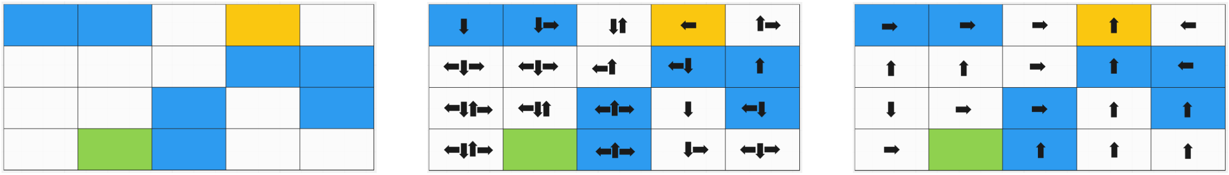
[Re]: Supplementary Material for Value Alignment Verification

Anonymous Author(s)
Affiliation
Address
`email`

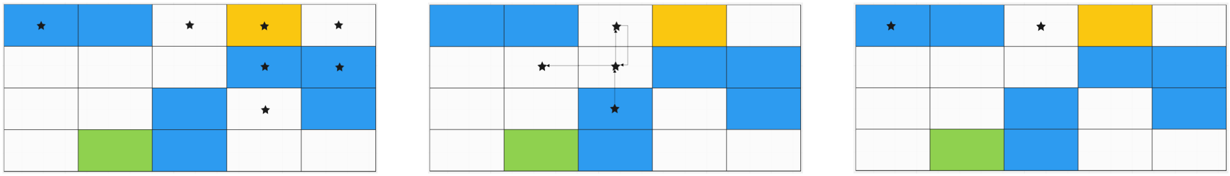
1 Sample Gridworld

We present a sample gridworld (Figure 1a), a sample human agent's optimal policy (Figure 1b), a sample robot's optimal policy (Figure 1c). In the gridworld, there are three different kinds of state (blue, yellow, and white, with state in green color denoting the terminal state). The bold arrows show the movements as per the optimal policy. Clearly, the robot is not aligned with the human agent. Further, we also depict the state queries (states marked with $\star$) asked by different testers as per the above human agent. Query states as per Critical States Tester in Figure 2a, Set Cover Optimal Teaching Tester (SCOT Tester) in Figure 2b and ARP Black Box Tester in Figure 2c. Additionally, with the aid of arrows, we depict the corresponding (maximally informative) trajectory with the state queries for SCOT Tester.



(a) Sample gridworld (b) Human policy (c) Robot policy

Figure 1: Sample Gridworld and Policies



(a) Critical States Tester (b) Set Cover Optimal Teaching Tester (c) ARP Black Box Tester

Figure 2: Query states for different heuristic testers on a sample gridworld

2 Additional Results

In the plots, tables, and following discussion, `rwt` indicates Reward Weight Queries Tester, `rt` indicates Reward Queries Tester, `vft` indicates Value Function Queries Tester, `ptt` indicates Preference Trajectory Queries Tester, `cst` indicates Critical States Tester, `scott` indicates SCOT Tester, and `arpbbt` indicates ARP Black Box Tester. Also note that, by performance metrics we refer accuracy, false positive rate, false negative rate, and number of test queries. `rwt`, `rt`, `rwt`, and `rt` are together referred to as algorithms. `cst`, `scott`, and `arpbbt` are together referred to as heuristics.

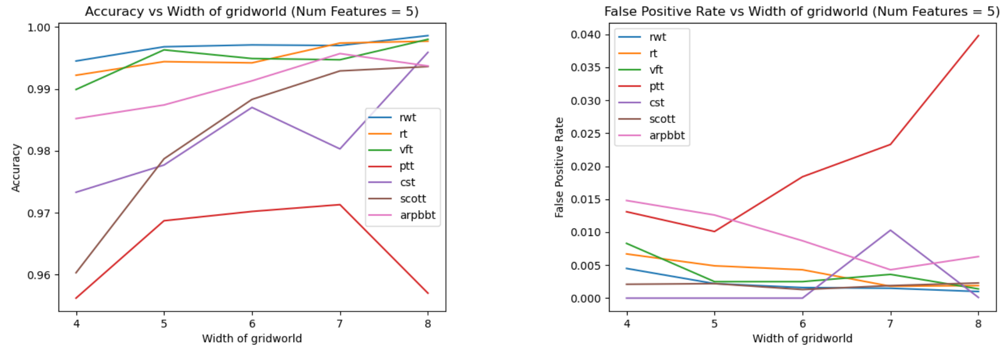
2.1 Algorithms and Heuristics

Table 1 details the performance metrics for all algorithms and heuristics. We fixed the feature size at 5. The feature size (or the number of features) is equal to the dimension of state-feature ϕ. Therefore, if $\phi : S \to \mathbb{R}^k \Rightarrow$ feature size $= k$.

