B SUPPLEMENTARY FILE FOR AAMAS PAPER: A HIERARCHICAL BAYESIAN PROCESS FOR INVERSE RL IN PARTIALLY-CONTROLLED ENVIRONMENTS

We discuss the complete definitions of the MDPs for the two domains, the observation variables, η distributions and α priors, and the input trajectories in the supplementary material. We also show a video of our simulation of the noisy onion sorting domain.

B.1 MDP Definitions

B.1.1 Gridworld.

States: 25 states corresponding to a 5 x 5 grid. The initial starting state is uniformly random.

Actions: Move Up, Move Down, Move Left, Move Right

Transition Function: If legal, an action "succeeds" (moves the agent in the named direction) with a 90% chance, the remaining probability mass is divided equally to the other three action's moves. In the event that an action would move the agent out of a gridworld, the current state receives the respective mass.

True Reward: Four reward features are defined:
Top Left
Top Right
Bottom Left
Bottom Right

In our experiments one of these features is chosen at random and given a weight of 1.0, all others receive 0.0.

B.1.2 Onion Sort.

States: There are 15 states total, with each state representing a tuple of (Onion quality, Onion Position, Gripper Position)
Unknown, On Conveyor, Conveyor
Unknown, Gripped, Conveyor
Unknown, Gripped, Inspection
Unknown, Gripped, Bin
Unknown, In Bin, Bin
Good, Gripped, Inspection
Good, Gripped, Bin
Good, In Bin, Bin
Good, Gripped, Conveyor
Good, On Conveyor, Conveyor
Blemished, Gripped, Inspection
Blemished, Gripped, Bin
Blemished, In Bin, Bin
Blemished, Gripped, Conveyor
Blemished, On Conveyor, Conveyor

The initial starting state is (Unknown, On Conveyor, Conveyor).

Actions: Grip Onion, Inspect Onion, Release Onion, Move to Inspection, Move to Conveyor, Move to Bin

Transition Function:

All actions may be performed in all states, however, most will deterministicly stay in the current state. The exceptions are as follows:
Grip Onion - Move from an "On Conveyor" state to the associated "Gripped" state, with probability 1
Inspect Onion - Change from an "Unknown" onion quality state to either "Good" or "Blemished", with a 50% chance of either. May only be performed in states where the gripper location is "Inspection" and the onion quality is "Unknown"
Release Onion - Move from a "Gripped" state to the associated "On Conveyor" or "In Bin" state (depending upon the gripper location),with probability 1
Move to Inspection - Move from a "Conveyor" or "Bin" gripper location state to the associated "Inspection" state with probability 1
Move to Conveyor - Move from an "Inspection" or "Bin" gripper location state to the associated "Conveyor" state with probability 1
Move to Bin - Move from an "Inspection" or "Conveyor" gripper location state to the associated "Bin" state with probability 1

True Reward: Four reward features are defined:
Release good onion on the conveyor
Release good onion in the bin
Release blemished onion on the conveyor
Release blemished onion in the bin

The true reward function assigns the value 1 to releasing good onions back on the conveyor, 1 to releasing blemished onions in the bin, and 0 for the others. This results in expert behavior of picking up each onion from the conveyor belt, inspecting it, on finding it blemished placing it in the bin, otherwise placing it back on the conveyor.

B.2 Experiment Observations (Ω)

This section describes the set of observations Ω for the two MDPs used in our experiments and how they are generated from simulation (Gridworld) or sensor stream (Onion sort). Additionally, we give the true observation function in Gridworld where it is known.

B.2.1 Gridworld.

Four observations are defined:
Top Left (0, 0), Top Right (0, 4) , Bottom Left (4, 0), Bottom Right (4, 4)

The true observation model for our Gridworld is as follows:

$$Pr(\omega|S) \propto exp(-\frac{\sqrt{(\omega_x - S_x)^2 + (\omega_y - S_y)^2}}{2})$$

where ω_x is the x coordinate of the observation.

Observation models for the confounding elements were randomly generated exponential distributions. When informative α hyperparameters are used, they are calculated as $\alpha_{s,a}[\omega] = \lceil 5 * Pr(\omega|s,a) \rceil \; \forall \; \omega, s, a$ and $\alpha_{\mathcal{E}}[\omega] = \lceil 3 * Pr(\omega|\mathcal{E}) \rceil \; \forall \; \mathcal{E}, s, a$

B.2.2 Onion sort.

Twenty-two observations are defined:
Gripper Position {1 - 20}, Bright Onion, Dark Onion

A blob finder trained on two colors was used to generate features (color blobs) from the learner's RGB video stream. Blue blobs were trained on the gripper, orange blobs were trained on onions. Each blob was then classified into one of the 22 observations as follows.

