Table 3: **Overview of notation.**

Variable	**Symbol**
Time	t
Principal	p
Agent	i
State	s
State vector	$\boldsymbol{s}$
State space	S
Agent's action space	A
Principal's action space	A^p
Action sequence	$a_{1:T} = \{a_1, a_2, \ldots, a_T\}$
Agent i's reward sequence	$\tilde{r}^i_{1:T} = \left\{\tilde{r}^i_1, \ldots, \tilde{r}^i_T\right\}$
Principal's reward sequence	$r^p_{1:T} = \{r^p_1, \ldots, r^p_T\}$
Transition function	$\mathcal{P}$
Agent i's policy	π^i
Principal's intervention policy	π^p
Agent's mean estimate of intervened rewards for action a	$\tilde{\mu}_a$
Number of adaptation steps	K
Number of meta-tasks for the planner	N
Principal's history of interventions and observed agent actions upto time t	$\mathcal{H}^p_t = \left\{a^p_1, a^i_1, a^p_2, a^i_2, \ldots, a^p_{t-1}, a^i_{t-1}\right\}$
Agent's history of actions taken and rewards observed upto time t	$\mathcal{H}^i_t = \left\{a^i_1, \tilde{r}^i_1, a^i_2, \tilde{r}^i_2, \ldots, a^i_{t-1}, \tilde{r}^i_{t-1}\right\}$

Table 4: **Notation for MERMAIDE** See Section Section 4 for their use.

Principal's policy parameter	$\theta \in \Theta$
Agent i's learning algorithm	$f^i \in \mathcal{F}$
Agent i's true action mean rewards	$\mu^i \sim \mathcal{U}$
Agent i's intervened action mean rewards	$\tilde{\mu}^i$
Principal's action at time t	$a^p_t \sim \pi^p_\theta\left(a^p_t \middle\| a^i_{t-1}, a^p_{t-1}, \hat{a}^i_t, h^p_{t-1}\right)$
Hidden state space of the principal's recurrent world model	H
Agent's action at time t	$a^i_t \sim \pi^i_t\left(a^i_t \middle\| \mathcal{H}^i_t\right), t = 1, \ldots, T$
Agent's reward at time t	$r^i_t \sim \mathcal{N}\left(\tilde{\mu}^i, \sigma^2\right)$
Principal's world model estimate of the agent's action probability distribution	$\hat{\pi}^i_\omega : A \times A^p \times H \to \Delta(A),\ \hat{\pi}^i_{\omega,0} \in A$
Principal's world model estimate of the latent state of the environment	$g^i_\omega : A \times A^p \times H \to H,\ g^i_{\omega,0} \in \Delta(H)$
Principal's world model hidden state embeding in the LSTM architecture	$h^i_t = g^i_\omega\left(a^i_{t-1}, a^p_{t-1}, h_{t-1}\right),\ t = 2, \ldots, T\ h^i_1 = g^i_{\omega,0}$

A Notation

For an overview of all symbols and variables used in this work, see Table 3 and Table 4.

B Additional Results

B.1 Description of the bandit algorithms

We provide a brief overview of the learning algorithms referred to in Section 6.

UCB. This is an Upper Confidence Bound based exploration-exploitation algorithm that follows the principle of optimism in the face of uncertainty. At each time step t, the bandit agent selects an action

$$a_t = \arg\max_a \tilde{\mu}_a + \beta \sqrt{\frac{\log t}{n_a}} \tag{5}$$

where n_a is the number of steps until t in which it previously selected the action a, $\tilde{\mu}_a$ is its corresponding mean estimate for the experienced rewards $\tilde{r}$ for action a and β is the exploration constant that balances the amount of exploration vs. exploitation across a time horizon T. A higher value of β makes the agent less optimistic and explore its action space more. The UCB agent's tendency to explore is also affected by the difference in the mean reward estimates of its actions. In the context of our principal - agent problem formulation, if the UCB agent has a larger value of $\delta = \max_a r_a - r_{a^*}$, without any intervention at the beginning of an episode, its confidence bounds would quickly converge to exploiting the action $\arg\max_a r_a$. So a principal that intervenes only towards the later stages of an episode with this agent would have to provide much more incentives (higher r') to alter the agent's preferred action to be a^*, thus incurring a larger cost c as compared to a principal that intervenes more at the beginning of an episode when the UCB agent is still exploring its action space. This is also illustrated in Section 5 with a simpler best response agent in the single round game setting. As shown in Figure 2b and Figure 2c, under observation noise (partial information), the meta-trained principal has a better one-shot meta-test-time performance when the agent's base payoff has a higher difference between the principal's preferred action and the agent's intrinsic preference without any intervention. Additionally, Appendix B.3 provides an illustration of this behavior.

