New Fuzzing Biases for Action Policy Testing

Jan Eisenhut\textsuperscript{1}, Xandra Schuler\textsuperscript{1}, Daniel Fišer\textsuperscript{1}, Daniel Höller\textsuperscript{1}, Maria Christakis\textsuperscript{2}, Jörg Hoffmann\textsuperscript{1,3}

\textsuperscript{1}Saarland University, Saarland Informatics Campus, Saarbrücken, Germany
\textsuperscript{2}TU Wien, Austria
\textsuperscript{3}German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany

\{eisenhut,danfis,hoeller,hoffmann\}@cs.uni-saarland.de, s8xaschu@stud.uni-saarland.de, maria.christakis@tuwien.ac.at

Abstract

Testing was recently proposed as a method to gain trust in learned action policies in classical planning. Test cases in this setting are states generated by a fuzzing process that performs random walks from the initial state. A fuzzing bias attempts to bias these random walks towards policy bugs, that is, states where the policy performs sub-optimally. Prior work explored a simple fuzzing bias based on policy-trace cost. Here, we investigate this topic more deeply. We introduce three new fuzzing biases based on analyses of policy-trace shape, estimating whether a trace is close to looping back on itself, whether it contains detours, and whether its goal-distance surface does not smoothly decline. Our experiments with two kinds of neural action policies show that these new biases improve bug-finding capabilities in many cases.

Introduction

Learned action policies, in particular ones represented by neural networks, are gaining traction in AI (e.g., Mnih et al. 2013; Silver et al. 2016, 2018), and are being intensively explored in AI planning (Issakkimuthu, Fern, and Tadepalli 2018; Groshev et al. 2018; Garg, Bajpai, and Mauw 2019; Toyer et al. 2020; Karia and Srivastava 2021; Stähler, Bonet, and Geffner 2022a,b). But how to gain trust that such a policy will yield desirable (or at least non-fatal) behavior?

Policy testing is one natural answer to this question. Most work so far views the environment as an adversary, controlled by the testing mechanism in a way that tries to make the system fail (e.g., Dreossi et al. 2015; Akazaki et al. 2018; Koren et al. 2018; Ernst et al. 2019; Lee et al. 2020). While this can be quite useful, it does not distinguish whether or not the failure is actually due to bad policy decisions, or is unavoidable given the environment behavior.

Recent work (Steinmetz et al. 2022) addresses this in the context of classical planning, defining test cases as states \(s\), and policy bugs as states on which the policy is sub-optimal (a better policy exists for \(s\), e.g., avoiding failure). Given a test case \(s\), test oracles are used to detect whether \(s\) is a bug, by evaluating sufficient criteria that avoid the need for a full optimal planning process (Eisenhut et al. 2023). To generate test cases in the first place, a fuzzing process iteratively builds up a pool of test states by conducting random walks from the initial state. A fuzzing bias attempts to bias these random walks towards policy bugs, thus making the overall testing machinery more effective in finding bugs. Specifically, we consider bias functions mapping states to numbers, with higher numbers indicating more promising states.

Prior work (Steinmetz et al. 2022) only explored a simple bias function, mapping a state \(s\) to the summed-up action cost of the policy trace in \(s\). Here, we investigate this topic more deeply. We introduce three new fuzzing biases that aim at analyzing the policy-trace shape. In what we call the loopiness bias, we measure how close a trace is to looping back on itself, a typical reason for failure to reach the goal (in dead-end free domains, this is indeed the only reason). In the detour bias, we measure the largest detour (the most sub-optimal segment) along the trace. In the surface bias, we consider the goal-distance along the trace as its surface, and measure the degree to which that surface does not smoothly decline along the path. All three measures rely on estimates of the distance between states (on the trace, respectively to the goal). We include an analysis of the idealized setting using the exact distance \(h^∗\); in practice, we use the delete relaxation heuristic \(h^{FF}\) (Hoffmann and Nebel 2001).

Using \(h^{FF}\) instead of \(h^∗\) of course makes the measurements inaccurate. In particular, “detours” and “goal distance increases” will often simply be due to \(h^{FF}\) estimation errors, but biases may still contain information useful for bug-finding. Indeed, this is what our experiments show. We consider two kinds of neural action policies: ASNets (Toyer et al. 2018, 2020), and Stähler et al.’s (2022a; 2022b) GNN policies. We test these policies across a set of domains where they perform well. Our new fuzzing biases improve bug-finding capabilities in many of these cases.

