In: Mexican International Conference on Artificial Intelligence. pp. 51–62. Springer (2009)
11. Halpern, J.Y., Pearl, J.: Causes and explanations: A structural-model approach. Part I: Causes. The British Journal for the Philosophy of Science **52**(3), 613–622 (2005)
12. Halpern, J.Y., Pearl, J.: Causes and explanations: A structural-model approach. Part II: Explanations. The British Journal for the Philosophy of Science **56**(4), 889–911 (2005)
13. Hayes, B., Shah, J.A.: Improving robot controller transparency through autonomous policy explanation. In: ACM/IEEE International Conference on Human-Robot Interaction (HRI). pp. 303–312 (2017)
14. Juozapaitis, Z., Koul, A., Fern, A., Erwig, M., Doshi-Velez, F.: Explainable reinforcement learning via reward decomposition. In: IJCAI/ECAI Workshop on Explainable Artificial Intelligence (2019)
15. Karimi, A.H., Schölkopf, B., Valera, I.: Algorithmic recourse: From counterfactual explanations to interventions. In: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency. pp. 353–362 (2021)
16. Khan, O., Poupart, P., Black, J.: Minimal sufficient explanations for factored Markov decision processes. In: International Conference on Automated Planning and Scheduling (ICAPS). vol. 19 (2009)
17. Leurent, E.: An environment for autonomous driving decision-making (2018)
18. Linegang, M.P., Stoner, H.A., Patterson, M.J., Seppelt, B.D., Hoffman, J.D., Crittendon, Z.B., Lee, J.D.: Human-automation collaboration in dynamic mission planning: A challenge requiring an ecological approach. Proceedings of the Human Factors and Ergonomics Society Annual Meeting **50**(23), 2482–2486 (2006)
19. Lucic, A., Haned, H., de Rijke, M.: Why does my model fail? Contrastive local explanations for retail forecasting. In: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency. pp. 90–98 (2020)
20. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. Advances in neural information processing systems **30** (2017)
21. Madumal, P., Miller, T., Sonenberg, L., Vetere, F.: Explainable reinforcement learning through a causal lens. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 2493–2500 (2020)
22. Mercado, J.E., Rupp, M.A., Chen, J.Y., Barnes, M.J., Barber, D., Procci, K.: Intelligent agent transparency in human-agent teaming for Multi-UxV management. Human Factors **58**(3), 401–415 (2016)
23. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence **267**, 1–38 (2019)
24. Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M.A.: Playing Atari with deep reinforcement learning. ArXiv **abs/1312.5602** (2013)
25. Molnar, C.: Interpretable Machine Learning. 2 edn. (2022), https://christophm.github.io/interpretable-ml-book
26. Mothilal, R.K., Sharma, A., Tan, C.: Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency. pp. 607–617 (2020)
27. Nashed, S.B., Mahmud, S., Goldman, C.V., Zilberstein, S.: Causal explanations for sequential decision making under uncertainty (2022)

28. Nisioi, S., Štajner, S., Ponzetto, S.P., Dinu, L.P.: Exploring neural text simplification models. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. pp. 85–91 (2017)
29. Panigutti, C., Perotti, A., Pedreschi, D.: Doctor XAI: An ontology-based approach to black-box sequential data classification explanations. In: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency. pp. 629–639 (2020)
30. Pouget, H., Chockler, H., Sun, Y., Kroening, D.: Ranking policy decisions. arXiv preprint arXiv:2008.13607 (2020)
31. Russell, J., Santos, E.: Explaining reward functions in Markov decision processes. In: Thirty-Second International FLAIRS Conference (2019)
32. Scharrer, L., Bromme, R., Britt, M.A., Stadtler, M.: The seduction of easiness: How science depictions influence laypeople's reliance on their own evaluation of scientific information. Learning and Instruction **22**(3), 231–243 (2012)
33. Shapley, L.S., et al.: A value for n-person games (1953)
34. Srikanth, N., Li, J.J.: Elaborative simplification: Content addition and explanation generation in text simplification. arXiv preprint arXiv:2010.10035 (2020)
35. Stubbs, K., Hinds, P.J., Wettergreen, D.: Autonomy and common ground in human-robot interaction: A field study. IEEE Intelligent Systems **22**(2), 42–50 (2007)
36. Sukkerd, R., Simmons, R., Garlan, D.: Tradeoff-focused contrastive explanation for mdp planning. In: 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN). pp. 1041–1048. IEEE (2020)
37. Štrumbelj, E., Kononenko, I.: Explaining prediction models and individual predictions with feature contributions. Knowledge and Information Systems **41**, 647–665 (2014)
38. Wang, N., Pynadath, D.V., Hill, S.G.: The impact of POMDP-generated explanations on trust and performance in human-robot teams. In: International Conference on Autonomous Agents and Multiagent Systems (AAMAS). pp. 997–1005 (2016)
39. Zhang, Y., Liao, Q.V., Bellamy, R.K.: Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency. pp. 295–305 (2020)