Formal Explanations of Neural Network Policies for Planning

Renee Selvey[1] , **Alban Grastien**[1] and **Sylvie Thiébaux**[2,1]
[1]School of Computing, The Australian National University
[2]LAAS-CNRS, ANITI, Université de Toulouse
{Renee.Selvey, Alban.Grastien}@anu.edu.au, Sylvie.Thiebaux@laas.fr

Abstract

Deep learning is increasingly used to learn policies for planning problems, yet policies represented by neural networks are difficult to interpret, verify and trust. Existing formal approaches to post-hoc explanations provide concise reasons for a *single* decision made by an ML model. However, understanding planning policies requires explaining *sequences* of decisions. In this paper, we formulate the problem of finding explanations for the sequence of decisions recommended by a learnt policy in a given state. We show that, under certain assumptions, a minimal explanation for a sequence can be computed by solving a number of single decision explanation problems which is linear in the length of the sequence. We present experimental results of our implementation of this approach for ASNet policies for classical planning domains.

1 Motivation

Deep learning has become the method of choice in the areas of AI that focus on perception, and is rapidly gaining traction in other areas that have traditionally been strongholds of reasoning, search, and combinatorial optimisation. In automated planning for instance, new work has emerged that uses deep learning to learn policies and heuristics in a range of planning domains. We refer the reader to [Toyer *et al.*, 2018; Groshev *et al.*, 2018; Garg *et al.*, 2020; Zhang and Geißer, 2022; Karia and Srivastava, 2022] for examples of work aiming at learning policies for planning domains, and to [Shen *et al.*, 2020; Ferber *et al.*, 2020; Karia and Srivastava, 2021; Ferber *et al.*, 2022; Gehring *et al.*, 2022] as representatives of work on learning heuristics to guide the search for a plan.

As the use of deep learning becomes more widespread in planning, the need to understand the solutions it produces becomes more pressing. Policies represented by neural networks are notoriously opaque, and difficult to understand, verify, and trust [Toyer *et al.*, 2020; Vinzent *et al.*, 2022]. Decisions could also rely on unexpected reasons and preferences/bias hidden in the data used for training. The neural network could also be prejudiced against certain groups of people and give different advice depending on gender, race, etc. [Darwiche and Hirth, 2020]. At the minimum, one would like to be able to explain why a particular course of action was recommended by the policy – by identifying the properties of the state of the world prior to the execution of the policy which led to that recommendation – so as to help the user decide whether this recommendation should be trusted. This is the problem addressed in this paper.

1.1 Running Example

Throughout the paper, we use the example of Yvette who needs to take a turnpike to get to her final destination where she will spend a couple of weeks holidays. The turnpike requires to either purchase a weekly pass online or pay with cash at a toll gate. The pass is expensive and should not be taken unless one expects to use the turnpike multiple times in a single week. Yvette does not currently have a pass or cash. Hence, the policy prescribes to drive to an ATM, withdraw some cash, drive to the toll gate, pay the toll, and drive to the destination.

Why choose this course of action? Firstly because i) Yvette is on the side of the turnpike opposite to her destination. Indeed this course of action would work in any state in which condition i) holds. However, this fails to explain *why* Yvette should not directly go to the toll gate. This is because ii) she has no pass and iii) she has no cash. A correct explanation should therefore include all three conditions.

If, instead, Yvette did hold a pass, then the policy would skip the visit to the ATM. The explanation would then mention the fact that Yvette holds a pass but would not mention her lack of cash as the policy would still have prescribed this course of action even if she had cash.

As mentioned earlier, explanations can also expose unexpected reasons for decisions. For instance, the explanation could include the proposition iv) it is sunny, which implicitly means that the policy would have decided differently if it was not. It is questionable whether iv) is relevant in this context – it might indicate that the neural network has learnt that you would not want to make a trip to the ATM under the rain.

1.2 Existing Work

The recent work on explainable planning has been focusing on a different set of problems in a different setting, in particular on model reconciliation and contrastive explanations for conventional model-based planning [Chakraborti *et al.*, 2020]. In model reconciliation, the aim is to generate explanations allowing a human user to update their model of the planning problem to make it consistent with the plan produced by a planning agent [Chakraborti *et al.*, 2017; Sreedharan *et al.*, 2021; Vasileiou *et al.*, 2022]. This latter question is concerned with the properties of the planning model, rather than those of the policy. The second prominent line of research in the explainable planning literature is the generation of contrastive explanations outlining why the planner chose a course of action over others within the space of possible plans [Eifler *et al.*, 2020; Kasenberg *et al.*, 2020; Krarup *et al.*, 2021]. These works are concerned with understanding the space of possible decisions and their respective merits, rather than a particular policy.

Therefore, as a starting point, we instead turn to prior work concerned with explaining deep learning and other data-driven models for *classification* tasks. Existing approaches typically either compute simpler models that locally approximate the classifier's behavior [Ribeiro *et al.*, 2016; Lundberg and Lee, 2017], or identify sufficient conditions on the inputs that led the neural network to produce a particular output [Ribeiro *et al.*, 2018; Ignatiev *et al.*, 2019]. One approach falling into the latter class is to compute *abductive* explanations that are *minimal* sufficient conditions for the decision. This has the advantage of providing formal guarantees of soundness and non-redundancy [Ignatiev *et al.*, 2019; Darwiche and Hirth, 2020; Marques-Silva and Ignatiev, 2022].

However, the above approaches are designed to explain a single decision, whereas understanding the recommendations of a planning policy requires explaining why a particular *sequence* of decisions was made. The latter is more challenging as it involves reasoning about repeated applications of the policy network and about the successive changes they induce in the state of the world in which the policy is executed.

1.3 Contribution

In this paper, we extend abductive explanations from single to sequential decisions. We restrict ourselves to classical planning policies and explanations of why the policy makes a certain sequence of decisions from a given state. We formally define explanations for a sequence of decisions, and show that, under certain assumptions, the problem of finding an explanation for the sequence can be decomposed into that of finding explanations for the individual decisions in the sequence. We provide an algorithm that exploits this decomposition to compute a minimal explanation for a sequence by making a number of consistency tests pertaining to individual decisions that is at most linear in the length of the sequence and in the number of state variables. We then discuss the implementation of our approach to explain policies represented by Action Schema Networks (ASNets) [Toyer *et al.*, 2020] and report on its performance on sparse ASNet policies for classical planning domains. We conclude by discussing the limits of our work and possible extensions.

2 Background

We start by introducing the type of planning problems we consider, their representation, and our notations.

Many of the recent works on deep learning for planning assume that the model of the planning domain is available to the learner. We assume that it is also available to the explainer. Here we represent the classical planning instance $I = \langle X, A, g \rangle$ under consideration using the SAS$^+$ formalism [Bäckström and Nebel, 1995]. X is a set of finite-domain *state variables*, where D_x is the *domain* of variable x. A *partial state* (or partial valuation) s is an assignment of value to a subset $X_s \subseteq X$ of the variables such that $s[x] \in D_x$ for $x \in X_s$. If $X_s = X$ then we say that s is a *state*; we write S for the set of states. A *proposition* $(x = v)$ is a partial valuation assigning a value to a single variable.

The goal g is a partial state. Given two partial states s and s', we write $s \subseteq s'$ when $s[x] = s'[x]$ for all $x \in X_s$. A *completion* of a partial state s is a state s' such that $s \subseteq s'$. The result of applying a partial valuation e to a partial state s is the partial state $s \oplus e$ over $X_s \cup X_e$ defined by $(s \oplus e)[x] = e[x]$ if $x \in X_e$ and $(s \oplus e)[x] = s[x]$ if $x \in X_s \backslash X_e$. We also define the binary operator $\ominus$ over partial states: $s \ominus e$ is the restriction of s to the variables $X_s \backslash X_e$. For a variable $x \in X_s$ will write $s - x$ as an abbreviation for $s \ominus (x = s[x])$.

A is the set of *actions*. Action $a \in A$ is characterised by two partial valuations representing its *precondition* $\text{pre}(a)$ and its *effect* $\text{eff}(a)$, respectively. We say that the action is *applicable* in a state $s \in S$ iff $\text{pre}(a) \subseteq s$ and write $A(s) \subseteq A$ for the subset of actions applicable in s. Moreover, given a partial state s and an action a, the *progression* of s through a is the partial state $\text{prg}_a(s) = s \oplus \text{eff}(a)$. Note that if s is a state and $a \in A(s)$ then $\text{prg}_a(s)$ is the state resulting from applying a in s.