ϵ-greedy. A simple exploration-exploitation strategy in the bandit setting is the ϵ-greedy rule (Sutton & Barto, 1998) wherein the agent selects with probability $1-\epsilon$ the action $a_t = \arg\max_a \tilde{\mu}_a$ and with probability ϵ it selects a random action. In our setting, we consider ϵ to be constant during an episode, which results in a uniform exploration rate throughout. In contrast to the UCB agent, the ϵ-greedy algorithm simulates a less optimistic, more exploratory agent for which the principal requires a robust belief representation of the agent's predicted behavior conditioned on the principal's past observations (Table 1). Since there is a uniform exploration rate for the agent, the principal has to continue intervening intermittently throughout an episode, especially when δ is large and the agent could obtain a higher reward for an action $a \neq a^*$ by exploring its action space when the principal does not intervene.

EXP3. The Exponential-weight algorithm for Exploration and Exploitation (EXP3) (Auer et al., 2002) follows a more pessimistic approach to exploration-exploitation in the bandit setting. It maintains a set of weights for each agent action $a \in A$ which are updated using the experienced rewards $\tilde{r}$ as follows:

$$\pi_t(a_t) = \frac{w}{|A|} + (1-w)\frac{\eta \exp\left(S_{a_t,t}\right)}{\sum_{a_t \in |A|} \eta \exp\left(S_{a_t,t}\right)}, \tag{6}$$

where

$$S_{a_t,t} = \sum_{l=1}^{t} \mathbf{1}\left\{a_l = a_t\right\} \frac{\tilde{r}_{a_t,l}}{\pi_l},\ \eta = \frac{w}{|A|}. \tag{7}$$

Here, w is the variable that determines the extent of uniform random exploration in the action space. This presents a very challenging problem to learn a suitable belief representation for such agents that can be utilized by a principal to guide its intervention policy. In Section 6, we exclude EXP3 from Table 1 since it is primarily designed for an adversarial bandit setup, whereas we do not consider an agent to have such biases under our current problem formulation.

B.2 Characterizing the distribution shift in our evaluation setup

Bandit agents having the same base reward r make different explore-exploit decisions depending on their algorithm (eg. UCB, ϵ-greedy) and also their prior observations. In Section 6, we consider agents with the same set of base rewards, but following different bandit algorithms. Both UCB and ϵ-greedy have tunable

Table 5: **Experiment design choice.** Frequency of agent selecting $a_t \neq \arg\max_a r_a$ with UCB and ϵ-greedy algorithms on the same set of base rewards (without any intervention) with a horizon $T = 200$, averaged across 3 random seeds.

β	UCB	ϵ-greedy	ϵ
0.17	33 (0)	33 (0)	0.10
0.27	47 (0)	47 (4)	0.20
0.42	70 (0)	68 (9)	0.30
0.50	80 (0)	81 (3)	0.40
0.67	99 (0)	99 (1)	0.50

parameters that determine their explore-exploit tradeoff. In order to measure the robustness of the learnt principal policy to different agent behavior (leading to different levels of non-stationarity in the principal's environment between training and test agents), we vary the amount of exploration performed by the agent by varying the respective parameters: β for the UCB agent and ϵ for the ϵ-greedy agent. Table 5 shows the average (and standard error) frequency of exploration by the agents for our choices of β and ϵ in Section 6. We vary β and ϵ such that they are pairwise comparable in Table 1 and would lead to similar change in exploration frequency for both UCB and ϵ-greedy agents. In other words, following Table 1, a principal trained with UCB agents having $\beta = 0.17$ when evaluated with UCB agents having $\beta \in \{0.17, 0.27, 0.42, 0.50, 0.67\}$ will encounter a similar shift in the agent's exploration frequency as in the case of training with ϵ-greedy agents with $\epsilon = 0.1$ and evaluating on ϵ-greedy agents having $\epsilon \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$. In that case, the difference in achieved scores between the UCB and ϵ-greedy agents can be attributed to the way in which they distribute their exploratory actions: UCB agent being more optimistic focuses most of its exploration at the beginning of an episode, whereas the ϵ-greedy agent is more stochastic with uniform random exploration throughout.

