A. Additional experimental data

In section 5, we presented data on the results of running various algorithms against a set of demonstrators, reporting the reward obtained according to the true reward function when using the inferred reward with an optimal planner, as a percentage of the maximum possible true reward. Table 1 shows the percentage reward obtained for all combinations of algorithms and demonstrators. We also measure the accuracy of the planner and reward at predicting the demonstrator's actions in new gridworlds where the rewards are the same but the wall locations have changed. These results are presented in Table 2. Note that there are often multiple optimal actions at a given state, which makes it challenging to get high accuracy.

Table 1. Percent reward obtained when the algorithm (column) is used to infer the bias of the demonstrator (row). The optimal and Boltzmann algorithms assume a fixed model of the demonstrator and train the VIN to mimic the model before performing reward inference (and were used in figure 4). We also include the four flavors of Algorithm 2 that were plotted in figure 5. The VI algorithm uses a differentiable implementation of soft value iteration as the planner instead of a VIN (used in section 5.3). The demonstrators are the optimal agent, the biased agents of figure 2, and versions of each of these agents with Boltzmann noise.

Agent	Optimal	Boltzmann	Algorithm 1	Coord w/ init	Joint w/ init	Coord w/o init	Joint w/o init	VI
Average	67.0 ± 2.7	79.7 ± 0.8	89.5 ± 0.7	85.0 ± 0.5	86.4 ± 0.6	-3.9 ± 0.7	2.6 ± 1.0	71.9 ± 3.0
Optimal	87.3 ± 1.0	73.9 ± 2.3	86.9 ± 1.6	86.2 ± 1.6	88.5 ± 1.1	-4.2 ± 1.2	2.6 ± 3.7	98.1 ± 0.1
Naive	86.4 ± 0.9	74.4 ± 1.6	91.1 ± 0.8	84.6 ± 1.2	87.5 ± 0.9	-3.2 ± 1.3	2.6 ± 3.7	96.1 ± 0.1
Sophisticated	87.5 ± 1.1	77.1 ± 1.6	91.8 ± 1.3	83.6 ± 1.3	87.9 ± 1.0	-3.6 ± 1.4	2.6 ± 3.7	96.7 ± 0.1
Myopic	82.8 ± 0.8	77.0 ± 1.2	81.0 ± 2.8	80.6 ± 0.8	82.6 ± 1.0	-5.5 ± 2.8	2.6 ± 3.7	87.5 ± 0.2
Overconfident	87.5 ± 1.2	70.7 ± 1.7	82.1 ± 1.4	83.9 ± 1.5	86.7 ± 1.2	-2.7 ± 1.1	2.6 ± 3.7	97.5 ± 0.1
Underconfident	88.0 ± 0.8	74.7 ± 1.6	86.7 ± 1.2	86.1 ± 1.5	88.5 ± 1.0	-2.4 ± 1.4	2.6 ± 3.7	98.9 ± 0.2
Boltzmann	8.5 ± 1.0	90.7 ± 1.3	91.4 ± 0.8	88.4 ± 1.6	91.3 ± 0.9	-3.0 ± 1.9	2.6 ± 3.7	8.7 ± 0.1
B-Naive	52.8 ± 2.3	77.3 ± 2.9	98.5 ± 0.1	82.5 ± 2.4	75.8 ± 2.9	-8.3 ± 4.5	2.6 ± 3.7	47.7 ± 0.2
B-Sophisticated	51.5 ± 2.1	74.5 ± 2.8	98.8 ± 0.2	80.1 ± 1.5	77.0 ± 2.3	-8.7 ± 3.9	2.6 ± 3.7	48.0 ± 0.2
B-Myopic	77.7 ± 1.1	90.8 ± 0.6	95.6 ± 1.0	91.5 ± 0.6	91.9 ± 0.5	-2.4 ± 2.1	2.6 ± 3.7	83.4 ± 0.1
B-Overconfident	7.0 ± 0.9	84.1 ± 2.3	79.2 ± 2.3	81.4 ± 2.8	86.3 ± 1.3	-0.8 ± 1.6	2.6 ± 3.7	8.7 ± 0.1
B-Underconfident	86.7 ± 0.9	91.3 ± 0.7	91.2 ± 0.7	90.7 ± 1.0	92.4 ± 0.8	-1.8 ± 1.2	2.6 ± 3.7	92.1 ± 0.1

Table 2. Accuracy when predicting the demonstrator's actions (row) on new gridworlds using the planner and reward inferred by the algorithm (column). Algorithms and demonstrators are the same as in Table 1.

Agent	Optimal	Boltzmann	Algorithm 1	Coord w/ init	Joint w/ init	Coord w/o init	Joint w/o init	VI
Optimal	61.3 ± 0.4	59.8 ± 0.4	62.0 ± 0.3	62.8 ± 0.2	63.6 ± 0.3	63.0 ± 0.2	72.4 ± 0.1	25.7 ± 0.1
Naive	60.1 ± 0.3	59.4 ± 0.3	58.6 ± 0.3	61.3 ± 0.3	61.8 ± 0.3	61.0 ± 0.3	71.1 ± 0.1	24.9 ± 0.1
Sophisticated	60.5 ± 0.4	59.2 ± 0.4	59.3 ± 0.3	61.0 ± 0.3	62.0 ± 0.4	61.2 ± 0.3	71.2 ± 0.1	24.9 ± 0.1
Myopic	54.1 ± 0.4	53.5 ± 0.5	54.9 ± 0.5	55.6 ± 0.2	56.1 ± 0.3	56.0 ± 0.1	62.8 ± 0.1	20.4 ± 0.1
Overconfident	61.6 ± 0.4	60.1 ± 0.4	61.8 ± 0.4	63.3 ± 0.3	63.7 ± 0.3	63.1 ± 0.2	72.8 ± 0.1	25.9 ± 0.1
Underconfident	60.9 ± 0.4	59.5 ± 0.4	61.4 ± 0.3	62.4 ± 0.3	62.9 ± 0.3	62.5 ± 0.3	72.0 ± 0.1	25.5 ± 0.1
Boltzmann	56.7 ± 1.1	60.5 ± 0.4	60.9 ± 0.3	60.3 ± 0.2	60.8 ± 0.3	62.3 ± 0.3	67.1 ± 0.5	24.2 ± 0.1
B-Naive	56.6 ± 0.8	59.8 ± 0.8	60.4 ± 0.1	60.3 ± 0.2	60.5 ± 0.7	59.9 ± 0.3	68.5 ± 0.3	23.7 ± 0.1
B-Sophisticated	57.6 ± 0.7	60.2 ± 0.7	60.5 ± 0.2	60.5 ± 0.2	61.2 ± 0.3	60.1 ± 0.3	68.5 ± 0.3	23.7 ± 0.1
B-Myopic	56.3 ± 0.2	56.9 ± 0.4	55.9 ± 0.2	56.5 ± 0.2	57.0 ± 0.2	56.3 ± 0.1	62.4 ± 0.1	20.3 ± 0.0
B-Overconfident	56.9 ± 1.1	60.7 ± 0.4	61.3 ± 0.3	60.9 ± 0.2	61.6 ± 0.3	62.7 ± 0.2	68.0 ± 0.5	24.2 ± 0.1
B-Underconfident	62.4 ± 0.3	63.1 ± 0.4	63.4 ± 0.2	63.0 ± 0.1	63.6 ± 0.1	63.5 ± 0.2	72.2 ± 0.1	25.4 ± 0.1