A *policy* for the planning instance is a function $\pi : S \mapsto A$ mapping states to applicable actions, i.e. $\pi(s) \in A(s)$. We define the *n-long trajectory* $\tau^n_\pi(s)$ *induced* by π from state s as follows: $\tau^n_\pi(s) = s_1 \xrightarrow{a_1} \ldots \xrightarrow{a_n} s_{n+1}$ such that $s_1 = s$ and for all $0 < i \leqslant n$, $a_i = \pi(s_i)$ and $s_{i+1} = \text{prg}_{a_i}(s_i)$. Finally, the *$n$-long sequence of actions recommended by* π in s is $\pi^n(s) = a_1, \ldots, a_n$, where the a_is are the successive actions in $\tau^n_\pi(s)$.

Given a planning instance I, a policy π for I, an *initial state* s, and an integer n, our problem is to explain why π recommended the sequence of actions $\pi^n(s)$ in s. The explanation is meant to shed light on the appropriateness of the recommendation.

The definition of explanations presented in the next section relies on the fact that the policy will recommend the same sequence of actions for certain states. To ensure that for any state s and any length n, a policy can recommend an n-long sequence from s, we make the following assumptions. First, we assume that there is no terminal state, i.e., $A(s) \neq \varnothing$ for all states. This is not a restriction as a default action could be to do nothing. Second, we assume that A includes a special goal action a_g which has no effect and which is only applicable in goal states (this will act as a marker of the goal being reached), and that π is a total function on S such that $\pi(s) = a_g$ iff $g \subseteq s$. This is purely for convenience as policies generally check whether a goal state has been reached before computing the next action. With these assumptions, our theory applies uniformly to any trajectory, regardless of whether it reaches the goal.

3 Explanations of Neural-Network Policies

An explanation of a decision (or sequence of decisions) in a state is a condition on this state that led to this decision being made: the decision was made because the condition was satisfied in this state. Said differently, the same decision would have been made in any other state that satisfies this condition. We could allow arbitrary conditions; e.g. the explanation could be the logical formula that describes exactly all the states in which this decision would be taken. However, such an explanation would not be very helpful. We aim instead for a 'simple' explanation, that is, an explanation that mentions as few propositions as possible and has the simple structure of a conjunction.

We use a definition similar to that of [Marques-Silva and Ignatiev, 2022]. An explanation is a partial state that entails the decision; in logic, this is known as an implicant.

Definition 1. *An* explanation of a single decision a for policy π *is a partial state* $\mathbf{z}$ *such that* π *yields the same decision for all completions of* $\mathbf{z}$*:*

$$\forall s \in S. \quad (\mathbf{z} \subseteq s) \implies \pi(s) = a.$$

When s completes $\mathbf{z}$, we say that $\mathbf{z}$ *explains* decision a in s.

Our goal is not to explain just the first decision of the policy, but the complete sequence of decisions. While the first decision was based on the initial state, later decisions were made based on the later states. These states, however, are fully determined by the initial state and the actions taken. Using the planning model, it is therefore possible to trace the sufficient condition that led to the full sequence of actions back to the initial state.

Definition 2. *An* explanation of the n-long sequence of decisions $a_1, \ldots, a_n$ for a policy π *is a partial state* $\mathbf{z}$ *such that* π *yields the same sequence of decisions for all completions of* $\mathbf{z}$*:*

$$\forall s \in S. \quad (\mathbf{z} \subseteq s) \implies \pi^n(s) = a_1, \ldots, a_n.$$

Similarly as before, when s completes $\mathbf{z}$, we say that $\mathbf{z}$ *explains* the sequence of decisions $a_1, \ldots, a_n$ in s. We note that if $a_n = a_g$ is the goal action, then the sequence $a_1, \ldots, a_{n-1}$ leads all completions of $\mathbf{z}$ to a goal state since π only recommends applicable actions and a_g is only applicable in a goal state. The sequence of actions recommended by the policy might not lead to the goal; in this case, the loop of the infinite trajectory induced by the policy does not occur at a goal state. If one wants to compute an explanation for this infinite sequence, it is possible to use Definition 2 with $n = L \times |S|$ where L is the length of the loop and $|S|$ the total number of states: if $\mathbf{z}$ explains $\pi^n(s)$, then it explains $\pi^{n+k}(s)$ for all $k \geqslant 0$. It may be possible to derive better bounds.

It should be clear that there can be multiple explanations in the same state. For instance, in our running example, Yvette needs either a pass or some cash to take the turnpike. Both the fact that she has a pass and the fact that she has cash would be acceptable explanations for driving directly to the toll gate. In a state where she has a pass and cash, these two explanations are therefore suitable. We also note that explanations enjoy monotonic properties: if $\mathbf{z}$ is an explanation, any superset of $\mathbf{z}$ is an explanation. In particular, the complete initial state is an explanation, although hardly a useful one.

Our goal is to compute the 'best' explanation for the sequence of decisions made from our initial state. Specifically, we want to compute a subset-minimal explanation (akin to a *prime* implicant in logic), i.e., an explanation $\mathbf{z}$ such that no strict subset of $\mathbf{z}$ is an explanation. Therefore, seeing a variable that should not be relevant in a minimal explanation should raise questions about the policy, while this phenomenon is unsurprising in a non-minimal explanation. Minimal explanations provide additional benefits: all variables mentioned in the explanation are required, in the sense that if any were removed, the resulting partial state would no longer be an explanation. While there can be multiple minimal explanations for the same sequence of decisions, we only consider the problem of finding one of them.

Definition 3. *Given a policy* π*, an integer* n*, and a state* s*, the* minimal explanation problem *is to find a minimal partial state that explains the sequence of decisions* $\pi^n(s)$ *in* s*.*

Ignatiev *et al.* [2019] have shown how to compute explanations for single decisions. The policy is translated into a set of constraints C_π over a set of variables that includes the state variables X which are the input to the policy, and the variable y which represents its output. The models of C_π are exactly all the pairs $\langle s', y \rangle$, $s' \in S$, $y = \pi(s')$. If π is represented by a neural network, C_π can be formulated as a set of sat modulo theory or mixed-integer programming constraints (see Section 5). In order to decide whether $\mathbf{z}$ is an explanation for a decision a, the two constraints that s' completes $\mathbf{z}$ and $\neg d$ are added to the set where $d \equiv (y = a)$. If the resulting set of constraints is consistent, then there exists a state s' that completes $\mathbf{z}$ and yields a decision different from a; hence, $\mathbf{z}$ does not explain a. Otherwise, all states that complete $\mathbf{z}$ lead to decision a, and $\mathbf{z}$ explains a:

$$\mathbf{z} \text{ explains } a \;\Leftrightarrow \neg \mathsf{CO}(C_\pi \wedge \bigwedge_{x \in X_\mathbf{z}} (s'[x] = \mathbf{z}[x]) \wedge \neg d). \quad (1)$$

Using monotonicity, it is then possible to greedily search for a minimal explanation: we start with an existing explanation $\mathbf{z}$, for instance the current state s, and repeatedly test whether for some variable $x \in X_\mathbf{z}$, $\mathbf{z}' := \mathbf{z} - x$ remains an explanation. If $\mathbf{z}'$ indeed explains a, we replace $\mathbf{z}$ with it; otherwise, we go on with $\mathbf{z}$. We test each state variable exactly once. The resulting partial state is a minimal explanation.

4 Computing a Minimal Explanation

We now turn to the problem of computing a minimal explanation for a sequence of decisions.

4.1 Naive Algorithm

For completeness, we first consider a naive algorithm, illustrated in Figure 1. We expect that this algorithm will only be practical for short sequences as it requires testing the consistency of constraint sets involving too many variables.