Background

A finite-domain representation planning task, short FDR planning task, is a tuple \(\Pi = (\mathcal{V}, \mathcal{A}, I, G)\), \(\mathcal{V}\) is a finite set of variables; each variable \(v \in \mathcal{V}\) has a finite domain \(\text{dom}(v)\). A partial state \(p\) is a variable assignment over some variables \(\mathcal{V}(p) \subseteq \mathcal{V}\); \(p[v]\) is the value assigned to \(v \in \mathcal{V}(p)\) in \(p\). A partial state \(s\) is a state if \(\mathcal{V}(s) = \mathcal{V}\); \(S\) denotes the set of all states. \(I\) is an initial state, \(G\) is a partial state called goal, and a state \(s\) is a goal state if \(G \subseteq s\).

\(\mathcal{A}\) is a finite set of actions, \(a \in \mathcal{A}\) is defined by its precondition \(\text{pre}(a)\), effect \(\text{eff}(a)\), and cost \(\text{cost}(a) \in \mathbb{R}_0^+\). Pre-
conditions and effects are partial states. $a \in \mathcal{A}$ is applicable in $s \in \mathcal{S}$ if $\text{pre}(a) \subseteq s$. Given $s \in \mathcal{S}$, $\mathcal{A}[s]$ is the set of all actions applicable in $s$. The resulting state of applying an applicable action $a$ in a state $s$ is $s' = s[a]$ s.t. $s'[v] = \text{eff}(a)[v]$ for every $v \in \mathcal{V}(\text{eff}(a))$, and $s'[v] = s[v]$ otherwise. A sequence of actions $\vec{a} = \langle a_1, \ldots, a_n \rangle$ is applicable in $s_0 \in \mathcal{S}$ if there are states $s_1, \ldots, s_n$ s.t. $a_i$ is applicable in $s_{i-1}$ and $s_i = s_{i-1}[a_i]$, the resulting state is $s_0[\vec{a}] = s_n$. Given $\vec{a}$ applicable in $s$, the sequence $\langle s_0, s_1, \ldots, s_n \rangle$ of the aforementioned states is called intermediate state sequence of $\vec{a}$. The cost of $\vec{a}$ is defined as $\text{cost}(\vec{a}) = \sum_{i=1}^{n} \text{cost}(a_i)$. If $\vec{a}$ is applicable in $s$ and $G \subseteq s[\vec{a}]$, $\vec{a}$ is called a plan for $s$.

We consider heuristic functions $h$ as estimators for costs of optimal plans, denoted as $h(s)$, and optimal paths between two states $s$ and $t$, denoted $h(s, t)$. $h^*(s)$ and $h^*(s, t)$ denote optimal costs. We assume all heuristics are safe, i.e., for all $s \in \mathcal{S}$, $h(s) = \infty$ implies $h^*(s) = \infty$.

Deterministic policies $\pi: \mathcal{S} \rightarrow \mathcal{A} \cup \{\emptyset\}$ map states to applicable actions or null ($\emptyset$) if there is no applicable action. The run (or trace) $\sigma^\pi(s)$ of a policy $\pi$ on a state $s$ is an action sequence $\sigma^\pi(s) = \langle a_1, \ldots, a_n \rangle$ applicable in $s$ (with its intermediate state sequence $\langle s_0, \ldots, s_n \rangle$) s.t. $a_i = \pi(s_{i-1})$ for all $1 \leq i \leq n$, all intermediate states differ from each other, and no $s_i$ is a goal state unless $i = n$. The cost $\sigma^\pi(s)$ is $\text{cost}(\sigma^\pi(s)) = \sum_{i=1}^{n} \text{cost}(a_i)$ if $\sigma^\pi(s)$ is a goal state, and $\text{cost}(\sigma^\pi(s)) = \infty$ otherwise; $\text{cost}^\pi_i(s_i, s_j)$ (i.e., $i < j$) denotes the cost of the subsequence of $\sigma^\pi(s_0)$ between $s_i$ and $s_j$.

We adopt the testing framework of Steinmetz et al. (2022): Given a task $I$ and a policy $\pi$, a state $s$ is called a bug in $\pi$ if $\text{cost}(s) > h^*(s)$.

To find bugs, policy testing (1) generates a pool $P \subseteq \mathcal{S}$ of test states using a fuzzing process, and (2) runs test oracles on each pool state. Here, we focus on (1) only. For (2), we fix a state-of-the-art oracle of Eisenhut et al. (2023).

