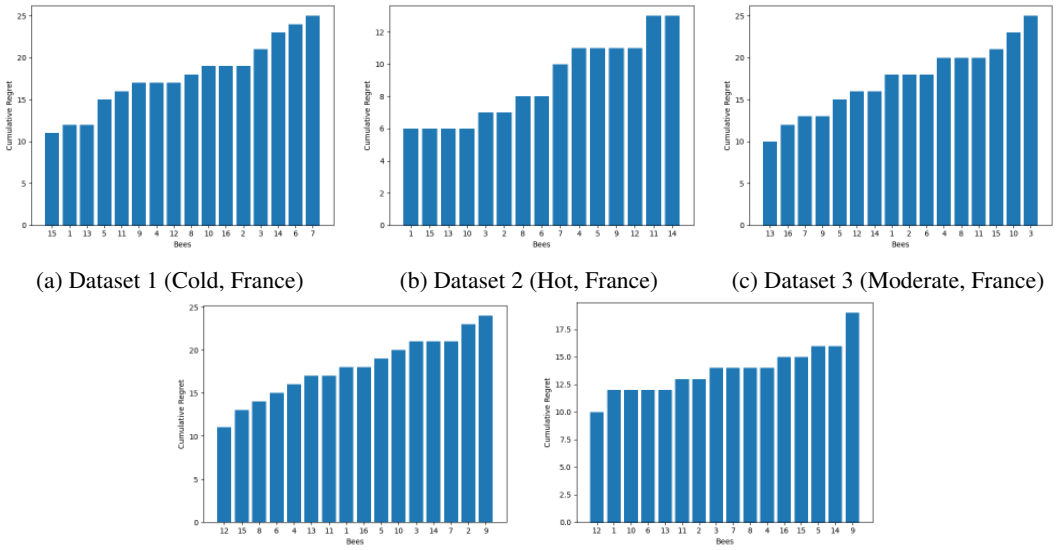
6 APPENDIX

6.1 DATASET DESCRIPTION

In this dataset, bees are confronted to a numerical discrimination task. Bees first enter the maze in an entrance chamber before flying in a hole and facing two images located at the end of each arm. The image has different number of dots : for example in dataset 1 and 2, one of the image has two dots while the other have four dots. If the bee chooses the correct image (i.e. the side with the highest number of dots), it will be rewarded by a sugar reward (50% sugar/water) placed in pipette in the middle of the image, alternatively if it chooses the incorrect image then it will be punished finding a bitter tasting solution (quinine solution) within the pipette. Bees cannot detect (neither visually nor by odor) which solution is located where. Then, they are only able to know image on each side before to choose. Between each trials the bee will go back to the hive to deliver the collected sugar, before reaching back the maze for another trial (typically lasting a few minutes). During this time the experimenter randomly changes the images or not, and varying the position of the dots. The localization of the correct image alternate between the right and left arm according to a pseudo-random sequence. Each dataset include 16 bees.

Table 3: Datasets summary

Dataset	nb indiv	T	Location	Weather
Dataset 1	16	40	France	Cold
Dataset 2	16	22	France	Hot
Dataset 3	16	40	France	Moderate
Dataset 4	16	40	Australia	Cold
Dataset 5	16	30	Australia	Hot



(a) Dataset 1 (Cold, France) (b) Dataset 2 (Hot, France) (c) Dataset 3 (Moderate, France)

(d) Dataset 4 (Cold, Australia) (e) Dataset 5 (Hot, Australia)

Figure 9: Proportion of cumulative regret for the five datasets, per bees

6.2 MSE AND MAE OF MAYA ACCORDING τ

side_window	MAYA_KL	MAYA_KL	MAYA_Wass	MAYA_Wass	MAYA_DTW	MAYA_DTW
	mean MSE	mean MAE	mean MSE	mean MAE	mean MSE	mean MAE
3.0	2.7 ± 2	1.2 ± 0.7	2.7±2	1.2 ± 0.6	7.4 ± 10	1.9 ± 1
4.0	4.2 ± 4	1.5 ± 0.8	3.0 ± 2	1.3 ± 0.6	8.0 ± 13	1.9 ± 1
5.0	4.0 ± 4	1.4 ± 0.9	3.8 ± 3	1.5 ± 0.6	6.8 ± 7	1.9 ± 0.9
6.0	4.1 ± 2	1.6 ± 0.6	2.8 ± 2	1.2 ± 0.5	7.5 ± 7	2.0 ± 1
7.0	4.2 ± 3	1.5 ± 0.7	2.5 ± 1	1.2 ± 0.5	6.7 ± 7	1.9 ± 1
8.0	5.5 ± 5	1.7 ± 0.9	3.7 ± 3	1.4 ± 0.7	7.2 ± 7.8	2.0 ± 1
9.0	3.9 ± 3	1.4 ± 0.7	2.9 ± 2	1.2 ± 0.6	8.8± 9	2.2 ± 1
10.0	5.5 ± 5	1.7 ± 0.9	4.1 ± 4	1.5 ± 0.8	8.7 ± 10	2.0 ± 1
20.0	5.4 ± 5	1.6 ± 0.8	4.8 ± 5	1.5 ± 0.9	8.7 ± 10	2.1 ± 1
30.0	4.3 ± 3	1.5± 0.6	4.4 ± 3	1.5 ± 0.7	8.4 ± 10	2.0 ± 1
$T = 40$	5.1 ± 5	1.6 ± 1	4.8 ±6	1.5 ± 0.9	9.7 ± 11	2.2±1