Similarly as in the single decision case, the idea of the algorithm is to build a single set of constraints which is consistent iff a specified partial state does not explain a specified sequence of decisions $a_1, \ldots, a_n$. Given an initial state s'_1, we define the set of constraints $C_{a_1,\ldots,a_n}$ which computes the

$$\begin{tikzcd}
s'_1 \arrow[r, "\mathrm{prg}_{a_1}"] \arrow[d, "\pi"] & s'_2 \arrow[r, "\mathrm{prg}_{a_2}"] \arrow[d, "\pi"] & s'_3 \arrow[r, "\mathrm{prg}_{a_3}"] \arrow[d, "\pi"] & \cdots \arrow[r, "\mathrm{prg}_{a_{n-1}}"] \arrow[d, "\pi"] & s'_n \arrow[d, "\pi"] \\
y_1 \arrow[drr] & y_2 \arrow[dr] & y_3 \arrow[d] & \cdots \arrow[dl] & y_n \arrow[dll] \\
 & & d & &
\end{tikzcd}$$

Figure 1: Graphical representation of $C_{a_1,\ldots,a_n}$, the set of constraints used to determine whether a partial state is an explanation. Nodes of the graph are sets of variables. Arrows represent constraints defined such that the target variables are a function of the source variables.

states $s'_1 \xrightarrow{a_1} s'_2 \xrightarrow{a_2} \ldots \xrightarrow{a_{n-1}} s'_n$ reached by applying the successive actions as well as the policy decisions $y_i = \pi(s'_i)$ in each of these states; we then compare the y_is with the a_i:

$$\begin{array}{ll}
s'_{i+1} = \mathrm{prg}_{a_i}(s'_i) & \forall i \in \{1, \ldots n-1\} \\
y_i = \pi(s'_i) & \forall i \in \{1, \ldots n\} \\
d_i = (y_i = a_i) & \forall i \in \{1, \ldots n\} \\
d = \bigwedge_{i=1}^{n} d_i &
\end{array}$$

Finally, we add the constraints that s'_1 should complete $\mathbf{z}$ and that d should be false. The resulting set of constraints,

$$C_{a_1,\ldots,a_n} \wedge \bigwedge_{x \in X_{\mathbf{z}}} (s'_1[x] = \mathbf{z}[x]) \wedge \neg d,$$

is consistent iff there exists a state s'_1 that completes $\mathbf{z}$ for which π generates a different sequence of decisions than the specified one; i.e., $\mathbf{z}$ does not explain $a_1, \ldots, a_n$. As before, starting with $\mathbf{z} = s_1$, one can then greedily try to remove propositions to compute a minimal explanation.

However in this set of constraints, the variables include n duplicates of the set of state variables X, which are needed to represent the successive states $s'_1 \ldots s'_n$, as well as the much larger set of variables required to represent n duplicates of the computation of π (via the constraints C_π mentioned above). As shown in our experiments in Section 6, this makes testing consistency impractical beyond short sequences.

4.2 Forward Decomposition

We now show that it is possible to decompose the explanation so that it is easier to compute a single minimal explanation. This decomposition leads to an algorithm that need only solve a number of single decision explanation problems that is in the worst case linear in the length of the sequence. In the following, $\circ$ is the function composition.

Theorem 1. *Let π be a policy, s be a state, and let $\tau^n_\pi(s) = s_1 \xrightarrow{a_1} \ldots \xrightarrow{a_n} s_{n+1}$ be the n-long trajectory induced by π from s. Let $\mathbf{z} \subseteq s$ be a partial state and let $\mathbf{z}_1, \ldots, \mathbf{z}_n$ be a sequence of partial states defined by $\mathbf{z}_{i+1} = prg_{a_i} \circ \cdots \circ prg_{a_1}(\mathbf{z})$. Then $\mathbf{z}$ explains $\pi^n(s)$ iff for all steps i, $\mathbf{z}_i$ explains a_i in s_i.*

The proof of Theorem 1 is in Appendix A. Theorem 1 gives us a clear procedure to verify whether a partial state $\mathbf{z}$ is an explanation, which is to compute the partial states $\mathbf{z}_i$ resulting from applying the actions from $\mathbf{z}$ (note: $\mathbf{z}_{i+1} = \mathrm{prg}_{a_i}(\mathbf{z}_i)$) and to verify whether each $\mathbf{z}_i$ explains a_i.

Algorithm 1 Computing a minimal explanation for a sequence of decisions.

```
1:  procedure MINIMALEXPLANATION($I, \pi, n, s$)
2:      $s_1 \xrightarrow{a_1} \ldots \xrightarrow{a_n} s_{n+1} = \tau^n_\pi(s)$
3:      $\mathbf{z}_1 := s$
4:      for $i \in \{1, \ldots, n-1\}$ do
5:          $\mathbf{z}_{i+1} := \mathrm{prg}_{a_i}(\mathbf{z}_i)$
6:      for $x \in X$ do
7:          $explains := True$
8:          for $i \in \{1, \ldots, n\}$ do          ▷ Testing condition of Th. 1
9:              $\mathbf{z}_i := \mathbf{z}_i - x$
10:             if $\neg(\mathbf{z}_i$ explains $a_i)$ then          ▷ Eq. 1
11:                 $explains := False$
12:                 break
13:             if $x \in X_{\mathrm{eff}(a_i)}$ then
14:                 break          ▷ Can stop now (cf. Eq. 2)
15:         if $\neg explains$ then          ▷ Must reinsert $x$
16:             for $j \in \{1, \ldots, i\}$ do $\mathbf{z}_j := \mathbf{z}_j \oplus (x = s_j[x])$
17:     return $\mathbf{z}_1$
```

In the algorithm that we propose next, we use the additional property: for any variable x, $\mathrm{prg}_{a_i} \circ \cdots \circ \mathrm{prg}_{a_1}(\mathbf{z} - x) =$

$$\begin{cases} \mathrm{prg}_{a_i} \circ \cdots \circ \mathrm{prg}_{a_1}(\mathbf{z}) & \text{if } x \in X_{\mathrm{eff}(a_i)} \cup \ldots \cup X_{\mathrm{eff}(a_1)} \\ \mathrm{prg}_{a_i} \circ \cdots \circ \mathrm{prg}_{a_1}(\mathbf{z}) - x & \text{otherwise.} \end{cases} \quad (2)$$

The first case is useful because it means $\mathrm{prg}_{a_i} \circ \cdots \circ \mathrm{prg}_{a_1}(\mathbf{z} - x)$ explains a_{i+1} iff $\mathrm{prg}_{a_i} \circ \cdots \circ \mathrm{prg}_{a_1}(\mathbf{z})$ does; so as soon as variable x appears in the effects of a_i, we will be able to ignore the condition of Theorem 1 that z_k should explain a_k in s_k for all $k > i$. The second case is useful because it makes it easy to compute $\mathrm{prg}_{a_i} \circ \cdots \circ \mathrm{prg}_{a_1}(\mathbf{z} - x)$ from $\mathrm{prg}_{a_i} \circ \cdots \circ \mathrm{prg}_{a_1}(\mathbf{z})$.

Our procedure is described in Algorithm 1. In Lines 3–5, the algorithm initialises the variables $\mathbf{z}_i$, mirroring the variables from Theorem 1 for explanation $\mathbf{z} := s$. Line 6 starts the loop in which each variable x is tentatively removed from the explanation. The variable $explains$ will record whether $\mathbf{z} - x$ explains the sequence; until proven otherwise, it is assumed it does. The inner loop from Line 8 verifies the condition of Theorem 1. After $\mathbf{z}_i$ is updated (using the second case of Eq. 2), the consistency check (Eq. 1) is performed on Line 10. If $\mathbf{z}_i$ does not explain action a_i, the inner loop is stopped; the loop from Line 16 will then undo the update. Otherwise, the inner loop moves to the next step $i + 1$ *except* if variable x is mentioned in the effects of a_i in which case the loop can be stopped (first case of Eq. 2).

This algorithm requires $O(n \times m)$ calls to the consistency solver where n is the length of the sequence of m is the number of state variables.

We illustrate this algorithm with our running example. We assume Yvette uses the policy described in Figure 2. We note that this policy recommends sub-optimal actions (e.g., driving to the ATM and then purchasing a pass online when it is raining), which our approach can also handle. The initial state, and so the initial explanation, is $\mathbf{z} = \{$*Friday*, *Sunny*, *NoCash*, *NoPass*, *AtHome*$, \ldots\}$.