The fuzzing process proposed by Steinmetz et al. uses random walks from $I$ guided by a bias $B$, which is a function mapping states to numbers where higher numbers indicate more promising states (more likely to lead to a bug state). The fuzzer evaluates $B$ on every successor state $s$, and obtains a probability distribution over the respective actions by normalizing the biases $B(s)$. We adopt this fuzzing process here, though with some refinements as described in the next section. Steinmetz et al. only explored a simple bias $\mathcal{B}^\pi(s) = \text{cost}^\pi(s)$: the cost of the policy run. The fuzzing process with $\mathcal{B}^\pi$ prefers to explore states with more costly policy traces. Under the assumption that such states are more likely to be bugs, this makes sense, and indeed it often increases the number of detected bugs in comparison to a uniform random fuzzer. Nevertheless, this bias encapsulates a rather naive notion of “good” vs. “bad” policy traces. Here, we investigate this topic more deeply, through new fuzzing biases that aim at analyzing the shape of policy runs.

The oracle we employ in the second phase of the testing process is a so-called bound maintenance oracle (BMO). A BMO maintains upper bounds $u(t)$ on $h^*(t)$ across the states $t$ it encounters during testing and attempts to propagate and decrease these bounds through state comparisons based on quantitative dominance functions (Torralba 2017). The underlying idea of how these bounds can be propagated is simple. Assume that we know that a previously encountered state $s$ is solvable with a cost of at most 5 (i.e., $h^*(s) \leq u(s) = 5$) and that $t$ dominates $s$ by cost 2 (i.e., $h^*(t) \leq h^*(s) - 2$). We can then easily conclude that 3 must be an upper bound on $h^*(t)$, update $u(t)$ accordingly, and flag $t$ as a bug if $h^*(t) \leq u(t) \leq \text{cost}^\pi(t)$. While the most basic BMO uses the cost of policy runs as the single initial source of upper bounds (as $h^*(t) \leq \text{cost}^\pi(t)$), any other sources such as search methods or external tools can be readily integrated. Crucially, this means that we do not require an optimal planning process in order to prove that the policy is suboptimal on a state.

### The Fuzzing Process

We design a refined version of the fuzzer introduced by Steinmetz et al. (2022), see Algorithm 1. Within a runtime limit $T$, we iteratively generate a pool $P$ of maximal size $N$ by random walks of maximal length $L$. Random walks start from a randomly selected pool state $s_0$ and are steered by a bias function $B$, influencing which successor state $s'$ is selected in each step of the random walk. Steinmetz et al. only considered the bias $\mathcal{B}^\pi$ returning the cost of the policy run. Here, we use bias functions $B: \mathcal{S} \rightarrow \mathbb{R} \cup \{-\infty, \infty\}$ allowing negative bias values as well as $-\infty$ and $\infty$.

The application of $B$ is still straightforward. In each step

---

1. It covers both “quantitative” ($\pi$ is sub-optimal on $t$) and “qualitative” ($\pi$ does not solve solvable $t$) bugs used by Steinmetz et al.
of the random walk, we compute a set $S'$ containing pairs $(s', b)$ of successor states $s'$ and bias values $b$, ignoring goal states and states with bias value $-\infty$. As Steinmetz et al., we employ the safe heuristic $h^{\text{max}}$ (Bonet and Geffner 2001) as a dead end detector, i.e., if $h^{\text{max}}(s') = \infty$, we ignore $s'$ as it is a dead end and hence cannot be a bug.

Having generated $S'$, we then pick one of the states in $S'$ as follows (function weightedSelect): if there is any $(s', \infty)$ in $S'$, we pick one $s'$ among all such $s'$ uniformly at random. Otherwise, we determine the minimal bias value $b_{\text{min}}$. If $b_{\text{min}} < 0$, we increase all $b$ by $|b_{\text{min}}|$ so that all $b$ are non-negative. Finally, we normalize the bias values to sum up to 1 and select $s'$ according to this probability distribution.

For biases that require computing the cost of partial policy runs ($B^\pi$ and all biases introduced next), our fuzzer includes further additions. For each successor state $s'$, we use the function PartialRun($s'$, $\pi$, $k$) to obtain the partial policy trace $\sigma$ of $\pi$ on $s'$ of at most $k$ steps, which is then analyzed by $B$. If we can infer from $\sigma$ that $\pi$ fails on $s'$, we ignore $(s', B)$ and assign $b = \infty$ instead, i.e., we consider only policy runs that could lead to a goal state. As we ignore goal states, this also means $B(s')$ is not computed if $\sigma$ is empty.