Submitted to ML Reproducibility Challenge 2021. Do not distribute.

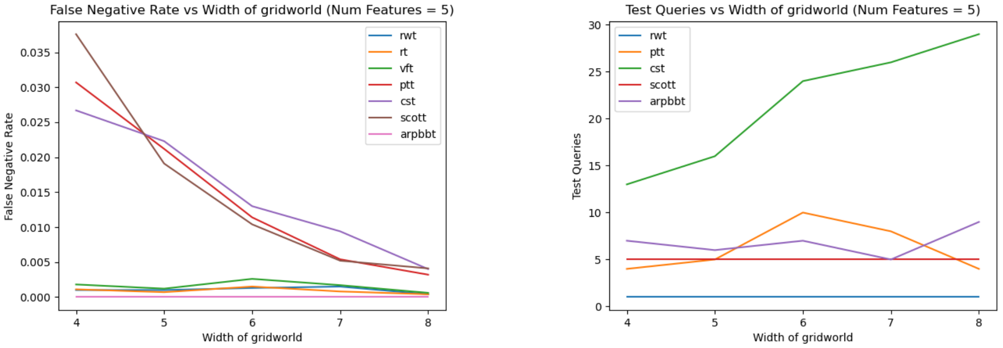
We observed that the varying width of a gridworld did not affect significantly the accuracy (Figure 3a), false positive (Figure 3b), and false negative rates (Figure 3c), while the number of test queries (Figure 3d) increased for heuristics except for `scott`, because the number of test queries for `scott` is equal to the maximum length of a trajectory (here, it is equal to 5). The accuracy for all the testers is extremely high, while the false positives and false negatives are exceedingly low, which indicates the ability of the algorithms and heuristics to identify alignment (or misalignment) between a robot and a human agent.

In table 2, we vary the feature size from 3 to 8. We fixed the gridworld width at 8. We observed no significant variation in accuracy (Figure 4a), false positive (Figure 4b), and false negative rates (Figure 4c). The trend for the number of test queries (Figure 4d) increases for `rt`, `vft`, and `arpbbt` while stays the same for `rwt` and `scott`. Unlike with different gridworld widths, the number of test queries stays similar with different feature sizes for `cst`.



(a) Accuracy

(b) False Positive Rate



(c) False Negative Rate

(d) Number of Test Queries

Figure 3: Tester performance for different gridworld widths

2.2 Diagonal Actions

Table 3 details the performance metrics for `rwt` and all heuristics in a gridworld with diagonal movements allowed. Again, since we varied the width of a gridworld, we fixed the feature size at 5. We did not perform additional experiments with `rt`, `vft`, and `ptt`, since they are reducible to `rwt`. We concluded there is no observable difference from the results in gridworlds without the diagonal movements, i.e., the accuracy (Figure 5a) is extremely high, and the false positive (Figure 5b), and false negative rates (Figure 5c) are significantly low. Additionally, the number of test queries (Figure 5d) increases for `cst` with an increase in the width of a gridworld.

2.3 Non-linear reward and state-feature relationship

Table 4 details the performance metrics for `rwt` and `cst` in gridworlds of different widths and non-linear relationship between reward and state-feature. The corresponding plot can be found in Figure 6 for *cubic* and Figure 7 for *exponential* relationship. These plots and tables contain all the gridworld widths ranging from 4 to 8 (naturally, the feature size is