1. If has passed the TollGate, has a pass, or has some cash, drive towards the destination.
2. Otherwise, if not at the ATM, drive towards the ATM.
3. Otherwise, if it is sunny, withdraw cash.
4. Otherwise, purchase a pass online.

Figure 2: The policy given to Yvette (without goal action a_g)

Algorithm 1 starts by removing *Friday* from the explanation, leading to $\mathbf{z}' = \mathbf{z}'_1 = \mathbf{z} \ominus \{\textit{Friday}\}$. We find that it is impossible to instantiate the only free variable (day of the week) in such a way that a decision different from *DriveToATM* is taken; in other words, $\mathbf{z}'_1$ explains the first decision. Progressing $\mathbf{z}'_1$ gives us $\mathbf{z}'_2 = \{\textit{Sunny}, \textit{NoCash}, \textit{NoPass}, \textit{AtATM}, \dots\}$. This partial state also explains the second decision (*WithdrawCash*), and so on until the end of the sequence. Therefore, $\mathbf{z}'$ explains the whole sequence.

Then, Algorithm 1 removes *Sunny* from the explanation. The partial state is therefore $\mathbf{z}'' = \mathbf{z}''_1 = \mathbf{z}' \ominus \{\textit{Sunny}\}$. Partial state $\mathbf{z}''$ explains the first decision (*DriveToATM*). Progressing $\mathbf{z}''_1$ gives us $\mathbf{z}''_2 = \{\textit{NoCash}, \textit{NoPass}, \textit{AtATM}, \dots\}$. This time, we find that a state in which Yvette is at the ATM and the weather is rainy will yield a different decision (*PurchasePass*) than the second decision (*WithdrawCash*). Therefore, $\mathbf{z}''$ does not explain the sequence and *Sunny* is added back to the explanation. Similarly, the algorithm would then try and fail to remove the subsequent propositions.

5 Implementation

We use Algorithm 1 to explain the recommendations of ASNet policies [Toyer *et al.*, 2018] for classical planning domains. This requires implementing the consistency test on line 10 of the algorithm for ASNets, which have a complex structure and nonlinear activation functions. Our encoding, presented below, is supported by mixed integer programming (MIP) solvers such as Gurobi [Gurobi Optimization, LLC, 2023], and builds on the principles underlying the MIP encodings of simpler networks used in verification or single-decision explanations (see e.g. [Fischetti and Jo, 2018]). Clearly the framework presented in the previous section is general enough to apply to any policy representation for which the consistency test can be implemented. We chose ASNets as they can learn *sparse* policies which are easier to scrutinise for testing purposes, and make the consistency test computationally feasible, at least for small problems.

5.1 ASNets

ASNets are neural networks dedicated to planning, trained to produce policies that apply to any problem instance from a given planning domain modelled in (P)PDDL [Younes and Littman, 2004]. An ASNet consists of L alternating *action* and *proposition layers*, beginning and ending with an action layer. In each action layer (resp. proposition layer), each action (resp. proposition) of the planning instance is represented by an action *module* (resp. proposition module). Modules in one layer are connected to *related* modules in the next layer, where a proposition p and action a are related, written $R(a, p)$ iff p appears in the precondition or effect of a. This enables relevant information to be efficiently propagated.

An ASNet is capable of learning policies that generalise to problem instances of different size from the same domain thanks to a weight sharing mechanism whereby the action (resp. proposition) modules from the same layer that are ground instances of the same action schema (resp. the same predicate), have the same weights. Here we omit details such as the use of skip connections and the more complex definition of relatedness in [Toyer *et al.*, 2020], but our implementation supports them.

Action Layers. At layer l, excluding the first and last layer, the module for action a takes as input a vector u_a^l which is constructed by enumerating the propositions $p_1, \dots, p_M$ related to a, and concatenating the outputs $\psi_{p_1}^{l-1}, \dots, \psi_{p_M}^{l-1}$ of these propositions' modules from the previous layer. That is $u_a^l = [\psi_{p_1}^{l-1} \dots \psi_{p_M}^{l-1}]^T$. The output of the module is the vector $\phi_a^l = f(W_a^l u_a^l + b_a^l)$ where W_a^l is a weight matrix, b_a^l a bias vector, and f a nonlinearity. ASNets use exponential linear units (ELU), i.e. $f(x) = x$ if $x \geqslant 0$ and $e^x - 1$ otherwise. Note that except for the final layer, the output vectors of all modules in the network have the same dimension d.

Proposition Layers. At layer l, the module for proposition p takes as input a vector v_p^l constructed by enumerating the actions that are related to the proposition, pooling from the outputs of their modules at layer l if they share the same action schema, and concatenating the results. That is $v_p^l = [\text{pool}(\{\phi_a^l \mid \text{op}(a) = o_1 \wedge R(a,p)\}) \dots \text{pool}(\{\phi_a^l \mid \text{op}(a) = o_S \wedge R(a,p)\})]^T$, where pool represents max-pooling and op(a) is the action schema of action a. Similarly as for action modules, the output of the proposition module is the vector $\psi_p^l = f(W_p^l v_p^l + b_p^l)$.

First Layer. The input to an action module of the first layer are the boolean values (0 or 1) of all its related propositions $p_1, \dots, p_M$ in the current state s, booleans indicating whether each of these propositions is true in the goal, and a boolean indicating whether the action is applicable in s. That is the input vector is $u_a^1 = [\sigma\ \gamma\ \epsilon]^T$, where $\sigma \in \{0,1\}^M$ with $\sigma_j = 1$ iff $p_j \in s$, $\gamma \in \{0,1\}^M$ with $\gamma_j = 1$ iff $p_j \in g$, and $\epsilon = 1$ iff $a \in A(s)$.

Last Layer. The output of an action module in the final layer $l = L$ is a single logit ϕ_a^L representing the unnormalised probability that this action should be taken in the current state s given as input to the network. When, as in this paper, a deterministic policy is sought, the applicable action $a \in A(s)$ with maximum ϕ_a^L is selected by the policy, breaking ties consistently.

5.2 MIP Encoding

The main purpose of the encoding is to answer consistency queries where only *some* of the inputs to the ASNet are given, and its output is constrained. The key decision variables in the MIP model fall into 3 categories: variables representing the network inputs, its outputs, and the action and proposition modules. As is well known from other MIP encodings of neural networks, one also needs auxiliary variables to encode the activation functions. Parameters include the weight matrices and bias vectors described above. The MIP has no objective function since we are only testing for consistency.

Policy Inputs. We encode each element in the input of an ASNet (current state propositions, goal proposition, applicable actions) using the following binary variables: S_p is true iff proposition p is true in the current state, G_p is true iff proposition p is true in the goal, and E_a is true iff a is applicable in the current state. Since a key feature of the MIP model is to allow specifying *partial* states as input, we must add constraints capturing when actions are applicable: action a is applicable when $S_p = 1$ for all $p \in \text{pre}(a)$.

$$\begin{array}{ll} E_a \leqslant S_p & \forall a \in A, \forall p \in \text{pre}(a) \\ E_a \geqslant 1 - |\text{pre}(a)| + \sum_{p \in \text{pre}(a)} S_p & \forall a \in A \end{array}$$

Moreover, ASNets only support *boolean* propositions $p \equiv (x = v)$ for $x \in X$ and $v \in D_x$. To encode SAS^+ state variables, we add mutex constraints ensuring that at most one proposition assigning a value to a given variable can be true.

$$\begin{array}{ll} \sum_{v \in D_x} S_{(x=v)} \leqslant 1 & \forall x \in X \\ \sum_{v \in D_x} G_{(x=v)} \leqslant 1 & \forall x \in X \end{array}$$

Action Modules. To encode action modules, we define the MIP variables $(\text{PAC}^l_a)_i$ representing the ith element of the pre-activation vector (omitting the non-linearity f) of the module for action a in layer l, and $(\text{OUT}^l_a)_i$ representing the ith element of the output vector ϕ^l_a of this module. For an action a with related propositions $p_1, \ldots, p_M$, we have, taking $j = (m-1)d + k$

$$(\text{PAC}^l_a)_i = \sum_{m=1}^{M} \sum_{k=1}^{d} (W^l_a)_{i,j} \cdot (\text{OUT}^{l-1}_{p_m})_k + (b^l_a)_i \text{ for } l > 1$$

$$(\text{PAC}^1_a)_i = \sum_{m=1}^{M} (W^1_a)_{i,m} \cdot S_{p_m} + (W^1_a)_{i,M+m} \cdot G_{p_m} + (W^1_a)_{i,2M+1} \cdot E_a + (b^1_a)_i$$