Since evaluating policies is often very time-consuming, we introduce simple mechanisms bounding the number of policy evaluations. We limit the number of steps of individual policy runs by $B_{\text{state}}$, and we limit the overall number of policy steps per successor state by $B_{\text{expansion}}$.

Note that Steinmetz et al.’s original fuzzer additionally allows to filter out states based on a novelty measure (Lipovetzky and Geffner 2012), which we do not use here to make the experimental results easier to interpret.

**Bias Functions Measuring Policy-Trace Shape**

The only bias from prior work preferred policy traces with high cost, which may be due to bad policy behavior, but may just as well simply be due to high plan cost. Here, we instead introduce measures of policy-trace shape, encapsulating deeper notions of policy-trace quality. Specifically, we introduce what we call the loopiness bias, detour bias, and surface bias. All three biases consider pairs of states $s_i$ and $s_j$ on the policy trace, and maximize a measure of “trace-shape badness” over all such state pairs. These measures all rely on heuristic functions $h$ estimating state distances; we consider both, the idealized setting where $h = h^*$, as well as $h = h^{\text{FF}}$ which we use in practice. As part of our experiments, we run the idealized biases with $h^*$ where feasible, to evaluate the impact of using $h^{\text{FF}}$ instead.

We next introduce each bias in turn. In the following, let $s$ denote the current state reached by a random walk as per Algorithm 1, let $\langle a_1, \ldots, a_n \rangle$ denote a non-empty partial run of the policy $\pi$ on $s$, i.e., the first $n$ steps of the policy run $\sigma^\pi(s)$, and let $\langle s_0, \ldots, s_n \rangle$ denote the respective intermediate state sequence where $s = s_0$.

**Loopiness Bias**

The loopiness bias is geared at estimating whether, at some point along the policy trace, that trace is close to forming a loop. The bias is defined as follows:

$$B^{\text{loop}}_h(s) = \max_{0 \leq i < j \leq n} \cosf^\pi(s_i, s_j) - h(s_j, s_i).$$

For every state $s_i$ and later state $s_j$ along the policy trace, we measure (a) the cost of the policy path from $s_i$ to $s_j$, minus (b) the estimated cost of a plan from $s_j$ back to $s_i$. This measure will be high if $s_i$ and $s_j$ are distant on the policy trace, but only few steps are needed to go back from $s_j$ to $s_i$. Note here that, while one may be tempted to use only (b) and minimize over its values, (a) is useful because states close to each other on the path may naturally have very small values for (b). Indeed, in invertible planning tasks where every action has a direct inverse of the same cost, this is necessarily the case as (b) will be the cost of a single action for every state pair $s_i$ and $s_j$ with $j = i + 1$ along the trace.

Invertible tasks also form a special case regarding the outcome of the maximization in the idealized setting:

**Proposition 1.** In invertible planning tasks, $B^{\text{loop}}_h(s) = \cosf^\pi(s_0, s_n) - h^*(s_n, s_0)$.

**Proof.** Let $i < j$ be arbitrary. If we decrease $i$, the value of $\cosf^\pi(s_i, s_j) - h^*(s_j, s_i)$ can only increase because $\cosf^\pi(s_{i-1}, s_j)$ is at most cost($a_i$) higher and $h^*(s_j, s_{i-1})$ is at most cost($a_i$) higher (we can take the inverse of $a_i$ to get from $s_i$ back to $s_{i-1}$). If we increase $j$, the value of $\cosf^\pi(s_i, s_j) - h^*(s_j, s_i)$ can also only increase because $\cosf^\pi(s_i, s_{j+1})$ is at most cost($a_{j+1}$) higher and $h^*(s_{j+1}, s_i)$ is at most cost($a_{j+1}$) higher (we can take the inverse of $a_{j+1}$ to get from $s_{j+1}$ back to $s_j$). It follows that the maximal value is obtained at $i = 0, j = n$. \hfill $\square$

Of course, this does not hold if we replace $h^*$ with $h^{\text{FF}}$.

We considered to save computational effort (avoiding maximization) for $B^{\text{loop}}_h$ by detecting invertible planning tasks automatically (e.g., Hoffmann 2005), but these automatic tests failed to find any invertible tasks in our benchmark set.