Table 1: Different testers versus gridworld widths

Tester	Width	Accuracy	False positive rate	False negative rate	Number of queries
`rwt`	4	0.995 ± 0.013	0.005 ± 0.011	0.001 ± 0.005	1
	5	0.997± 0.008	0.002± 0.006	0.006± 0.004	1
	6	0.997± 0.007	0.002± 0.005	0.001± 0.005	1
	7	0.997± 0.008	0.002± 0.005	0.002± 0.006	1
	8	0.999± 0.004	0.001± 0.004	0.000± 0.002	1
`rt`	4	0.992± 0.015	0.007± 0.014	0.001± 0.004	5
	5	0.994± 0.013	0.005± 0.012	0.001± 0.004	5
	6	0.994± 0.012	0.004± 0.010	0.002± 0.007	5
	7	0.997± 0.006	0.002± 0.005	0.001± 0.004	5
	8	0.998± 0.005	0.002± 0.004	0.000± 0.002	5
`vft`	4	0.990± 0.019	0.008± 0.017	0.002± 0.007	25
	5	0.996± 0.008	0.003± 0.007	0.001± 0.004	25
	6	0.995± 0.011	0.003± 0.007	0.003± 0.009	25
	7	0.995± 0.021	0.004± 0.017	0.002± 0.007	25
	8	0.998± 0.005	0.001± 0.005	0.001± 0.002	25
`ptt`	4	0.956± 0.041	0.013± 0.028	0.031± 0.034	4
	5	0.969± 0.035	0.010± 0.026	0.021± 0.027	5
	6	0.970± 0.039	0.018± 0.039	0.011± 0.019	10
	7	0.971± 0.040	0.023± 0.041	0.005± 0.012	8
	8	0.957± 0.054	0.040± 0.055	0.003± 0.006	4
`cst`	4	0.973± 0.043	0.000± 0.000	0.027± 0.043	13
	5	0.978± 0.048	0.000± 0.000	0.022± 0.048	16
	6	0.987± 0.018	0.000± 0.000	0.013± 0.018	24
	7	0.980± 0.100	0.010± 0.099	0.009± 0.019	26
	8	0.996± 0.007	0.000± 0.001	0.004± 0.007	29
`scot`	4	0.960± 0.041	0.002± 0.005	0.04± 0.041	5
	5	0.979± 0.024	0.002± 0.005	0.019± 0.023	5
	6	0.988± 0.017	0.001± 0.004	0.010± 0.017	5
	7	0.993± 0.010	0.002± 0.005	0.005± 0.009	5
	8	0.994± 0.010	0.002± 0.005	0.004± 0.008	5
`arpbbt`	4	0.985± 0.036	0.015± 0.036	0.000± 0.000	7
	5	0.987± 0.040	0.013± 0.040	0.000± 0.000	6
	6	0.991± 0.027	0.009± 0.027	0.000± 0.000	7
	7	0.996± 0.016	0.004± 0.016	0.000± 0.000	5
	8	0.994± 0.022	0.006± 0.022	0.000± 0.000	9

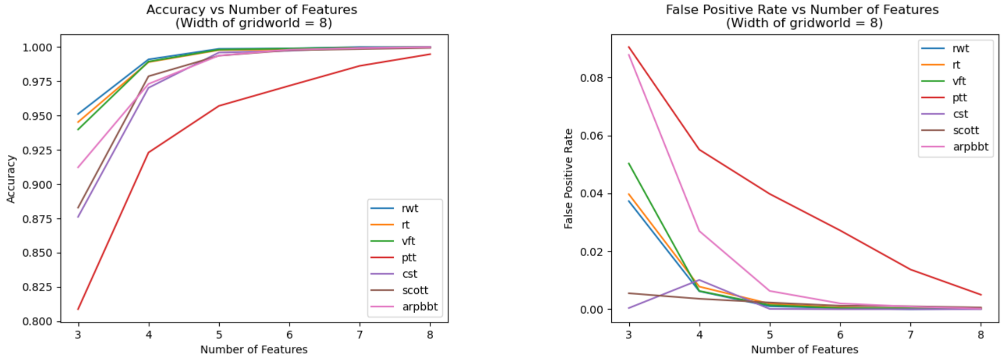
fixed at 5), which are not present in the Results section of the paper. Note that, in *cubic* we approximate the *linear* relationship only when $w^T\phi(s) \approx 0$ while *exponential* completely ignores the *linear* relationship. We observed that since the accuracy is high for *cubic* and low for *exponential* relationships, the false positive rates are low and high respectively.

2.4 Critical States Tester with different thresholds

Table 5 details the performance metrics for `cst` in different gridworld widths and different threshold values (0.0001, 0.2, and 0.8). As noted earlier, we observed that the number of queries decreases with strict threshold values such as 0.8, which results in reduced verification abilities. The corresponding plots for performance metrics are present in Figure 8.