Proposition Modules. Similarly, we define $(\text{PAC}^l_p)_i$ and $(\text{OUT}^l_p)_i$ for each proposition p. For proposition modules we additionally need variables $(\text{POOL}^l_{p,o})_i$ to represent the ith element of the pooled vector over actions related to p sharing the action schema o. If related actions belong to S action schemas $o_1, \ldots o_S$, we have, taking $j = (s-1)d + k$

$$\begin{aligned} (\text{PAC}^l_p)_i &= \sum_{s=1}^{S} \sum_{k=1}^{d} (W^l_p)_{i,j} \cdot (\text{POOL}^l_{p,o_s})_k + (b^l_p)_i \\ (\text{POOL}^l_{p,o})_i &= \max(\{(\text{OUT}^l_a)_i \mid \text{op}(a) = o \wedge R(a,p)\}) \end{aligned}$$

Activations. The encoding of f as the ELU function is very similar to the encoding of ReLU in MIP [Fischetti and Jo, 2018] and is the same for proposition and action modules. Given a proposition or action b we define the auxiliary variables $(t^l_b)_i$ and $(s^l_b)_i$ to store the positive and negative component of the variable $(\text{PAC}^l_b)_i$, respectively, and $(z^l_b)_i$ which is an indicator variable for the sign of $(\text{PAC}^l_b)_i$. This leads to the following constraints (note that Gurobi linearises the exponential, see [Gurobi Optimization, LLC, 2023, p584-586])

$$\begin{array}{ll} (z^l_b)_i = 1 \implies (t^l_b)_i \leqslant 0, & (z^l_b)_i = 0 \implies (s^l_b)_i \leqslant 0 \\ (t^l_b)_i \geqslant 0, & (s^l_b)_i \geqslant 0 \\ (\text{PAC}^l_b)_i = (t^l_b)_i - (s^l_b)_i, & (\text{OUT}^l_b)_i = (t^l_b)_i + e^{-(s^l_b)_i} - 1 \end{array}$$

Policy Output. The chosen action is the applicable action a maximising the output ϕ^L_a of the final layer's modules. We enforce this using the following additional variables: C_a which is a binary variable true iff action a is chosen, and $Vmax$ which is the maximal value of ϕ^L_a for *applicable* actions.

$$\begin{array}{ll} \sum_{a \in A} C_a = 1 & \forall a \in A \\ C_a = 1 \implies E_a = 1 & \forall a \in A \\ E_a = 1 \implies Vmax \geqslant \text{OUT}^L_a & \forall a \in A \\ C_a = 1 \implies Vmax \leqslant \text{OUT}^L_a & \forall a \in A \end{array}$$

Moreover, ties must be broken consistently to prevent the solver to select another action tied at the maximum value with the one currently tested. This issue frequently occurs when the problem has some symmetries, as e.g. in the Gripper domain where our ASNet policy often gives equal value to picking up the various balls in a room. Let M_a be a boolean variable true iff action a has maximum value, and $<$ be the total ordering over the action set A used by the policy to break ties, we break ties by favouring the action coming earlier in the ordering.

$$\begin{array}{l} (\text{OUT}^L_a = Vmax \wedge E_a = 1) \iff M_a = 1 \\ C_a = 1 \implies M_a - \sum_{a' < a} M_{a'} = 1 \end{array}$$

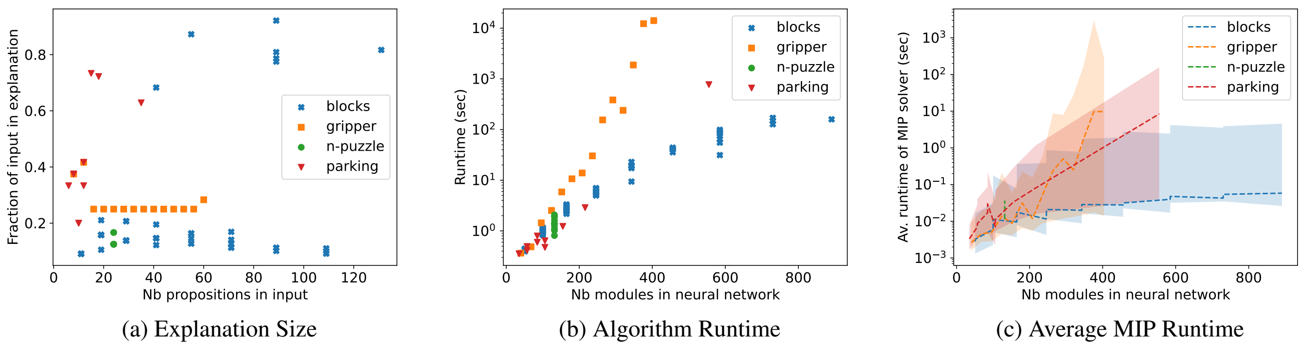
These additional constraints greatly increase the runtime of the MIP solver. Therefore we apply them only when needed. We first run the MIP without them, and if the solver finds a solution which selects an action tied with the action tested, we re-run the MIP with the tie-breaking constraints.

Temporary Constraints. The above constraints constitute the encoding C_π of the ASNet policy and only needs to be built once for a given sets X and A of state variables and actions. For each consistency query involving a planning instance $I = \langle X, A, g \rangle$, an action decision a, and a candidate explanation $\mathbf{z}$, it suffices to temporarily add the following constraints to set g as the goal, prevent a to be chosen by the policy, and look for a possible completion s of $\mathbf{z}$.

$$\begin{array}{ll} C_a = 0 & \\ S_{(x=\mathbf{z}[x])} = 1 & \forall x \in X_{\mathbf{z}} \\ G_{(x=g[x])} = 1 & \forall x \in X_g \\ \sum_{v \in D_x} G_{(x=v)} = 0 & \forall x \in X \backslash X_g \end{array}$$

6 Experimental Results

The goal of our experiments is to evaluate how effective abductive explanations are in explaining why a policy recommends a particular course of actions. In particular, we consider what fraction of the input appears in the minimal explanation: the smaller the explanation, the easier it is for a human to interpret. We also focus on the scalability of Algorithm 1 as the size of the policy increases, taking that of the naive algorithm in Section 4.1 as a baseline. Finally, we evaluate the sensitivity of the explanation size and runtime to the ordering in which variables are processed by our algorithm. For reproducibility, our repository https://github.com/Renee-Selvey/policy-explanations provides our algorithm implementation, benchmarks used, learnt policies, and the scripts to learn them and run the experiments.



(a) Explanation Size (b) Algorithm Runtime (c) Average MIP Runtime

Figure 3: Performance of Algorithm 1: size of the explanation produced as a fraction of the neural network input size (3a), runtime of the algorithm (3b) and average runtime of the MIP solver (3c) as a function of the number of network modules.

6.1 Experimental Setup

Hardware. All experiments were run on a machine with an AMD Ryzen Threadripper 3990X CPU, with 64 cores/128 threads, a clock speed of 2.9 GHz base, 4.3 GHz max boost, and 128 GB of memory of which we used 64 GB.

MIP Configuration. Gurobi version 9.1.2 is the MIP solver used for the experiments. To ensure the model is accurate enough for our experiments, we set the integer feasibility tolerance (`IntFeasTol`) to 10^{-9} and the error for function approximations (`FuncPieceError`) to 10^{-6}.

ASNet policies. We took all deterministic domains and training instances from the code distributions of [Toyer *et al.*, 2020] and [Steinmetz *et al.*, 2022]. We ran the script provided by Toyer using 1 core, an 8h timeout, and 64 GB of memory to learn, from these instances, two-layer (i.e. two proposition layers and three action layers) *sparse* policies with skip connections and without heuristic inputs. As described in [Toyer *et al.*, 2020, Sec. 6], the ℓ_1-regulariser used to train sparse policies results in many modules having coefficient so close to zero that they are insignificant to the output of the network and are prunned by the ASNet sparsification procedure. Hence, these modules do not appear in the MIP model, making a simpler model to solve.