**Detour Bias**

The detour bias aims to directly measure whether a policy trace is sub-optimal. We consider every segment from $s_i$ to $s_j$ on the trace and compare its cost under the policy to the estimated cost of a path from $s_i$ to $s_j$:

$$B^{\text{detour}}_h(s) = \max_{0 \leq i < j \leq n} \cosf^\pi(s_i, s_j) - h(s_j, s_i).$$

Large values of $\cosf^\pi(s_i, s_j) - h(s_i, s_j)$ indicate that the policy trace “takes a detour” from $s_i$ to $s_j$. Note that, syntactically, the only difference to the loopiness bias is the order of $s_i$ and $s_j$ in the call to $h$. That difference can lead to arbitrary differences in bias value however.

For $h^*$ (but not for $h^{\text{FF}}$), maximization here is not needed as any subpath detour is contained in the overall path:

**Proposition 2.** $B^{\text{detour}}_h(s) = \cosf^\pi(s_0, s_n) - h^*(s_0, s_n)$.

**Proof.** Same argument as for Proposition 1, except we do not need to rely on inverse actions. \hfill $\square$
It is furthermore easy to see that the idealized detour bias carries exact information about bugs and their optimality gap (note that unsolvable states can by definition not be bugs):

**Proposition 3.** Let \( s \) be a solvable state. If \( B_{h^*}^{\text{detour}}(s) > 0 \), then \( s \) is a bug. If \( s \) is a bug and \( \langle a_1, \ldots, a_n \rangle \) reaches the goal, then \( B_{h^*}^{\text{detour}}(s) > 0 \) and there exists a plan for \( s \) whose cost is at least \( \text{cost}^*_s(s_0, s_n) - B_{h^*}^{\text{detour}}(s) \).

**Proof.** If \( B_{h^*}^{\text{detour}}(s) > 0 \) then the partial run of \( \pi \) on \( s \) is sub-optimal so the overall run cannot be optimal.

For the second part of the claim, say that \( \langle a_1, \ldots, a_n \rangle \) reaches the goal. If \( s \) is a bug, then \( \langle a_1, \ldots, a_n \rangle \) is a sub-optimal plan and hence \( B_{h^*}^{\text{detour}}(s) > 0 \). Furthermore, by Proposition 2 we have \( \text{cost}^*_s(s_0, s_n) - B_{h^*}^{\text{detour}}(s) = \text{cost}^*_s(s_0, s_n) - (\text{cost}^*_s(s_0, s_n) - h^*(s_0) - h^*(s_n)) = h^*(s_0) - h^*(s_n) \). As \( s_0 \) is a goal state we have \( h^*(s_0) = 0 \), so \( \text{cost}^*_s(s_0, s_n) - B_{h^*}^{\text{detour}}(s) = h^*(s) \). In particular, if \( s \) is a bug, then \( \langle a_1, \ldots, a_n \rangle \) is a sub-optimal plan and hence \( B_{h^*}^{\text{detour}}(s) > 0 \). \( \square \)

**Surface Bias**

Search space surface has previously been investigated as a way of examining the quality of heuristic functions (e.g., Hoffmann 2005, 2011). Here we take a related perspective, considering the surface of the policy trace when taking \( h \)-values as the vertical dimension. On an optimal path to the goal and if \( h = h^* \), this surface will decline smoothly, i.e., it will reduce by \( \text{cost}(a_i) \) in each step. The surface bias aims at measuring the degree to which this is not the case: \( B_{h^*}^{\text{surf}}(s) = \)

\[
\begin{cases}
-\infty & h(s) = \infty \\
\max_{0 \leq i < j \leq n} \text{cost}^*_s(s_i, s_j) - (h(s_i) - h(s_j)) & h(s) \neq \infty
\end{cases}
\]

If \( h(s) = \infty \), \( s \) is unsolvable, so we return the minimal bias value \(-\infty\). Otherwise, this measure includes (a) the cost of the policy path from \( s_i \) to \( s_j \), minus (b) the decrease in \( h \)-value from \( s_i \) to \( s_j \). If (a) is larger than (b), then the policy-trace surface between \( s_i \) and \( s_j \) does not decline as smoothly as it should. In particular, if \( s_j \) is far from \( s_i \) on the path, but is estimated to be further from the goal, i.e., \( h(s_j) > h(s_i) \), the surface bias will be very high, indicating that the trace is meandering around, not making progress towards the goal. Of course, outside the idealized \( h^* \) setting such an indication may be misguided merely due to local minima in \( h \). Again, maximization here is never needed for \( h = h^* \): **Proposition 4.** \( B_{h^*}^{\text{surf}}(s) = \text{cost}^*_s(s_0, s_n) - (h^*(s_0) - h^*(s_n)) \).