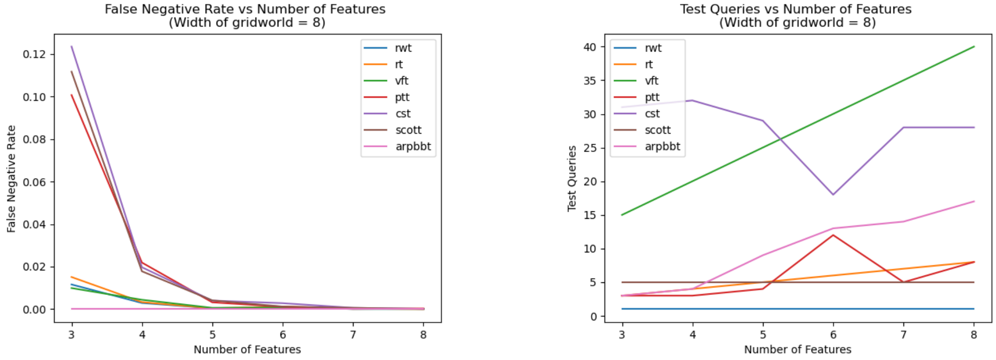
2.5 Time profile for algorithms and heuristics

Table 6 details the time taken by all algorithms and heuristics to verify 100 different robots for a single human agent. We observed that due to the complexity of Set Cover Optimal Teaching Heuristic, `scott` takes the maximum time



(a) Accuracy

(b) False Positive Rate



(c) False Negative Rate

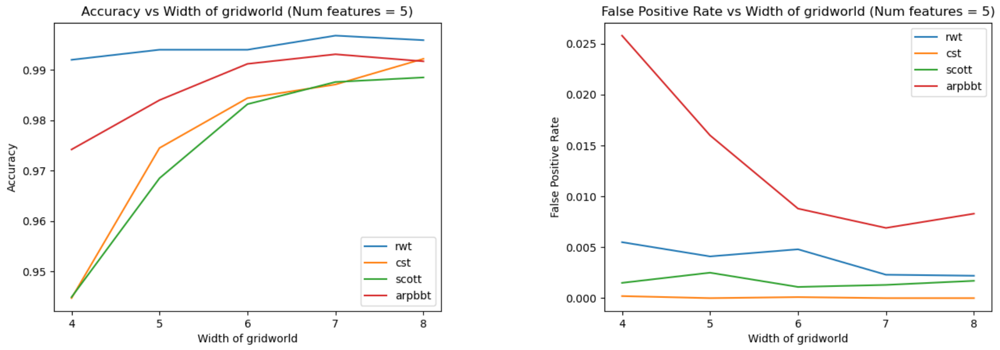
(d) Number of Test Queries

Figure 4: Tester performance for different number of features

to complete the verification while `rwt` takes minimum time. Also, as expected, with the increase in the width of a gridworld, the time taken to verify increases. For most of the algorithms and heuristics, the time taken increases 2X when width of a gridworld increases from 4 to 8 while for `scott`, the time taken increases by at least 3X.

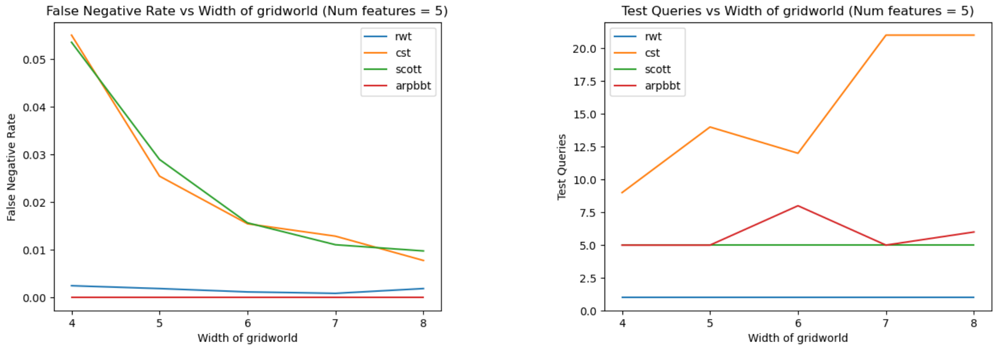
Table 2: Different testers versus features sizes