B.3 Visualizing the effect of β and the effect of different types of principal's interventions on the behavior of a UCB agent

We will consider three different instances of base rewards for a UCB agent and characterize its behavior when

- unintervened
- the principal intervenes once every 10 time steps for $T = 200$ (strategy S1)
- the principal intervenes continuously until the first 20 time steps for $T = 200$ (strategy S2)

for $\beta \in \{0.17, 0.27, 0.42, 0.50, 0.67\}$. Instead of a learned stochastic intervention policy, we will analyze the effect of the deterministic policies S1 and S2 in aligning the preferred action of an agent with the preferred action of the principal. Note that both S1 and S2 incur the same total intervention cost.

First, consider a UCB agent whose base reward is $\boldsymbol{r} = [0.16, 0.11, 0.66, 0.14, 0.20, 0.37, \mathbf{0.82}, 0.10, \mathbf{0.84}, 0.10]$ (Figure 4). The principal prefers the action with base reward 0.82, whereas the unintervened agent would prefer the action with base reward 0.84. For such small value of $\delta = \max_{a \in A} \boldsymbol{r}^i[a] - \boldsymbol{r}^i[a^*] = 0.02$, the agent can be incentivized to align its preferred action with a^* more easily than if δ were larger. Figure 4 depicts the behavior of the UCB agents with different exploration coefficients β under different principal-agent interaction conditions over $T = 200$ as follows. In Figure 4a, a value of 1 indicates the time step when the unintervened agent selects the action with base reward 0.84, whereas Figure 4b similarly shows its frequency of selecting a^* without the principal's intervention. As expected, these frequency distributions are quite similar since $\delta = 0.02$. Given the same budget for interventions, Figure 4c shows the agent's response to the principal's intervention strategy S1 and Figure 4d shows its response to the intervention strategy S2. In these figures, a value of 1 indicates the time step when the agent selects a^*. Both the principal intervention policies are able to align the action preference of the UCB agent with the principal's preference. But S2 receives a higher score, especially for agents with lower values for β. This is because the UCB algorithm gradually shifts from

exploration to exploitation as the episode progresses, and since S2 uses its intervention budget in the initial 20 time steps of the episode, the principal is able to incentivize the agent by effectively changing the observed reward $\tilde{r}$ at the beginning of the episode. It experimentally verifies our discussion in Section 6: in sequential learners, the time step *when* the principal intervenes determines how effective the intervention will be in helping align the agent's action preference with that of the principal. This becomes more pronounced for higher values of δ, which we analyze next.

Figure 5 demonstrates the behavior of a UCB agent with base reward $\boldsymbol{r} = [0.32, 0.67, 0.13, 0.72, 0.29, 0.18, \mathbf{0.59}, 0.02, \mathbf{0.83}, 0.01]$. The principal prefers the action with base reward 0.59. Without any interventions, the agent prefers the action with base reward 0.83. In this case, $\delta = \max_{a \in A} \boldsymbol{r}^i[a] - \boldsymbol{r}^i[a^*] = 0.24$. Figure 5b indicates the frequency with which the unintervened agent selects a^* whereas Figure 5a shows the frequency of selecting the action with base reward 0.83 without the principal's intervention. Note that the unintervened agent would rarely pick a^*, even less so for smaller values of the exploration coefficient β. Figure 5c indicates the frequency of the agent selecting a^* with principal's intervention policy S1 and Figure 5d indicates its frequency of selection a^* with principal's intervention policy S2. We observe that S2 outperforms S1 in aligning the agent's preferred action with the principal's preferred action. Since UCB agents tend to explore their action space more at the beginning of the episode, intervening on the agent's experienced reward during the initial time steps (S2) has a more noticeable effect in influencing the agent's preferred action than intervening with a fixed interval (S1).

Similar observations also hold in Figure 6 where the UCB agent has a base reward $\boldsymbol{r} = [0.79, 0.53, 0.57, \mathbf{0.93}, 0.07, 0.09, \mathbf{0.02}, 0.83, 0.78, 0.87]$ and the principal prefers the action with base reward 0.02. In this case, $\delta = \max_{a \in A} \boldsymbol{r}^i[a] - \boldsymbol{r}^i[a^*] = 0.91$. As Figure 6b shows, this implies that the unintervened agent would almost never select a^* even when it has a higher exploration coefficient β. Even with intervention policy S1, over $T = 200$, the principal wouldn't be able to align the agent's preferred action with its own as shown in Figure 6c. In contrast, Figure 6d shows that a principal with intervention policy S2 would outperform S1 and achieve a higher score, but the agent eventually discovers its own preference when the intervention stops in S2 and then it no longer selects a^*. This further highlights the extent of non-stationarity in the environment that affects the intervention policy of the principal and also the learning behavior of the agent. It also demonstrates the difficulty of the learning problem in our setup and the importance of learning a cost-efficient few-shot adaptable principal intervention policy to effectively intervene on unknown adaptive agents.