Table 4: Dataset 1 (Cold weather, France)

side_window	MAYA_KL	MAYA_KL	MAYA_Wass	MAYA_Wass	MAYA_DTW	MAYA_DTW
	mean	mean	mean	mean	mean	mean
3.0	1.2± 1	0.7± 0.4	1.3± 1	0.8± 0.4	2.4± 1	1.1± 0.4
4.0	1.3± 0.8	0.8± 0.3	1.5± 1	0.8 ± 0.4	2.7± 2	1.1± 0.6
5.0	2.1± 1	1.0± 0.4	1.9± 2	1.0± 0.5	2.4± 2.7	1.0± 0.6
6.0	2.1± 1	1.0± 0.5	1.5± 1	0.8 ± 0.4	3.5 ± 3	1.3± 0.7
7.0	1.6± 1	0.9± 0.4	1.5± 1	0.8± 0.4	3.0± 3	1.2± 0.6
8.0	1.9± 1	1.0± 0.3	1.8± 1	0.9± 0.4	2.8± 2	1.2± 0.6
9.0	1.8± 1	0.9± 0.4	2.2± 2	1.0± 0.6	2.5± 2	1.1± 0.6
10.0	2.3 ± 2	1.0± 0.5	2.1± 2	1.0± 0.6	2.7± 1	1.2± 0.6
20.0	2.3± 1	1.0± 0.4	2.6± 1	1.1± 0.3	2.0± 1	1.0± 0.4
$T = 22$	3.2± 3	1.2± 0.6	2.8± 1	1.2± 0.4	2.1± 1	1.0± 0.5

Table 5: Dataset 2 (Hot weather, France)

side_window	MAYA_KL	MAYA_KL	MAYA_Wass	MAYA_Wass	MAYA_DTW	MAYA_DTW
	mean MSE	mean MAE	mean MSE	mean MAE	mean MSE	mean MAE
3.0	3.0±2	1.3±0.6	4.0±4	1.4 ±0.8	8.4±12	2.0±1.4
4.0	4.4±4	1.5±0.9	3.9±4	1.4±0.8	7.4±11	1.9±1
5.0	4.3±4	1.5±0.7	3.0±3	1.2±0.7	7.3±11	1.9±1
6.0	4.4±4	1.5±0.8	3.1±2	1.3±0.6	7.7±9	2.0±1
7.0	3.7±3	1.4±0.6	2.6±1	1.2±0.5	5.7±5	1.8±0.8
8.0	4.1±3	1.5±0.7	2.5±1	1.1±0.4	8.3±9	2.1±1
9.0	5.8±5	1.8±0.8	4.2±2	1.6±0.6	8.1±8	2.1±1
10.0	3.6±3	1.4±0.7	4.9±5	1.6±1	7.1±9	1.9±1
20.0	5.3±4	1.7±0.8	5.2±5	1.7±0.7	6.5±8	1.9±1
30.0	3.6±2	1.4±0.5	4.4±3	1.6±0.7	8.7±9	2.2±1
$T = 40$	4.2 ± 4	1.5 ± 0.8	3.45±3	1.3±0.6	9.3 ± 11	2.2±1

Table 6: Dataset 3 (Moderate weather, France)

side_window	MAYA_KL	MAYA_KL	MAYA_Wass	MAYA_Wass	MAYA_DTW	MAYA_DTW
	mean MSE	mean MAE	mean MSE	mean MAE	mean MSE	mean MAE
3.0	4.5 ± 4	1.6 ± 0.9	3.0 ± 4	1.2 ±0.9	7.1 ± 10	1.8 ± 1.3
4.0	3.8 ± 3	1.5±0.7	3.6 ±3	1.5 ± 0.6	7.1 ± 9	1.9 ± 1
5.0	4.9 ± 3	1.7 ±0.7	2.6 ± 3	1.2 ± 0.7	7.6±11	1.9 ±1
6.0	4.1±3	1.5 ± 0.7	2.6 ± 1	1.2 ± 0.4	7.8±9	2.0 ± 1
7.0	3.7±3	1.4±0.6	3.6±2	1.5 ±0.5	8.3 ±10	2.1 ±1
8.0	6.2±8	1.7 ±1	3.4±2	3.4±2	6.2±7	1.8 ± 1
9.0	4.6 ±3	1.6 ±0.7	3.1 ±2	1.3±0.5	8.1±7	2.1 ±1
10.0	7.7±7	2.0 ±1	4.8±4	1.6±0.8	8.4±10	2.0±1
20.0	5.4 ±4	1.7 ±0.8	4.4 ± 2	1.6 ±0.5	8.5 ± 11	2.1 ±1.2
30.0	5.5 ±4	1.7±0.7	6.7 ±7	1.9 ±0.9	9.0 ±12	2.1 ± 1
$T = 40$	4.2 ± 5	1.4 ± 0.8	3.3 ± 2	1.3 ± 0.6	9.0 ± 10	2.2 ± 1