Tester	**Feature size**	**Accuracy**	**False positive rate**	**False negative rate**	**Number of queries**
`rwt`	3	0.951± 0.051	0.037± 0.045	0.012± 0.034	1
	4	0.991± 0.015	0.006± 0.013	0.003± 0.007	1
	5	0.999± 0.004	0.001± 0.004	0.000± 0.002	1
	6	0.999± 0.006	0.000± 0.002	0.001± 0.005	1
	7	1.000± 0.001	0.000± 0.000	0.000± 0.001	1
	8	1.000± 0.002	0.000± 0.001	0.000± 0.001	1
`rt`	3	0.945± 0.065	0.040± 0.056	0.015± 0.04	3
	4	0.989± 0.017	0.008± 0.013	0.003± 0.010	4
	5	0.998± 0.005	0.002± 0.004	0.000± 0.002	5
	6	0.998± 0.007	0.001± 0.003	0.001± 0.006	6
	7	1.000± 0.002	0.000± 0.001	0.000± 0.002	7
	8	1.000± 0.004	0.000± 0.004	0.000± 0.000	8
`vft`	3	0.940± 0.069	0.050± 0.068	0.010± 0.030	15
	4	0.989± 0.018	0.006± 0.012	0.004± 0.014	20
	5	0.998± 0.005	0.001± 0.005	0.001± 0.002	25
	6	0.999± 0.004	0.000± 0.002	0.001± 0.004	30
	7	1.000± 0.002	0.000± 0.002	0.000± 0.001	35
	8	1.000± 0.001	0.000± 0.001	0.000± 0.001	40
`ptt`	3	0.809± 0.094	0.091± 0.109	0.101± 0.080	3
	4	0.923± 0.075	0.055± 0.073	0.022± 0.040	3
	5	0.957± 0.054	0.040± 0.055	0.003± 0.006	4
	6	0.972± 0.042	0.027± 0.042	0.001± 0.004	12
	7	0.986± 0.030	0.014± 0.030	0.000± 0.001	5
	8	0.995± 0.012	0.005± 0.012	0.000± 0.002	8
`cst`	3	0.876± 0.097	0.000± 0.002	0.124± 0.097	31
	4	0.970± 0.101	0.010± 0.099	0.020± 0.025	32
	5	0.996± 0.007	0.000± 0.001	0.004± 0.007	29
	6	0.997± 0.008	0.000± 0.000	0.003± 0.008	18
	7	0.999± 0.002	0.000± 0.000	0.001± 0.002	28
	8	1.000± 0.001	0.000± 0.000	0.000± 0.001	28
`scott`	3	0.883± 0.090	0.006± 0.011	0.112± 0.089	5
	4	0.979± 0.029	0.004± 0.008	0.018± 0.029	5
	5	0.994± 0.010	0.002± 0.005	0.004± 0.008	5
	6	0.998± 0.005	0.001± 0.003	0.001± 0.004	5
	7	0.998± 0.004	0.001± 0.003	0.001± 0.003	5
	8	0.999± 0.003	0.001± 0.002	0.000± 0.001	5
`arpbbt`	3	0.912± 0.110	0.088± 0.110	0.000± 0.000	3
	4	0.973± 0.063	0.027± 0.063	0.000± 0.000	4
	5	0.994± 0.022	0.006± 0.022	0.000± 0.000	9
	6	0.998± 0.008	0.002± 0.008	0.000± 0.000	13
	7	0.999± 0.003	0.001± 0.003	0.000± 0.000	14
	8	1.000± 0.001	0.000± 0.001	0.000± 0.000	17



(a) Accuracy

(b) False Positive Rate



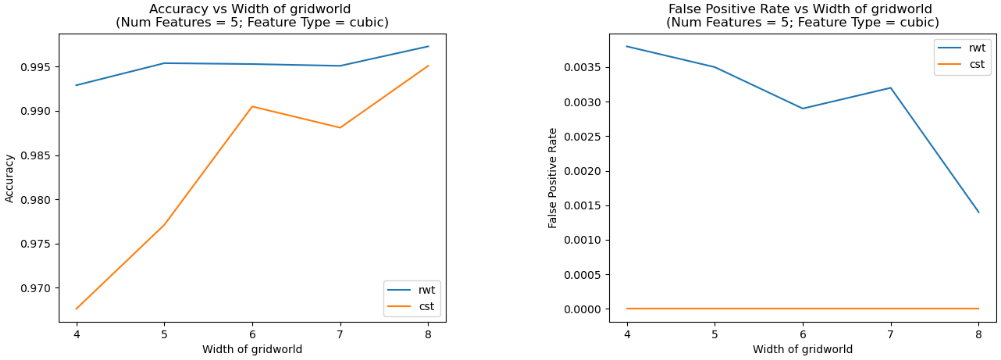
(c) False Negative Rate

(d) Number of Test Queries

Figure 5: Tester performance for different gridworld widths with an extended action space