6.2 Domains and Problems

We kept the domains for which the learnt policy could solve any problem within 150 execution steps, amongst a test set of 20 randomly generated small problems for that domain. This resulted in 4 policies for the domains of Blocksworld, Gripper, n-Puzzle, and Parking, which are present in the ASNets code distribution. In order to evaluate our approach, we randomly generated, for each of these domains, the set of problems described below. Explanations were only computed for problems where the policy was able to reach the goal within 150 steps. Each explanation problem was run with a time limit of 3h, except for Gripper for which the timeout was 4h.

Blocksworld. Experiments were conducted on 10 problems of each size (2-10 blocks) for a total of 90 problems. The policy was able to solve 71% of problems, all of which had explanations computed within the time limit.

Gripper. Explanations were computed using problems with 2 rooms and 1-15 balls. The policy was able to solve all problems, and explanations could be computed for 93%.

n-Puzzle. Experiments were conducted on all solvable and non-trivial combinations of 3-puzzle problems (on a 2x2 grid). Out of these 11 problems, the policy was correctly able to solve all instances and our approach could compute all explanations.

Parking. Problems in the parking domain comprised of 2 cars and 2-4 curbs. Out of the 22 total problems, the policy reached the goal for 55% and we were able to compute explanations for all of these within the time limit.

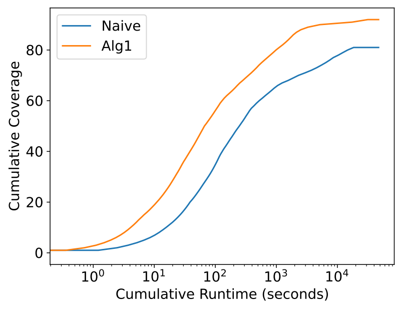
To avoid numerical instability issues, we excluded from our results below any problem for which a MIP call exceeded a constraint violation tolerance of 5e-01. This led to 8 Blocksworld and 2 Parking problems being excluded. Note moreover that we would have liked to evaluate our algorithm on larger n-Puzzle or Parking instances. However, we couldn't learn sparse policies capable of solving such instances, and the non-sparse (standard) ASNet policies produced with the ℓ_2-regulariser were too large for the MIP solver.

6.3 Results

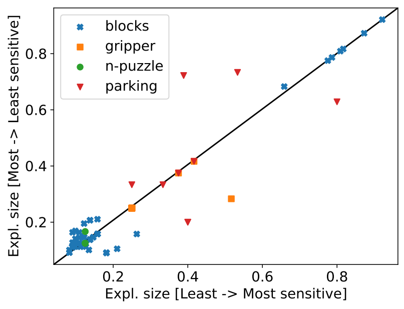
Size of Explanations

Figure 3a shows the size of the explanation obtained as a fraction of the size of the input. In all domains, Algorithm 1 produces an explanation smaller than ASNet's input. For n-Puzzle, the small explanations are due to many static propositions. The grid in the problem is defined as propositions listing the neighbours of each position. As this grid is static, this will not appear in the explanation.

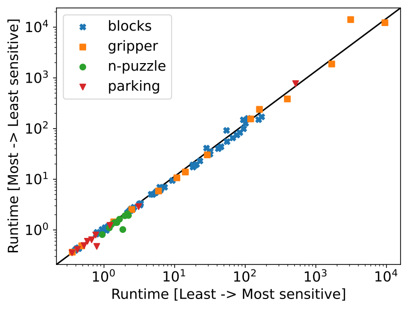
Many Blocksworld's explanations are also particularly small as many propositions remain false along the plan execution and are irrelevant, and some of the true propositions can be deduced from a small set of others. For instance, consider the optimal plan for a 3 block problem with b_2 on b_1 on b_3 initially and b_3 on b_2 and b_1 on the table in the goal. The explanation $clear(b_2), on(b_2, b_1), \neg ontable(b_1)$ returned by Algorithm 1 suffices to make the plan applicable and does not need to mention b_3 which is obviously on the table.



(a) Cumulative Coverage



(b) Sensitivity of Explanation Size



(c) Sensitivity of Algorithm Runtime

Figure 4: Cumulative Coverage Comparison: number of problems solved as a function of computation time for Algorithm 1 and Naive (4a); Sensitivity to Removal Ordering: explanation size (4b) and runtime (4c) when removing variables by increasing vs decreasing sensitivity.

The bulk of Gripper's explanations (i.e. all except for $n = 1, 2, 14$ balls) are steady at 1/4 of the input propositions. These explanations are obtained when the algorithm makes the best choices about which propositions to remove. They mention the actual location of the balls and Robby. The other propositions representing locations can be deduced to be false, and at that point the status of the grippers become irrelevant.

Scalability

Most of the runtime is spent in MIP consistency tests. As can be seen from the size of problems we can address, reasoning about neural networks using MIP is a serious bottleneck. Across all domains, we have observed that, as input propositions progressively get removed from the explanation, consistency tests generally become harder. This is because proving inconsistency generally becomes harder, and whilst proving consistency becomes easier, it leads us to reinsert the proposition, so the algorithm never reaches the easy region of the consistent side of the phase boundary. This behaviour is common with optimisation problems.

There are differences in scalability amongst the domains however. Figure 3b shows the runtime of the algorithm as a function of the size of the neural network policy for the problem. Gripper scaled the worst as a function of the number of modules in the network. This is due to two main factors. Firstly gripper plans are longer than those recommended by the policies for the other domains (from length 3 for $n = 1$ to 41 for $n = 14$) and therefore lead to a large number of MIP calls (from 16 to 1442). Secondly, and more importantly, the average runtime of MIP calls (shown in Figure 3c) is also larger than for the other domains, due to a large amount of tie-breaking being necessary for many problems. For $n = 1$ to 9, only the odd numbered problems required tie-breaking as is visible on both Figures 3b and 3c.

Blocksworld scaled the best. It has reasonably long plans (2 to 26 actions), and leads to even more MIP calls than Gripper (17 to 2176) because propositions survive longer through the inner loop before being reinstated in the explanation or definitely ruled out. However, as shown in Figure 3c, the MIP problems are relatively easy to solve, in part because, except for two problems, almost no tie-breaking was necessary.

We implemented the naive approach as a baseline for scalability comparison. The naive algorithm can solve the Parking and n-Puzzle problems as these have short plans (1 to 6 actions). It could not solve Gripper problems after $n = 4$ nor the largest Blocksworld problems. Moreover, as shown in Figure 4a, it is 5 to 10 times slower than Algorithm 1 on our benchmark problems.

Sensitivity to Removal Ordering

As explained in Section 3, there are potentially many minimal explanations for a policy trajectory. The explanation obtained, as well as its size and the runtime of Algorithm 1, depend on the order in which input variables are selected for removal on line 6 of the algorithm. In order to provide an idea of how stable the minimal explanation size and runtime are to the ordering, we use a heuristic inspired by [Wu *et al.*, 2022; Vinzent *et al.*, 2023], which computes the sensitivity of the network output to changes of value to each input variable. We then compare the explanations obtained by removing variables in increasing and decreasing order of their sensitivity, respectively. We measure variable sensitivity as follows.

$$\text{sen}(x) = \min_{i=1}^{n} \left(\pi(a_i, s_i) - \max_{\substack{a \in A(s_i)\setminus\{a_i\} \\ v \in D_x \setminus \{s_i[x]\}}} \pi(a, s_i \oplus (x = v))\right)$$

where, by slight abuse of notation, $\pi(a, s)$ is the (unnormalised) probability of the policy predicting action a in state s, before determinisation (this corresponds to the value of OUT_a^L in our MIP encoding). Hence, for each state in the trajectory, we take the difference between the output of the neural network for the action prescribed, and that of the best other action that would be prescribed if we changed the value of the variable x. The variable sensitivity sen(x) is the minimum of this difference over the states in the trajectory.

Figures 4b and 4c compare the explanation size and algorithm runtime, respectively, obtained under the two removal orderings. These figures suggest that except for Parking, the runtime and explanation size are only very moderately sensitive to the ordering.

7 Conclusion, Limits, and Future Work

We have extended abductive explanations to sequences of decisions recommended by neural network policies. Our decomposition approach to find a single minimal explanation incurs no overhead in comparison with the single decision case, once the length of the sequence is factored in.