**Proof.** Same argument as for Proposition 2, except we consider \( h^*(s_i) \) and \( h^*(s_j) \) separately. \( \square \)

The bug information in the idealized setting is even stronger than for the detour bias:

**Proposition 5.** Let \( s \) be a solvable state. If \( B_{h^*}^{\text{surf}}(s) > 0 \), then \( s \) is a bug. If \( s \) is a bug and \( \langle a_1, \ldots, a_n \rangle \) reaches the goal, then \( B_{h^*}^{\text{surf}}(s) > 0 \) and the optimal plan cost for \( s \) is exactly \( \text{cost}^*_s(s_0, s_n) - B_{h^*}^{\text{surf}}(s) \).

**Proof.** The first part of the claim holds because optimal paths are exactly those where \( h^* \) decreases by action cost in every step.

For the second part of the claim, say that \( \langle a_1, \ldots, a_n \rangle \) reaches the goal. By Proposition 4 we have \( \text{cost}^*_s(s_0, s_n) - B_{h^*}^{\text{surf}}(s) = \text{cost}^*_s(s_0, s_n) - (\text{cost}^*_s(s_0, s_n) - h^*(s_0) - h^*(s_n)) = h^*(s_0) - h^*(s_n) \). As \( s_0 \) is a goal state we have \( h^*(s_0) = 0 \), so \( \text{cost}^*_s(s_0, s_n) - B_{h^*}^{\text{surf}}(s) = h^*(s) \). In particular, if \( s \) is a bug, then \( \langle a_1, \ldots, a_n \rangle \) is a sub-optimal plan and hence \( B_{h^*}^{\text{surf}}(s) > 0 \). \( \square \)

The surface bias also has a desirable property in a non-idealized setting:

**Proposition 6.** For any safe heuristic function \( h \), \( B_{h^*}^{\text{surf}}(s) = \infty \) only if \( \langle a_1, \ldots, a_n \rangle \) ends in a dead-end state.

**Proof.** By definition we have that \( h(s) < \infty \). So \( B_{h^*}^{\text{surf}}(s) = \infty \) iff \( h(s_i) = \infty \) for at least one state \( s_i, i > 0 \). \( \square \)

**Experiments**

For our experiments, we extend the testing framework by Eisenhut et al. (2023), which builds on the one by Steinmetz et al. (2022). As in this previous work, we test ASNet policies (Toyer et al. 2018, 2020). We use our own, less resource intensive, C implementation of ASNets. We also experiment with a more comprehensive benchmark set for ASNets, which contains additional IPC domains, as well as a new “Beluga” domain. The latter domain is named after the Airbus transport aircraft used for moving aircraft components, and is an abstract encoding of a problem occurring in the Airbus manufacturing process. The problem involves unloading parts from a Beluga airplane; storing them on intermediate storage racks that behave like queues with access to front and end element; and transporting parts from the racks to a factory in the required order.

We also consider the GNN policies by Stålberg et al. (2022a; 2022b). Here, we test the original policies provided by Stålberg, Bonet, and Geffner (2022b) on a subset of their benchmark set (excluding Logistics and Spanner). That is, we include only the domains in which the policies work well without augmenting states with derived atoms.

Both kinds of policies are run in separate processes with their own memory limit of 2 GiB for ASNets and 8 GiB for GNNs. The testing engine uses a limit of 2 GiB. All experiments were run on a cluster with Intel E5-2660 processors.2

We compare our new biases (in \( h^\text{PP} \) and \( h^* \) configuration) with two baselines: the constant bias \( B^0 = 0 \) and \( B^\text{surf} \) defined as \( B^\text{surf}(s) = \text{cost}^*_s(s_0, s_n) \). Given our additions to the fuzzier, \( B^\text{surf} \) differs from Steinmetz et al.’s definition in that it only computes the cost of partial runs.

For each bias \( B \) and planning task, we run the fuzzers from Algorithm 1 for at most one hour to generate a pool \( \mathcal{P} \) of up to \( N = 100 \) states. We limit the length of random walks to \( L = 5 \) and use the budgets \( B_{\text{state}} = 50 \) and \( B_{\text{expansion}} = 200 \). For the test oracle, we use the best-performing oracle of Eisenhut et al. (2023) (“BMO-100 + Aras/EHC”). We run this oracle for at most two hours per test pool \( \mathcal{P} \), to determine how many states in \( \mathcal{P} \) can be classified as bugs.