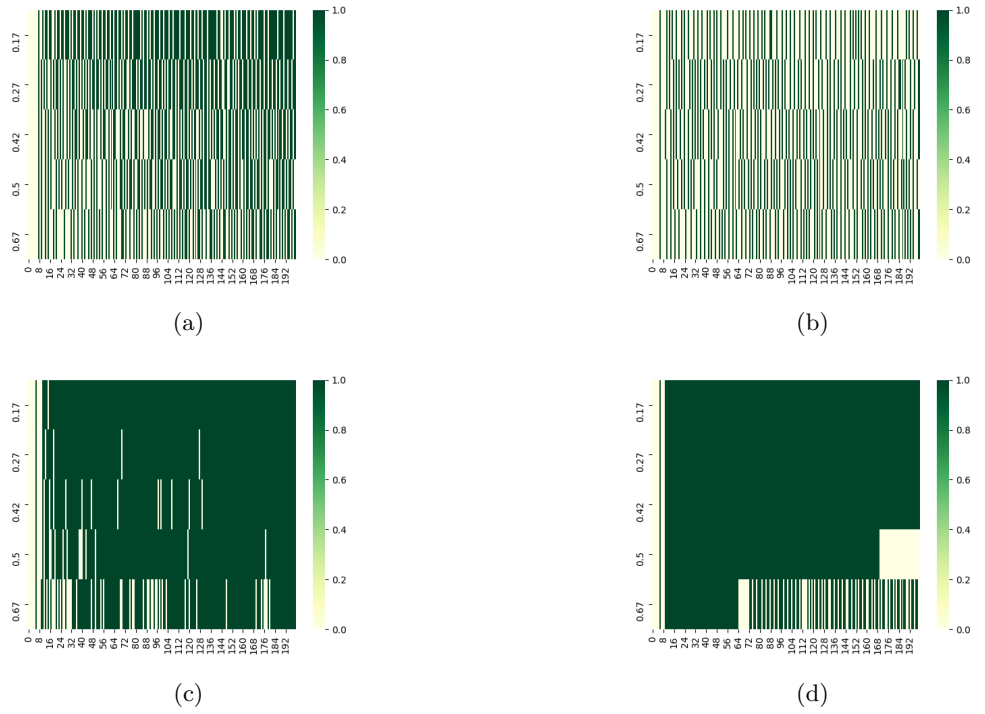
B.4 Description of baselines

We now describe the details of our evaluated baselines in Section 6 along with their variations that assume access to an agent state oracle.

Rule based intervention with an agent state oracle (RB): Given an oracle that correctly identifies the action a_t to be taken by an agent in the next time step, a simple rule based approach is for the principal to intervene at time t when $a_t \neq a^*$. We assume that the principal always intervenes with a fixed incentive ($r' = 0.5$ or 1) and we compute the principal's maximum possible score. Note that this is not a realistic solution for the principal since it is impractical to expect the availability of such an oracle, especially for out of distribution test agents.

Model-free learning based intervention policy: In this framework, we assume that the planner has a recurrent intervention policy that outputs a distribution over interventions $a^p_t \sim \pi^p_\theta\left(a^p_t | a^i_{t-1}, a^p_{t-1}, h^p_{t-1}\right)$, conditioned on the planner's intervention and observed agent action at $t-1$. The policy network is trained using REINFORCE for the MF-RL baseline and using MAML for the MF-MAML baseline.