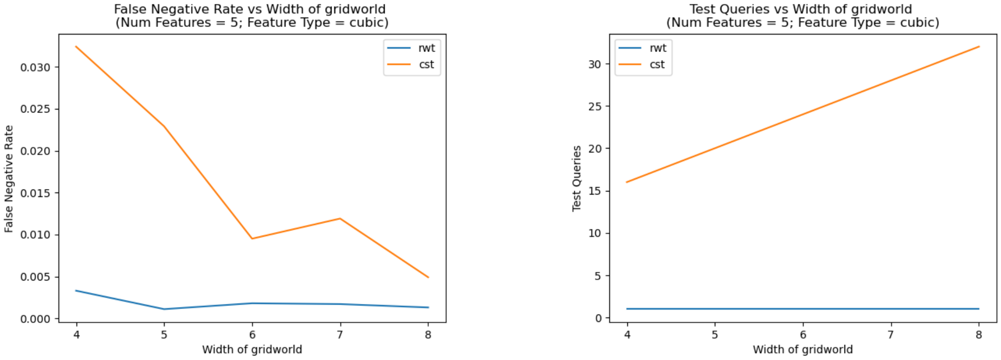
Table 3: Different testers versus gridworld widths with extended action space

Tester	**Width**	**Accuracy**	**False positive rate**	**False negative rate**	**Number of queries**
`rwt`	4	0.992± 0.017	0.006± 0.015	0.003± 0.010	1
	5	0.994± 0.013	0.004± 0.009	0.002± 0.007	1
	6	0.994± 0.013	0.005± 0.011	0.001± 0.004	1
	7	0.997± 0.008	0.002± 0.006	0.001± 0.003	1
	8	0.996± 0.008	0.002± 0.005	0.002± 0.005	1
`cst`	4	0.945± 0.055	0.000± 0.001	0.055± 0.055	9
	5	0.975± 0.039	0.000± 0.000	0.026± 0.039	14
	6	0.984± 0.017	0.000± 0.001	0.016± 0.017	12
	7	0.987± 0.022	0.000± 0.000	0.013± 0.022	21
	8	0.992± 0.011	0.000± 0.000	0.008± 0.011	21
`scott`	4	0.945± 0.049	0.002± 0.005	0.054± 0.049	5
	5	0.969± 0.036	0.003± 0.009	0.029± 0.034	5
	6	0.983± 0.023	0.001± 0.004	0.016± 0.023	5
	7	0.988± 0.023	0.001± 0.003	0.011± 0.022	5
	8	0.989± 0.017	0.002± 0.005	0.010± 0.017	5
`arpbbt`	4	0.974± 0.068	0.026± 0.068	0.000± 0.000	5
	5	0.984± 0.059	0.016± 0.059	0.000± 0.000	6
	6	0.991± 0.039	0.009± 0.039	0.000± 0.000	8
	7	0.993± 0.029	0.007± 0.029	0.000± 0.000	5
	8	0.992± 0.037	0.008± 0.037	0.000± 0.000	6