---

2Our code, the benchmarks, and the policies are available at https://github.com/fai-saarland/bughive/tree/icaps24-biases.
We refer to the percentage of tested pool states (across the entire domain) classified as bugs as *oracle coverage*. To foster comparison of oracle coverage across biases, we use the same number of test states across all biases for each problem instance. If we have, e.g., 90 pool states for \( B_1 \) but only 50 for \( B_2 \), we consider only the first 50 pool states for \( B_1 \).

Table 1 shows the results. The left part shows the comparison of practical biases (not using \( h^* \)). In general, the behavior is highly domain specific and there is no bias that works best universally. However, in 8 of the 12 domains considered for AS Nets and in 4 out of 6 domains for GNNs, one of our new biases achieves a higher oracle coverage than both baselines. The new biases perform significantly better than \( B^0 \) in most cases, though they often provide only a slight advantage over \( B^* \). Oracle coverage varies, however, greatly between domains. For example, coverage for AS Nets in Elevator is at least 75% across all biases, but we do not detect a single bug state for AS Nets in Gripper (presumably, the policy there is optimal). We are able to increase coverage by 10 percentage points (from \( B^* \)) in Beluga and Woodworking, but in Blocks the difference between all biases is negligible.

The right part of Table 1 shows the results when including all biases. As this includes \( h^* \) based biases, the number of considered states \( S \) is mostly much lower (e.g., the pools for AS Nets in Blocksworld contain the initial state only). While there are cases where the \( h^* \) variant outperforms its \( h^{FP} \) proxy (e.g. \( B_{h^{FP}}^{\text{detour}} \) for AS Nets in Scanalyzer), this is not the case in general, indicating that \( h^{FP} \) works reasonably well here in replacing \( h^* \). Note also that, here, there is only a single case (GNNs in Visitall) where a baseline achieves better results than all configurations of our new biases.