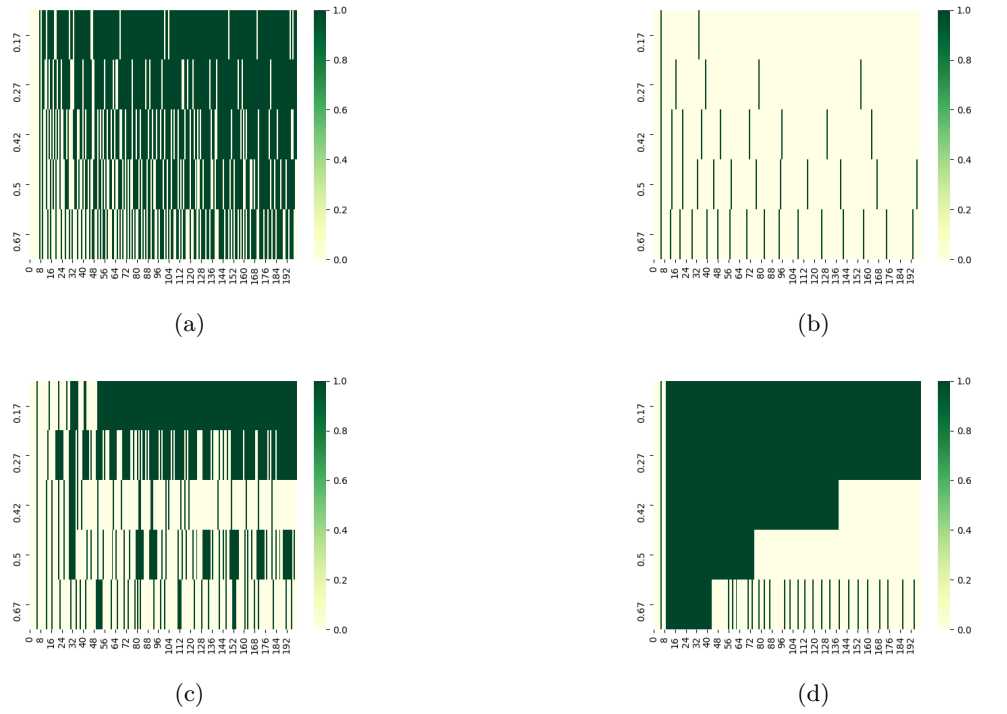
Learned intervention policy with an agent state oracle: In this setting, the principal learns a recurrent intervention policy that outputs a distribution over interventions $a^p_t \sim \pi^p_\theta\left(a^p_t | a^i_t, h^p_{t-1}\right)$ conditioned on the true agent action at time t provided by an oracle. The policy network is trained using REINFORCE for the SB-RL baseline and MAML for the SB-MAML baseline.



(a) (b)

(c) (d)

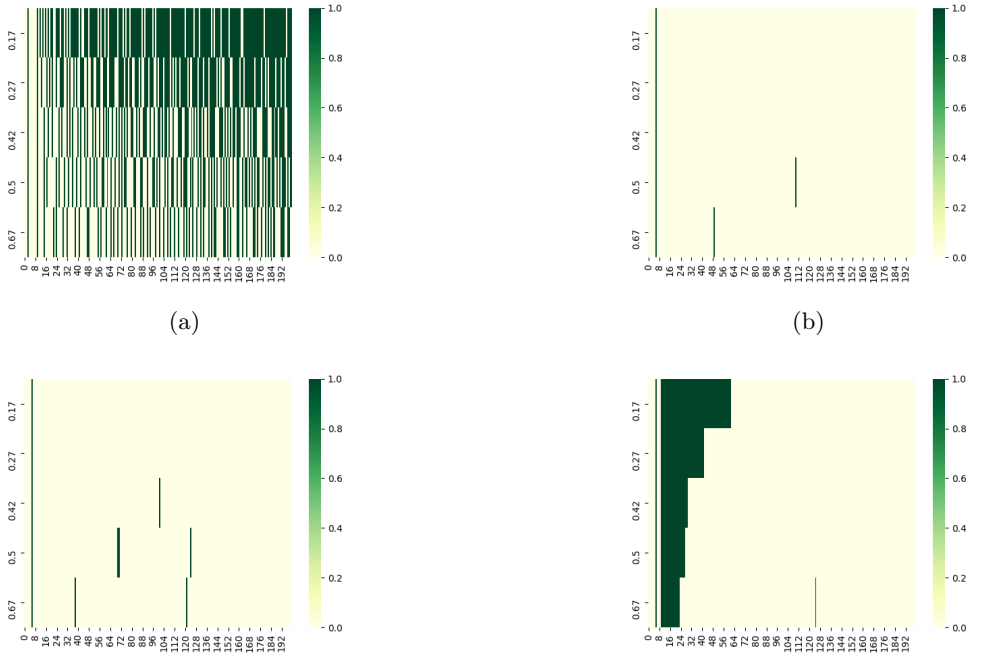
Figure 4: **Characterizing agent's behavior.** UCB agent with base rewards $[0.16, 0.11, 0.66, 0.14, 0.20, 0.37, \mathbf{0.82}, 0.10, \mathbf{0.84}, 0.10]$. The agent prefers the action with base reward 0.84, while the principal prefers the action with base reward 0.82. Horizontal axis indicates time steps $t = \{1, \ldots, 200\}$. Vertical axis indicates agents following UCB with different exploration coefficient β. Values are either 0 or 1. (a) Frequency distribution of agent selecting its unintervened preferred action with base reward 0.84. (b) Frequency distribution of agent selecting a^* without principal's intervention. (c) Frequency distribution of agent selecting a^* under principal's intervention S1. (d) Frequency distribution of agent selecting a^* under principal's intervention S2. For a small $\delta = \max_{a \in A} \boldsymbol{r}^i[a] - \boldsymbol{r}^i[a^*] = 0.02$, both S1 and S2 affect the agent's behavior quite similarly.