Table 7: Dataset 4 (Cold weather, Australia)

side_window	MAYA_KL	MAYA_KL	MAYA_Wass	MAYA_Wass	MAYA_DTW	MAYA_DTW
	mean MSE	mean MAE	mean MSE	mean MAE	mean MSE	mean MAE
3	6.6 ± 9	1.6 ± 1	6.3 ± 9	1.6 ± 1	8.6 ± 10	1.9 ± 1
4	8.1 ± 8	2.0 ± 1	10.4 ± 12	2.2 ± 1	9.4 ± 8	2.1 ± 1
5	4.3 ± 5	1.4 ± 0.9	8.4 ± 10	2.0 ± 1	10.4 ± 12	2.2 ± 1
6	3.6 ± 3	1.4 ± 0.7	3.9 ± 8	1.2 ± 1	12.0 ± 11	2.3 ± 1
7	3.4± 3	1.2 ± 0.9	4.5 ± 5	1.5 ± 1	10.3 ± 11	2.1 ± 1
8	4.1 ±3	1.5±0.6	4.4±5	1.5±0.9	10.3 ± 12	2.2±1
9	5.5±8	1.6±1	5.7±6	1.7±1	12.9 ± 16	2.4 ± 1
10	**3.3** ± 3	**1.3** ± 0.6	**3.3** ± 3	**1.3** ± 0.7	9.6 ± 10	2.1 ± 1
20	6.4±6	1.8 ±1	4.7 ± 5	1.5 ±0.8	11.8±13	2.3 ± 1
$T = 30$	6.1±5	1.8±0.9	6.1±5	1.8±0.9	9.2±10	2.1±1

Table 8: Dataset 5 (Hot weather, Australia)

Table 9: MSE and MAE of MAYA as a function of the window size τ. The T row denotes the no-window setting ($\tau = T$), where at each trial the full trajectory up to time t is used.

7 UNDERSTANDING THE LEARNING PROCESS

7.1 MAYA EXPLAINABILITY WITH $\tau = 7$

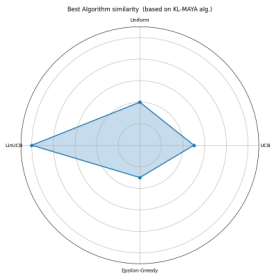


Figure 10: MAYA-KL

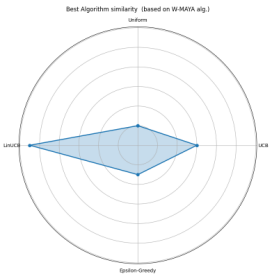


Figure 11: MAYA-Wass

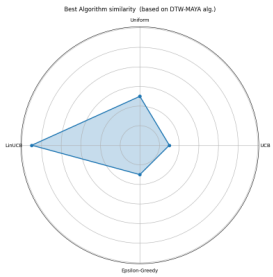


Figure 12: MAYA-DTW

Figure 13: For bee 1 (fast learner, low regret) from dataset 2 we report choice interpretability for MAYA-variants ($\tau = 7$).

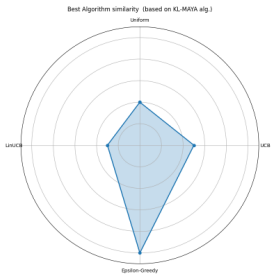


Figure 14: MAYA-KL

Figure 15: MAYA-Wass

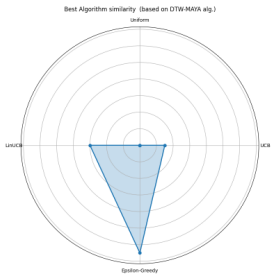


Figure 16: MAYA-DTW

Figure 17: For bee 15 (slow learner, high regret) from dataset 2 we report choice interpretability for MAYA-variants ($\tau = 7$).