(a) Accuracy

(b) False Positive Rate



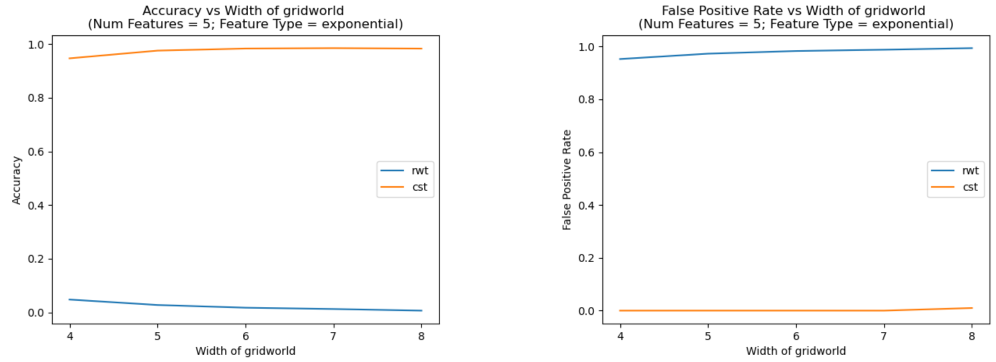
(c) False Negative Rate

(d) Number of Test Queries

Figure 6: Tester performance for cubic reward - state features relationship

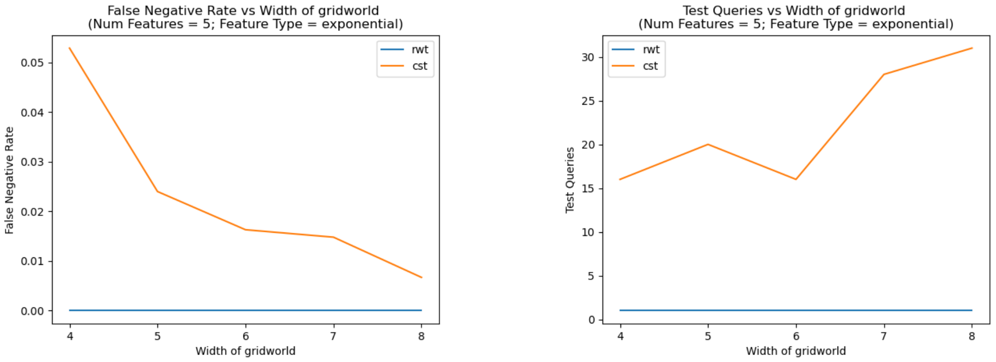
Table 4: Different testers versus gridworld widths with non-linear reward state-feature relationships

Tester	Width	Accuracy	False positive rate	False negative rate	Number of queries
`rwt` (cubic)	4	0.993± 0.013	0.004± 0.007	0.003± 0.011	1
	5	0.995± 0.011	0.004± 0.009	0.001± 0.005	1
	6	0.995± 0.008	0.003± 0.006	0.002± 0.005	1
	7	0.995± 0.008	0.003± 0.006	0.002± 0.005	1
	8	0.997± 0.006	0.001± 0.005	0.001± 0.004	1
`rwt` (exponential)	4	0.048± 0.052	0.953± 0.052	0.000± 0.000	1
	5	0.027± 0.037	0.973± 0.037	0.000± 0.000	1
	6	0.017± 0.021	0.983± 0.021	0.000± 0.000	1
	7	0.012± 0.019	0.988± 0.019	0.000± 0.000	1
	8	0.006± 0.012	0.994± 0.012	0.000± 0.000	1
`cst` (cubic)	4	0.968± 0.040	0.000± 0.000	0.032± 0.040	16
	5	0.977± 0.031	0.000± 0.000	0.023± 0.031	20
	6	0.991± 0.015	0.000± 0.000	0.010± 0.015	24
	7	0.988± 0.019	0.000± 0.000	0.012± 0.019	28
	8	0.995± 0.010	0.000± 0.000	0.005± 0.010	32
`cst` (exponential)	4	0.947± 0.051	0.000± 0.001	0.053± 0.051	16
	5	0.976± 0.023	0.000± 0.001	0.024± 0.023	20
	6	0.984± 0.022	0.000± 0.001	0.016± 0.022	16
	7	0.985± 0.027	0.000± 0.001	0.015± 0.027	28
	8	0.983± 0.099	0.010± 0.099	0.007± 0.010	31