(a) (b)

(c) (d)

Figure 5: **Characterizing agent's behavior.** UCB agent with base rewards $[0.32, 0.67, 0.13, 0.72, 0.29, 0.18, \mathbf{0.59}, 0.02, \mathbf{0.83}, 0.01]$. The agent prefers the action with base reward 0.83, while the principal prefers the action with base reward 0.59. Horizontal axis indicates time steps $t = \{1, \ldots, 200\}$. Vertical axis indicates agents following UCB with different exploration coefficient β. Values are either 0 or 1. (a) Frequency distribution of agent selecting its unintervened preferred action with base reward 0.83. (b) Frequency distribution of agent selecting a^* without principal's intervention. (c) Frequency distribution of agent selecting a^* under principal's intervention S1. (d) Frequency distribution of agent selecting a^* under principal's intervention S2. For different values of β, the UCB agent acts differently based on *when* the principal intervened following S1 or S2. S1 intervenes periodically whereas S2 intervenes only at the beginning. The action selected by the agents in an episode clearly reflects the effect that this has in being able to align the agent's preference with that of the principal.



(a) (b)

(c) (d)

Figure 6: **Characterizing agent's behavior.** UCB agent with base rewards $[0.79, 0.53, 0.57, \mathbf{0.93}, 0.07, 0.09, \mathbf{0.02}, 0.83, 0.78, 0.87]$. The agent prefers the action with base reward 0.93, while the principal prefers the action with base reward 0.02. Horizontal axis indicates time steps $t = \{1, \ldots, 200\}$. Vertical axis indicates agents following UCB with different exploration coefficient β. Values are either 0 or 1. (a) Frequency distribution of agent selecting its unintervened preferred action with base reward 0.93. (b) Frequency distribution of agent selecting a^* without principal's intervention. (c) Frequency distribution of agent selecting a^* under principal's intervention S1. (d) Frequency distribution of agent selecting a^* under principal's intervention S2. Since $\delta = \max_{a \in A} \boldsymbol{r}^i[a] - \boldsymbol{r}^i[a^*] = 0.91$ is large and the UCB agent explores its action space in the initial time steps, S2 is able to intervene more effectively than S1 and achieves a higher score. Since the agent is a sequential learner, it discovers its own preferred action once the principal stops intervening and does not select a^*.

Table 6: **Principal (with oracle agent state input) scores across 3 random seeds.** These baselines are not applicable in practice since they cheat by assuming access to an oracle that always informs them of the agent's next action. We include them here as a form of standardization with respect to a (perfect) system that does not face the challenges of partial observability or out-of-distribution generalization in our setting.

Train on UCB, $\beta = 0.17$	Test on $\beta = 0.17$	$\beta = 0.27$	$\beta = 0.42$	$\beta = 0.5$	$\beta = 0.67$
No intervention	3 (0)	5 (0)	8 (0)	10 (0)	12 (0)
RB	173 (0)	166 (0)	154 (0)	146 (0)	126 (0)
SB-RL	168 (3)	138 (27)	128 (26)	122 (24)	107 (22)
SB-MAML	169 (3)	169 (1)	155 (2)	148 (1)	128 (2)
Train on ϵ**-greedy,** $\epsilon = 0.1$	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.3$	$\epsilon = 0.4$	$\epsilon = 0.5$
No intervention	3 (0)	4 (1)	7 (0)	9 (1)	11 (0)
RB	156 (3)	130 (1)	105 (4)	87 (4)	62 (6)
SB-RL	148 (2)	119 (3)	87 (4)	75 (6)	50 (2)
SB-MAML	152 (1)	126 (2)	105 (3)	66 (3)	30 (9)
Train on UCB, $\beta = 0.67$	$\beta = 0.17$	$\beta = 0.27$	$\beta = 0.42$	$\beta = 0.5$	$\beta = 0.67$
No intervention	3 (0)	5 (0)	8 (0)	10 (0)	12 (0)
RB	173 (0)	166 (0)	154 (0)	146 (0)	126 (0)
SB-RL	166 (3)	163 (2)	150 (3)	146 (2)	128 (2)
SB-MAML	173 (1)	170 (0)	159 (0)	152 (0)	133 (0)
Train on ϵ**-greedy,** $\epsilon = 0.5$	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.3$	$\epsilon = 0.4$	$\epsilon = 0.5$
No intervention	3 (0)	4 (1)	7 (0)	9 (1)	11 (0)
RB	156 (3)	130 (1)	105 (4)	87 (4)	62 (6)
SB-RL	49 (46)	51 (35)	64 (29)	61 (15)	28 (17)
SB-MAML	93 (45)	62 (32)	32 (13)	58 (25)	24 (17)