Our approach makes a number of limiting assumptions which we discuss here together with possible extensions and future work. The first assumption is the availability of a symbolic planning model. An interesting avenue for future work is the extension of this approach to **learnt planning models**, also represented as neural networks.

We also assumed that actions have simple, unconditional effects. **Conditional effects** can be handled by our naive algorithm. However, Theorem 1 does not allow for them because it is impossible to apply conditional effects to a partial state. We assumed that we have no **background knowledge**, i.e. constraints that restrict the set of possible states [Thiébaux *et al.*, 2005; Rintanen, 2017]. These can greatly simplify explanations. Yu et al. [2023] showed that in the single decision case, adding the background knowledge K as another conjunct to the consistency tests performed by the greedy algorithm preserves the minimality of explanations. This property carries over to the multiple decision case and our naive algorithm. However, our decomposition algorithm may not return a minimal explanation with this augmented consistency test, because the set of reachable states cannot be represented by intersections of partial states with K. We leave an efficient treatment of conditional effects and background knowledge for future work.

We have assumed that the actions and the policy are deterministic. Handling **stochastic actions and/or policies** could be achieved by generating explanations that pertain to a finite tree of actions reachable with non-zero probability from the initial state. This would require applying progression and consistency tests at each branch of the tree. Handling stochastic policies would additionally require a more complex consistency test, as the decision d becomes a probability distribution over actions and the test must establish that it is not possible for the policy to return a different distribution.

Another avenue for future work is the generation of all (set-inclusion) **minimal explanations** and of **minimum cardinality explanations**. This is likely to require a different decomposition of the problem than the one presented here, as well as effective heuristics to guide search. Finally, even in the single decision case, methods for computing formal explanations of neural networks suffer from scalability issues due to the expensive consistency test. New breakthroughs in MIP and SMT methods for analysing neural networks, new problem abstractions, and approximate explanation methods will be needed for these approaches to flourish.

A Proof of Theorem 1

Lemma 1. *If* $\mathbf{z}$ *explains* $a_1, \dots, a_n$*, then for all* $i \in \{0, \dots, n-1\}$, $\mathbf{z}_{i+1} := prg_{a_i} \circ \dots \circ prg_{a_1}(\mathbf{z})$ *explains* a_{i+1}.

We shall prove that for all $i \in \{0, \dots, n-1\}$, for all state s_{i+1} that completes $\mathbf{z}_{i+1}$, there exists s_i such that i) a_i is applicable in s_i, ii) $s_i \xrightarrow{a_i} s_{i+1}$ is a transition of the planning domain, and iii) s_i completes $\mathbf{z}_i$. By recursion, we end up with a state s_1 that completes $\mathbf{z}_1$ such that $s_1 \xrightarrow{a_1} s_2 \xrightarrow{a_2} \dots \xrightarrow{a_i} s_{i+1}$. Since s_1 completes $\mathbf{z}_1 = \mathbf{z}$, we know that $\pi^n(s_1) = a_1, \dots, a_n$. Therefore, $\pi(s_{i+1}) = a_{i+1}$. As this is true for all s_{i+1}, $\mathbf{z}_{i+1}$ explains a_{i+1}.

The proof will be done by induction, i.e., we assume that it is true for i. (Base case for $i = 0$ is trivially true as $\mathbf{z}_1 = \mathbf{z}$.) Given state s_{i+1}, we choose s_i as one of the states satisfying

$$s_i \supseteq \mathbf{z}_i \oplus ((s_{i+1} \ominus \text{eff}(a_i)) \oplus \text{pre}(a_i)) . \qquad (3)$$

We prove that s_i satisfies the three points above.
i) Eq. 3 clearly enforces the precondition pre(a_i), so a_i is indeed applicable in s_i.
ii) We note

$$\text{prg}_{a_i}(s_i) \supseteq (s_{i+1} \ominus \text{eff}(a_i)) \oplus \text{pre}(a_i) \oplus \text{eff}(a_i).$$

which can be simplified by

$$\text{prg}_{a_i}(s_i) \supseteq s_{i+1} \oplus \text{pre}(a_i) \oplus \text{eff}(a_i). \qquad (4)$$

We also note that all variables $x \in X$ are specified in prg$_{a_i}(s_i)$, regardless of the choice of s_i. Consider any variable x; we show that prg$_{a_i}(s_i)[x] = s_{i+1}[x]$ holds by looking at 3 cases:

- If $x \in X_{\text{eff}(a_i)}$, then prg$_{a_i}(s_i)[x] = \text{eff}_{a_i}(x)$ and, since $\mathbf{z}_{i+1} = \text{prg}_{a_i}(\mathbf{z}_i)$, $s_{i+1}[x] = \mathbf{z}_{i+1}[x] = \text{eff}_{a_i}(x)$.
- If $x \in X_{\text{pre}(a_i)} \backslash X_{\text{eff}(a_i)}$, then prg$_{a_i}(s_i)[x] = s_i(x) = \text{pre}(a_i)[x]$ since a_i is applicable in s_i.
  Furthermore, we note that s_{i+1} completes prg$_{a_1}(\mathbf{z}_i)$ and that $\mathbf{z}_i[x] = \text{pre}(a_i)[x]$ since (by induction) $\mathbf{z}_i$ explains a_i; therefore $s_{i+1}[x] = \text{pre}(a_i)[x] = \text{prg}_{a_i}(s_i)[x]$.
- If $x \in X \backslash X_{\text{pre}(a_i)} \backslash X_{\text{eff}(a_i)}$, then prg$_{a_i}(s_i)[x] = s_i[x] = s_{i+1}[x]$.

iii) Does s_i complete $\mathbf{z}_i$? Since, by construction, s_i completes $\mathbf{z}_i \oplus ((s_{i+1} \ominus \text{eff}(a_i)) \oplus \text{pre}(a_i))$, the question is whether some assignment in the right operand of $\mathbf{z}_i \oplus$ contradicts an assignment in s_i. We know that $\mathbf{z}_i$ explains a_i, so pre$_{a_i}$ will not contradict $\mathbf{z}_i$. State s_{i+1} completes $\mathbf{z}_{i+1}$ which differs with $\mathbf{z}_i$ only on eff(a_i). However, the expression above removes eff(a_i) from s_{i+1}, so that the right operand does not map any variable to a different value than $\mathbf{z}_i$.

Lemma 2. *Let* $\mathbf{z}$ *be a partial state, and for all* $i \in \{0, \dots, n-1\}$, $\mathbf{z}_{i+1} := prg_{a_i} \circ \dots \circ prg_{a_1}(\mathbf{z})$*. If for all i,* $\mathbf{z}_i$ *explains* a_i*, then* $\mathbf{z}$ *explains* $a_1, \dots, a_n$.

Assume that $\mathbf{z}_i$ explains a_i for all i. Let s be a state completing $\mathbf{z}$ and let $s_1 \xrightarrow{a_1} \dots \xrightarrow{a_n} s_n$ be the trajectory obtained by applying the sequence of actions $a_1, \dots, a_n$ from $s_1 = s$. We shall prove $\pi(s_i) = a_i$ for all i (which proves that π recommends $a_1, \dots, a_n$); this is proven by showing that s_i completes $\mathbf{z}_i$.

The state s_{i+1} is defined as prg$_{a_i} \circ \dots \circ \text{prg}_{a_1}(s)$. Do we have $\forall x \in X_{\mathbf{z}_{i+1}}.\ s_{i+1}[x] = \mathbf{z}_{i+1}[x]$? Let j be the largest index in $\{1, \dots, i\}$ such that $x \in X_{\text{eff}(a_j)}$. If j does not exist, then $s_{i+1}[x] = s[x] = \mathbf{z}[x] = \mathbf{z}_{i+1}[x]$. Otherwise, $s_{i+1}[x] = \text{eff}(a_j)[x] = \mathbf{z}_{i+1}[x]$. Either way, the variables of s_{i+1} map to the same value as those of $\mathbf{z}_{i+1}$. Therefore, $\mathbf{z}_{i+1}$ explains a_{i+1} in s_{i+1}.

Theorem 1 is a consequence of Lemmas 1 and Lemma 2.