The learner's viewport was divided into a 5 x 4 grid of equal size rectangles, each one corresponding to a single Gripper Position observation. For blue blobs, we classify the observation using the center point of a blob on the 2d view port, whichever rectangle it

appeared inside would decide which Gripper Position observation to generate.

To classify an orange blob into **Bright Onion** and **Dark Onion** we converted the blob's pixels to grayscale using RGB-to-Luma ITU BT.709, then found the proportion of pixels whose value is less than 30, **Dark Onion** is selected if $\geq 12\%$.

B.3 η Definitions

These procedures were used to compute the η distribution for each observation ω in our experiments.

B.3.1 Gridworld.

All four observations used the same procedure:

Let $s \in Z$ be the true source of an observation and $U \in Z$ be a random choice. Initialize all values of η to $\frac{0.4}{|\mathcal{E}|+1}$.

Then, $\begin{cases} \eta[s] = 0.6 & \text{with 80\% probability} \\ \eta[U] = 0.6 & \text{with 20\% probability} \end{cases}$

B.3.2 Onion sort.

Gripper Position

As the apparent size of the gripper changed very little during the experiments we used the area of a gripper blob to set the η distribution, with areas closer to the average getting the most probability mass assigned to the subject.

$\eta_{subject} = 0.5 + 1842.35 * \mathcal{N}(4000, 2100)$

where $\mathcal{N}$ is the PDF for the normal distribution. Leftover probability mass was assigned to the person-in-a-blue-shirt confounding element.

Onion Observation

Both **Bright Onion** and **Dark Onion** use the same η distributions. For a given onion blob first find the closest gripper blob. Compute the longest cross section of both blobs; if the distance between the two blobs' centers is less than the sum of these cross sections, call them adjacent.

If adjacent to the gripper, the η for the subject is set to 0.8, with remaining mass given to other onions (0.1) and the foreground person (0.1) confounding elements. Otherwise, if the area of the blob is greater than 500 pixels the other onions confounding element was set to 0.6, foreground person 0.25, and subject 0.15. Otherwise, the foreground person is set to 0.6, other onions set to 0.25, and subject 0.15.

B.4 Example α priors

Here we show some α values for the onion sort experiment, learned from the control trajectories.

State	Action	Gripper Positions	Bright Onion	Dark Onion
Good Gripped Conveyor	Release Onion	0,0,142,26,0, 0,290,84,0,0, 198,340,0,0,20, 194,0,0,0,0	469	123
Blemished Gripped Inspection	Move to Bin	0,0,517,0,0, 0,312,0,0,0, 0,0,0,0,0, 0,0,0,0,0	186	208

Notice the large values; this is due to the large amount of observations received from 10 trajectories. During experiments, we scale each α vector so that it sums to 100.

B.5 Example Grid world Observations

We show an example trajectory for the grid world with 9 confounding elements present and 20 observations per timestep. Underlined observations are those caused by the subject.

State	Action	Observations
0 x 3	Right	1, 1, 1, 2, 1, 3, 0, 1, 2, 1, 1, 1, 1, 3, 1, 1, 3, 1, 0, 3
1 x 3	Right	0, 1, 1, 1, 2, 0, 0, 3, 2, 1, 1, 0, 1, 3, 3, 1, 1, 3, 1, 0
2 x 3	Right	1, 3, 1, 1, 1, 1, 1, 2, 0, 2, 1, 1, 1, 0, 3, 2, 1, 0, 1, 3
3 x 3	Right	3, 2, 3, 3, 0, 2, 1, 0, 1, 2, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0
4 x 3	Down	1, 2, 0, 1, 0, 1, 0, 0, 0, 1, 1, 3, 1, 1, 0, 2, 1, 1, 1, 3
4 x 4	Down	3, 3, 0, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 2, 2, 0, 0, 0
4 x 3	Down	3, 3, 0, 0, 0, 1, 3, 0, 0, 2, 1, 1, 1, 2, 1, 0, 1, 0, 0, 1
4 x 4	Down	3, 3, 3, 3, 0, 0, 1, 0, 2, 1, 1, 0, 3, 2, 1, 1, 1, 2, 1, 1
4 x 4	Down	3, 3, 1, 3, 0, 0, 3, 2, 0, 1, 1, 0, 1, 0, 3, 1, 1, 1, 2, 1
4 x 4	Down	3, 0, 1, 3, 0, 2, 0, 0, 1, 0, 2, 1, 0, 3, 2, 0, 1, 2, 1, 1