Learned intervention policy with a world model without meta-learning (WM-RL): In this setting, we use our proposed recurrent world model with a recurrent intervention policy trained using REINFORCE. Here, the policy network outputs a distribution over interventions $a_t^p \sim \pi_\theta^p \left(a_t^p | \hat{a}_t^i, h_{t-1}^p\right)$ where $\hat{a}_t^i = \arg\max_a \hat{\pi}_\omega \left(a_t^i | a_{t-1}^i, a_{t-1}^p, h_{t-1}^i\right)$.

We would like to highlight an implementation detail in our baselines indicated 'RL' in Section 6. Since we evaluate our learnt principal policy in the K-shot adaptation setting which is common in the meta-learning literature, we ensure that the principal policies that are not meta-trained are also allowed to K-shot adapt at test time. This means that the 'RL' policies are also updated at test time, before evaluation, using K rounds of principal-agent interactions. This is in contrast to Section 5 where 'RL' was trained from scratch during test time adaptation. It further shows that even with pre-training (on the same set of train agents as used by 'MAML'), standard policy gradient update does not lead to effective test time K-shot adaptation on test agents.

In Table 6, we compare the test time scores for the principal policy having access to a state based oracle. We observe that overall, the meta-trained principal policy (SB-MAML) achieves a higher score even with distribution shift across different bandit algorithms and different levels of exploration, compared to the SB-RL baseline. The rule based baseline also shows strong performance but we note its scores do not reflect adaptation to distribution shift. However, none of these baselines that assume the principal has access to an oracle that correctly predicts the agent's action at the next time step are realistic. We can only treat the scores in Table 6 as gold standards in a perfect system that does not account for the challenges faced by a principal in practice.

Training details. In Section 5, the principal policy π^p is a fully connected neural network (MLP) with one hidden layer and ReLU activation. Given an (noisy) observed value of the agent type as input, it predicts the probability of intervention: π_t^p. The principal's action at time t is $a_t^p \sim \text{Bern}\left(\pi_t^p\right)$.

Algorithm 2 MERMAIDE (K-shot Adaptation)

Initialize principal with trained parameters $(\theta_{\text{meta}}, \omega_{\text{train}})$, and hidden states h_0^i, h_0^p.
for agents (tasks) $i = 1, \ldots, n_{\text{test}}$ **do**
  Initialize agent: (μ^i, π_0^i), task specific intervention policy parameter $\theta\left(\tau_0^i\right) = \theta_{\text{meta}}$.
  for $k = 1, \ldots, K$ **do**
    for time t $= 1, \ldots, T$ **do**
      Predict $\hat{a}_t^i = \arg\max_{a_t^i} \hat{\pi}_{\omega_{\text{train}}}\left(a_t^i | a_{t-1}^i, a_{t-1}^p, h_{t-1}^i\right)$ using the world model.
      Intervention: $\tilde{\mu}^i = \mu^i + a_t^p, \quad a_t^p \sim \pi^p_{\theta\left(\tau_k^i\right)}\left(a_t^p | a_{t-1}^i, a_{t-1}^p, \hat{a}_t^i, h_{t-1}^p\right)$.
      Agent acts: $a_t^i \sim \pi_t^i$ and receives reward $r_t^i \sim \mathcal{N}\left(\tilde{\mu}^i, \sigma^2\right)$. $\pi_t^i \mapsto \pi_{t+1}^i$.
    end for
    Locally update $\theta\left(\tau_k^i\right) \mapsto \theta\left(\tau_{k+1}^i\right)$. {Using REINFORCE.}
  end for
  for $t = 1, \ldots, T$ **do**
    Predict $\hat{a}_t^i = \arg\max_{a_t^i} \hat{\pi}_{\omega_{\text{train}}}\left(a_t^i | a_{t-1}^i, a_{t-1}^p, h_{t-1}^i\right)$ using the world model.
    Intervention: $\tilde{\mu}^i = \mu^i + a_t^p, \quad a_t^p \sim \pi^p_{\theta\left(\tau_K^i\right)}\left(a_t^p | a_{t-1}^i, a_{t-1}^p, \hat{a}_t^i, h_{t-1}^p\right)$.
    Agent acts: $a_t^i \sim \pi_t^i$, receives reward $r_t^i \sim \mathcal{N}\left(\tilde{\mu}^i, \sigma^2\right)$. Updates $\pi_t^i \mapsto \pi_{t+1}^i$.
    Update principal's score.
  end for
end for