Acknowledgements

We thank the anonymous reviewers for their useful comments which helped to improve the paper. This work has received funding from the European Union's Horizon Europe Research and Innovation program under the grant agreement TUPLES No 101070149. S. Thiébaux was also supported by ARC grant DP220103815. A. Grastien and S. Thiébaux are members of the Humanising Machine Intelligence project at ANU.

References

[Bäckström and Nebel, 1995] Christer Bäckström and Bernhard Nebel. Complexity results for SAS+ planning. *Comput. Intell.*, 11:625–656, 1995.

[Chakraborti *et al.*, 2017] Tathagata Chakraborti, Sarath Sreedharan, Yu Zhang, and Subbarao Kambhampati. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. In *Proc. IJCAI*, pages 156–163, 2017.

[Chakraborti *et al.*, 2020] Tathagata Chakraborti, Sarath Sreedharan, and Subbarao Kambhampati. The emerging landscape of explainable automated planning & decision making. In *Proc. IJCAI*, pages 4803–4811, 2020.

[Darwiche and Hirth, 2020] Adnan Darwiche and Auguste Hirth. On the reasons behind decisions. In *Proc. ECAI*, pages 712–720, 2020.

[Eifler *et al.*, 2020] Rebecca Eifler, Marcel Steinmetz, Álvaro Torralba, and Jörg Hoffmann. Plan-space explanation via plan-property dependencies: Faster algorithms & more powerful properties. In *Proc. IJCAI*, pages 4091–4097, 2020.

[Ferber *et al.*, 2020] Patrick Ferber, Malte Helmert, and Jörg Hoffmann. Neural network heuristics for classical planning: A study of hyperparameter space. In *Proc. ECAI*, pages 2346–2353, 2020.

[Ferber *et al.*, 2022] Patrick Ferber, Florian Geißer, Felipe W. Trevizan, Malte Helmert, and Jörg Hoffmann. Neural network heuristic functions for classical planning: Bootstrapping and comparison to other methods. In *Proc. ICAPS*, pages 583–587, 2022.

[Fischetti and Jo, 2018] Matteo Fischetti and Jason Jo. Deep neural networks and mixed integer linear optimization. *Constraints*, 23(3):296–309, 2018.

[Garg *et al.*, 2020] Sankalp Garg, Aniket Bajpai, and Mausam. Symbolic network: Generalized neural policies for relational mdps. In *Proc. ICML*, pages 3397–3407, 2020.

[Gehring *et al.*, 2022] Clement Gehring, Masataro Asai, Rohan Chitnis, Tom Silver, Leslie Pack Kaelbling, Shirin Sohrabi, and Michael Katz. Reinforcement learning for classical planning: Viewing heuristics as dense reward generators. In *Proc. ICAPS*, pages 588–596, 2022.

[Groshev *et al.*, 2018] Edward Groshev, Maxwell Goldstein, Aviv Tamar, Siddharth Srivastava, and Pieter Abbeel. Learning generalized reactive policies using deep neural networks. In *Proc. ICAPS*, pages 408–416, 2018.

[Gurobi Optimization, LLC, 2023] Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual. https://www.gurobi.com/wp-content/plugins/hd_documentations/documentation/current/refman.pdf, 2023. Accessed: 2023-03-23.

[Ignatiev *et al.*, 2019] Alexey Ignatiev, Nina Narodytska, and João Marques-Silva. Abduction-based explanations for machine learning models. In *Proc. AAAI*, pages 1511–1519, 2019.

[Karia and Srivastava, 2021] Rushang Karia and Siddharth Srivastava. Learning generalized relational heuristic networks for model-agnostic planning. In *Proc. AAAI*, pages 8064–8073, 2021.

[Karia and Srivastava, 2022] Rushang Karia and Siddharth Srivastava. Relational abstractions for generalized reinforcement learning on symbolic problems. In *Proc. IJCAI*, pages 3135–3142, 2022.

[Kasenberg *et al.*, 2020] Daniel Kasenberg, Ravenna Thielstrom, and Matthias Scheutz. Generating explanations for temporal logic planner decisions. In *Proc. ICAPS*, pages 449–458, 2020.

[Krarup *et al.*, 2021] Benjamin Krarup, Senka Krivic, Daniele Magazzeni, Derek Long, Michael Cashmore, and David E. Smith. Contrastive explanations of plans through model restrictions. *J. Artif. Intell. Res.*, 72:533–612, 2021.

[Lundberg and Lee, 2017] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proc. NeurIPS*, pages 4765–4774, 2017.

[Marques-Silva and Ignatiev, 2022] João Marques-Silva and Alexey Ignatiev. Delivering trustworthy AI through formal XAI. In *Proc. AAAI*, pages 12342–12350, 2022.

[Ribeiro *et al.*, 2016] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proc. KDD*, pages 1135–1144, 2016.

[Ribeiro *et al.*, 2018] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proc. AAAI*, pages 1527–1535, 2018.

[Rintanen, 2017] Jussi Rintanen. Schematic invariants by reduction to ground invariants. In *Proc. AAAI*, pages 3644–3650, 2017.

[Shen *et al.*, 2020] William Shen, Felipe W. Trevizan, and Sylvie Thiébaux. Learning domain-independent planning heuristics with hypergraph networks. In *Proc. ICAPS*, pages 574–584, 2020.

[Sreedharan *et al.*, 2021] Sarath Sreedharan, Tathagata Chakraborti, and Subbarao Kambhampati. Foundations of explanations as model reconciliation. *Artif. Intell.*, 301:103558, 2021.

[Steinmetz *et al.*, 2022] Marcel Steinmetz, Daniel Fiser, Hasan Ferit Eniser, Patrick Ferber, Timo P. Gros, Philippe

Heim, Daniel Höller, Xandra Schuler, Valentin Wüstholz, Maria Christakis, and Jörg Hoffmann. Debugging a policy: Automatic action-policy testing in AI planning. In *Proc. ICAPS*, pages 353–361, 2022.

[Thiébaux *et al.*, 2005] Sylvie Thiébaux, Jörg Hoffmann, and Bernhard Nebel. In defense of PDDL axioms. *Artif. Intell.*, 168(1-2):38–69, 2005.

[Toyer *et al.*, 2018] Sam Toyer, Felipe W. Trevizan, Sylvie Thiébaux, and Lexing Xie. Action schema networks: Generalised policies with deep learning. In *Proc. AAAI*, pages 6294–6301, 2018.

[Toyer *et al.*, 2020] Sam Toyer, Sylvie Thiébaux, Felipe W. Trevizan, and Lexing Xie. Asnets: Deep learning for generalised planning. *J. Artif. Intell. Res.*, 68:1–68, 2020.

[Vasileiou *et al.*, 2022] Stylianos Loukas Vasileiou, William Yeoh, Tran Cao Son, Ashwin Kumar, Michael Cashmore, and Daniele Magazzeni. A logic-based explanation generation framework for classical and hybrid planning problems. *J. Artif. Intell. Res.*, 73:1473–1534, 2022.

[Vinzent *et al.*, 2022] Marcel Vinzent, Marcel Steinmetz, and Jörg Hoffmann. Neural network action policy verification via predicate abstraction. In *Proc. ICAPS*, pages 371–379, 2022.

[Vinzent *et al.*, 2023] Marcel Vinzent, Min Wu, Haoe Wu, and Jörg Hoffmann. Policy-specific abstraction predicate selection in neural policy safety verification. In *Proc. 2nd Workshop on Reliable Data-Driven Planning and Scheduling (RDDPS@ICAPS)*, 2023.

[Wu *et al.*, 2022] Min Wu, Haoze Wu, and Clark W. Barrett. Verix: Towards verified explainability of deep neural networks. *CoRR*, abs/2212.01051, 2022.

[Younes and Littman, 2004] Håkan Younes and Michael Littman. PPDDL1.0: an extension to PDDL for expressing planning domains with probabilistic effects. Technical report, CMU, 2004.

[Yu *et al.*, 2023] Jianqiang Yu, Alexey Ignatiev, Peter J. Stuckey, Nina Narodytska, and João Marques-Silva. Eliminating the impossible, whatever remains must be true. In *Proc. AAAI*, 2023.

[Zhang and Geißer, 2022] Ziqi Zhang and Florian Geißer. Extending graph neural networks for generalized stochastic planning. In *Proc. ICAPS-21 Planning and Reinforcement Learning Workshop (PRL)*, 2022.