(a) Accuracy

(b) False Positive Rate



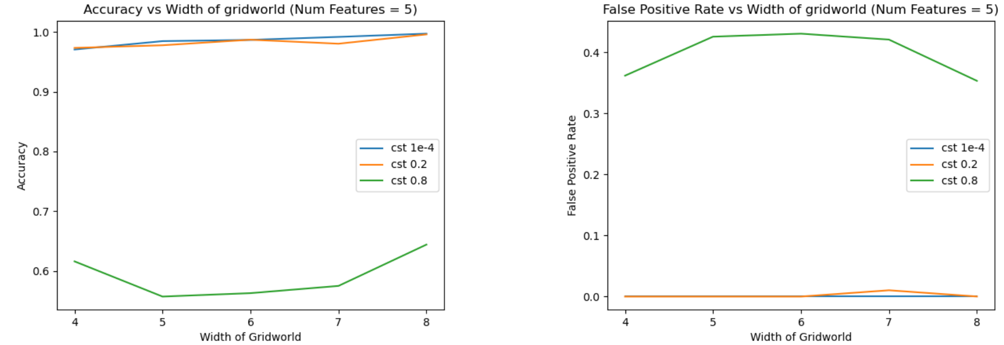
(c) False Negative Rate

(d) Number of Test Queries

Figure 7: Tester performance for exponential reward - state features relationship

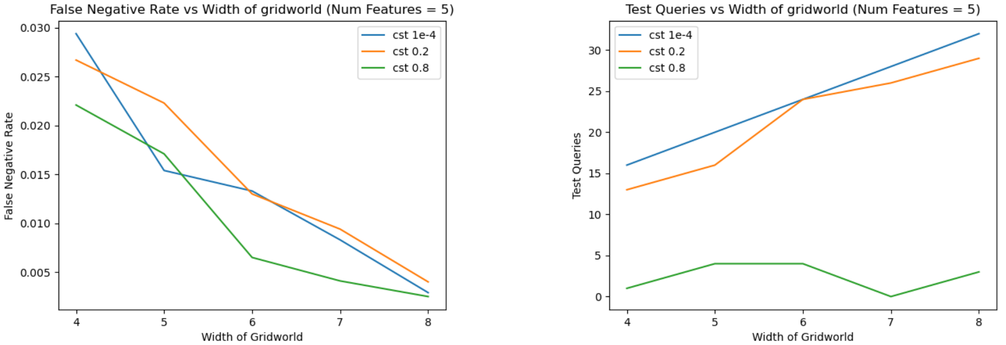
Table 5: Critical states tester with different thresholds

Tester	**Width**	**Accuracy**	**False positive rate**	**False negative rate**	**Number of queries**
`cst` (threshold = 0.0001)	4	0.971± 0.036	0.000± 0.000	0.029± 0.036	16
	5	0.985± 0.023	0.000± 0.000	0.015± 0.023	20
	6	0.987± 0.018	0.000± 0.000	0.013± 0.018	24
	7	0.992± 0.016	0.000± 0.000	0.008± 0.016	28
	8	0.997± 0.007	0.000± 0.000	0.003± 0.007	32
`cst` (threshold = 0.2)	4	0.973± 0.043	0.000± 0.000	0.027± 0.043	13
	5	0.978± 0.048	0.000± 0.000	0.022± 0.048	16
	6	0.987± 0.018	0.000± 0.000	0.013± 0.018	24
	7	0.980± 0.100	0.010± 0.099	0.009± 0.019	26
	8	0.996± 0.007	0.000± 0.001	0.004± 0.007	29
`cst` (threshold = 0.8)	4	0.616± 0.447	0.362± 0.463	0.022± 0.032	1
	5	0.557± 0.469	0.426± 0.483	0.017± 0.031	4
	6	0.563± 0.482	0.431± 0.488	0.007± 0.013	4
	7	0.575± 0.485	0.421± 0.488	0.004± 0.009	0
	8	0.644± 0.468	0.354± 0.470	0.003± 0.008	3



(a) Accuracy

(b) False Positive Rate



(c) False Negative Rate

(d) Number of Test Queries

Figure 8: Critical states tester with different thresholds

Table 6: Different testers versus time taken

Tester	**Width**	**Time (in sec)**
`rwt`	4	15.980
	5	21.966
	6	30.335
	7	31.186
	8	35.093
`rt`	4	34.399
	5	53.089
	6	54.039
	7	66.959
	8	73.450
`vft`	4	39.110
	5	55.968
	6	41.702
	7	69.879
	8	70.993
`ptt`	4	21.164
	5	23.315
	6	29.121
	7	33.048
	8	40.644
`cst`	4	18.714
	5	24.487
	6	30.871
	7	32.206
	8	36.563
`scott`	4	95.301
	5	143.555
	6	195.014
	7	257.184
	8	322.146
`arpbbt`	4	20.763
	5	25.800
	6	28.041
	7	34.407
	8	38.532