For the 'RL' principal, it is trained on the test agents starting from scratch over K episodes before evaluation. For the MAML principal, it is meta-trained to learn an initial parameterization with a different set of training agents and evaluated with K-shot adaptation on the test agents.

In Section 6, the recurrent world model and policy networks are GRUs with 2 layers and hidden state dimension 128. For meta-training, the inner gradient update loop uses SGD optimizer with a learning rate of 7×10^{-4} whereas the meta-update step uses Adam with a learning rate of 0.001. The world model is trained only with the set of training agents, it is not adapted at test time: only the policy network is K-shot adapted.

We plan to release the code for our implementation with the published paper.

B.5 Overview of K-shot adaptation with MERMAIDE:

Algorithm 2 outlines our framework for K-shot adaptation of the meta-trained principal to test agents. In our experiments, $K = 1$.

C Additional experimental results with bandit agents

Comparing the stability of MERMAIDE for different sets of random seeds. We trained and evaluated MERMAIDE for 3 additional random seeds in the $K = 1$-shot setting from Table 1. In Table 7, we show the scores (mean and standard error) for both sets of seeds and all the six seeds combined. Our results indicate that the training and evaluation of MERMAIDE is stable, with reasonable and explainable variability across random seeds. More specifically, models trained with different sets of random seeds result in similar mean scores with small standard error when $K = 1$-shot evaluated on the less stochastic UCB agents. In contrast, evaluation with the more stochastic ϵ-greedy agents results in comparably larger standard error and more variation in the mean scores for models trained with different sets of random seeds. Due to the computational costs involved, we were unable to evaluate all baselines with six seeds but on the basis of these observations, we do not expect a significant deviation from the claims made in Section 6 and the values originally reported in Table 1 and Table 2, even with more seeds.

Table 7: **Comparing the variability of MERMAIDE test scores with a trained model across different sets of random seeds.** Set 1 uses the seeds {11, 26, 90} and Set 2 uses the seeds {12, 27, 91}. 'Combined' indicates mean and s.e. scores for models trained with the seeds {11, 12, 26, 27, 90, 91}. Overall, we observe that training MERMAIDE with different sets of random seeds shows little variance in the evaluation scores within error bounds, especially with the less stochastic UCB agent. For the ϵ-greedy agent, we observe a higher standard error in the $K = 1$-shot evaluation scores, but it is in line with the stochasticity associated with ϵ-greedy action selection in the bandit agent.

Train on UCB, $\beta = 0.17$	Test on $\beta = 0.17$	$\beta = 0.27$	$\beta = 0.42$	$\beta = 0.5$	$\beta = 0.67$
MERMAIDE (Set 1; 3 seeds)	154 (2)	151 (1)	129 (1)	129 (1)	87 (0)
MERMAIDE (Set 2; 3 seeds)	144 (3)	133 (3)	122 (2)	108 (2)	81 (2)
MERMAIDE (Combined; 6 seeds)	151 (3)	142 (4)	125 (2)	118 (5)	84 (2)
Train on UCB, $\beta = 0.67$	$\beta = 0.17$	$\beta = 0.27$	$\beta = 0.42$	$\beta = 0.5$	$\beta = 0.67$
MERMAIDE (Set 1; 3 seeds)	132 (1)	130 (1)	123 (2)	115 (2)	99 (1)
MERMAIDE (Set 2; 3 seeds)	119 (7)	118 (5)	110 (3)	104 (3)	87 (2)
MERMAIDE (Combined; 6 seeds)	126 (4)	124 (3)	116 (3)	110 (3)	93 (3)
Train on ϵ**-greedy,** $\epsilon = 0.1$	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.3$	$\epsilon = 0.4$	$\epsilon = 0.5$
MERMAIDE (Set 1; 3 seeds)	138 (1)	112 (2)	85 (3)	66 (2)	37 (4)
MERMAIDE (Set 2; 3 seeds)	133 (2)	133 (4)	133 (4)	133 (4)	133 (4)
MERMAIDE (Combined; 6 seeds)	136 (2)	123 (5)	109 (11)	99 (15)	85 (22)