<table>
<thead>
<tr>
<th>Policy</th>
<th>Domain</th>
<th>#II</th>
<th>( S )</th>
<th>Oracle coverage (practical biases)</th>
<th>( S )</th>
<th>Oracle coverage (all biases)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beluga</td>
<td>67</td>
<td>4476.4</td>
<td>35.9 41.7</td>
<td>51.8 44.4 41.1</td>
<td>2509.8</td>
<td>14.0 19.7 31.6 33.6 23.7 18.5 18.8 19.2</td>
</tr>
<tr>
<td>Blocks</td>
<td>21</td>
<td>1488.0</td>
<td>7.6 7.2</td>
<td>7.3 7.4 7.6</td>
<td>21.0 4.8 4.8 4.8 4.8 4.8 4.8 4.8</td>
<td></td>
</tr>
<tr>
<td>Elevator</td>
<td>58</td>
<td>5279.8</td>
<td>75.0 89.8</td>
<td>92 92.6 90.6 91.4</td>
<td>341.2 45.4 66.7 73.7 73.6 68.4 68.4 71.0 73.6</td>
<td></td>
</tr>
<tr>
<td>Floortile</td>
<td>14</td>
<td>1400.0</td>
<td>11.3 5.2</td>
<td>5.2 4.0 4.1</td>
<td>182.4 8.5 4.0 4.0 3.9 3.9 3.9 3.9 3.9</td>
<td></td>
</tr>
<tr>
<td>Gripper</td>
<td>35</td>
<td>3967.6</td>
<td>0.0 0.0</td>
<td>0.0 0.0 0.0</td>
<td>752.4 0.0 0.0 0.0 0.0 0.0 0.0 0.0</td>
<td></td>
</tr>
<tr>
<td>MB Blocks</td>
<td>16</td>
<td>1477.0</td>
<td>15.1 13.9</td>
<td>14.0 14.3 13.5</td>
<td>524.6 13.3 12.5 13.3 13.5 13.5 13.5 66.4 12.9 13.5</td>
<td></td>
</tr>
<tr>
<td>Satellite</td>
<td>15</td>
<td>1210.0</td>
<td>67.1 83.6</td>
<td>86.0 89.1 85.9</td>
<td>388.4 84.8 69.7 74.3 81.8 78.3 68.3 73.7 83.1</td>
<td></td>
</tr>
<tr>
<td>Scanalyzer</td>
<td>48</td>
<td>2796.2</td>
<td>42.4 68.8</td>
<td>73.5 70.1 71.1</td>
<td>543.0 26.3 57.9 62.8 89.3 59.0 56.1 59.1 59.1</td>
<td></td>
</tr>
<tr>
<td>Storage</td>
<td>19</td>
<td>1137.4</td>
<td>47.6 78.7</td>
<td>79.5 80.0 81.4</td>
<td>486.2 32.0 59.3 58.8 59.2 61.3 59.6 61.3 59.2</td>
<td></td>
</tr>
<tr>
<td>Transport</td>
<td>26</td>
<td>1946.8</td>
<td>62.1 79.5</td>
<td>81.9 82.6 83.0</td>
<td>944.4 36.5 66.6 71.1 70.5 71.4 63.2 73.5 70.5</td>
<td></td>
</tr>
<tr>
<td>VisitAll</td>
<td>30</td>
<td>1812.0</td>
<td>67.6 81.1</td>
<td>85.9 80.7 81.6</td>
<td>566.4 44.2 56.1 64.1 64.4 56.4 53.7 58.4 58.5</td>
<td></td>
</tr>
<tr>
<td>Woodwork</td>
<td>38</td>
<td>3700.6</td>
<td>55.6 78.2</td>
<td>81.1 88.5 75.9</td>
<td>1779.4 46.7 68.4 71.6 68.8 83.6 66.3 66.0 64.7</td>
<td></td>
</tr>
<tr>
<td>AS Nets</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blocks</td>
<td>35</td>
<td>2944.6</td>
<td>34.9 34.6</td>
<td>36.0 36.1 36.4</td>
<td>1646.6 14.1 14.7</td>
<td>28.2 15.7 15.0 16.0</td>
</tr>
<tr>
<td>Delivery</td>
<td>26</td>
<td>2113.8</td>
<td>18.2 17.5</td>
<td>21.6 17.3 23.3</td>
<td>1925.2 13.5 13.5</td>
<td>16.9 22.2 14.6 14.6</td>
</tr>
<tr>
<td>Gripper</td>
<td>17</td>
<td>986.8</td>
<td>24.1 25.7</td>
<td>21.4 21.1 24.4</td>
<td>312.0 26.7 27.9</td>
<td>20.5 18.8 9.9</td>
</tr>
<tr>
<td>Miconic</td>
<td>30</td>
<td>4020.0</td>
<td>34.2 32.3</td>
<td>38.1 34.5 32.6</td>
<td>3866.4 33.1 34.1</td>
<td>37.0 38.5 33.3</td>
</tr>
<tr>
<td>Reward</td>
<td>20</td>
<td>1329.2</td>
<td>39.6 38.7</td>
<td>41.3 45.8 41.5</td>
<td>1298.4 39.7 38.7</td>
<td>40.4 43.8 45.8</td>
</tr>
<tr>
<td>Visitall</td>
<td>16</td>
<td>431.0</td>
<td>41.3 59.4</td>
<td>51.9 51.1 58.6</td>
<td>206.2 17.7 33.6</td>
<td>19.6 32.2 18.3</td>
</tr>
</tbody>
</table>

Table 1: Oracle coverage (in %) for two sets of biases. \#II is the number of problem instances per domain; we only include tasks where \( \pi \) solves the initial state. \( S \) is the overall number of considered states (sum over domain). Oracle coverage is the percentage of states in the respective combined pool (of size \( S \)) identified as bugs (see text). Results averaged over five runs. Beluga is a new domain, abstractly encoding a problem occurring in the Airbus manufacturing process (see text).

**Conclusion**

With the proliferation of learned action policies, how to meaningfully test them is becoming very relevant. We contribute more advanced fuzzing biases for policy testing in classical planning, measuring path-shape features instead of merely path cost. The results with respect to previous simpler biases show that advanced biases can be useful, though often to only a small degree.

We are skeptical whether much more can be gained by investigating additional fuzzing biases. In particular, our \( h^* \) based biases have ideal theoretical properties — carrying exact information about whether a state is a bug, and even on the optimality gap (Propositions 3 and 5). Yet, while they do sometimes outperform our practical biases, the performance gap is generally not huge.

Beyond investigating further techniques to increase testing performance, the primary challenges for policy testing in our view remain fault analysis (which specific policy decisions cause sub-optimality), extension to richer planning paradigms, and connections to policy re-training.

**Acknowledgments**

This work was funded by the European Union’s Horizon Europe Research and Innovation program under the grant agreement TUPLES No 101070149, as well as by DFG Grant 389792660 as part of TRR 248 (CPEC, https://perspicuous-computing.science). Thanks to Rebecca Eifler for providing the Beluga PDDL domain and training the respective ASNet policy.
References