7.2 MAYA EXPLAINABILITY WITH $\tau = 3$

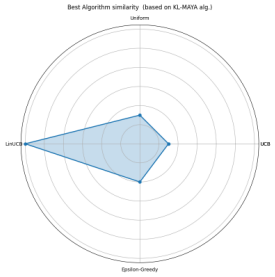


Figure 18: MAYA-KL

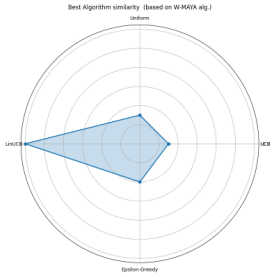


Figure 19: MAYA-Wass

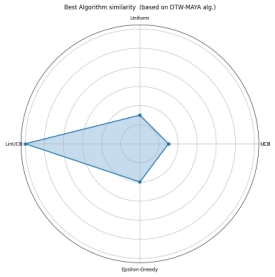


Figure 20: MAYA-DTW

Figure 21: For bee 1 (fast learner, low regret) from dataset 2 we report choice interpretability for MAYA-variants ($\tau = 3$).

Figure 22: MAYA-KL

Figure 23: MAYA-Wass

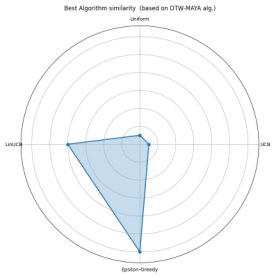


Figure 24: MAYA-DTW

Figure 25: For bee 15 (slow learner, high regret) from Dataset 2 we report choice interpretability for MAYA-variants ($\tau = 3$).

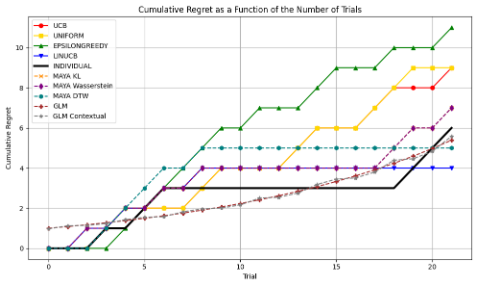


Figure 26: Bee 1

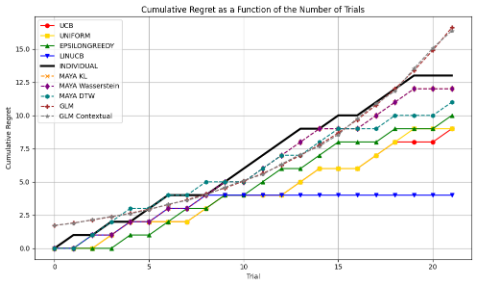


Figure 27: Bee 15

Figure 28: Regret modelization for bee 1 (lower cumulative regret) and bee 15 (higher cumulative regret) of Dataset 2, with $\tau = 3$

8 COMPARATIVE METHODES DESCRIPTION

- Generative Adversarial Imitation Learning (GAIL) GAIL learns a policy by simultaneously training it with a discriminator that aims to distinguish expert trajectories against trajectories from the learned policy. Ho & Ermon (2016)
- Behavioral Cloning (BC) Behavioral cloning directly learns a policy by using supervised learning on observation-action pairs from expert demonstrations. It is a simple approach to learning a policy, but the policy often generalizes poorly and does not recover well from errors. Foster et al. (2024).
- AIRL, similar to GAIL, adversarially trains a policy against a discriminator that aims to distinguish the expert demonstrations from the learned policy. Unlike GAIL, AIRL recovers a reward function that is more generalizable to changes in environment dynamics. Fu et al. (2018).
- DAgger (Dataset Aggregation) iteratively trains a policy using supervised learning on a dataset of observation-action pairs from expert demonstrations (like behavioral cloning), runs the policy to gather observations, queries the expert for good actions on those observations, and adds the newly labeled observations to the dataset. DAgger improves on behavioral cloning by training on a dataset that better resembles the observations the trained policy is likely to encounter, but it requires querying the expert online Ross et al. (2011).
- Density-based reward modeling is an inverse reinforcement learning (IRL) technique that assigns higher rewards to states or state-action pairs that occur more frequently in an expert's demonstrations. The key intuition behind this method is to incentivize the agent to take actions that resemble the expert's actions in similar states Dumoulin et al. (2024).
- Maximum Causal Entropy Inverse Reinforcement Learning (MCE IRL) : The principle of maximum causal entropy is a method that extends the classical maximum entropy idea to sequential settings. Instead of considering probabilities in isolation, it uses causally conditioned probabilities, which means that the model explicitly accounts for the fact that information is revealed step by step over time. This allows us to properly capture how side information becomes available and how it influences decisions at each stage Biernaskie et al. (2009).
- Preference Comparisons : The preference comparison algorithm learns a reward function from preferences between pairs of trajectories. The comparisons are modeled as being generated from a Bradley-Terry (or Boltzmann rational) model, where the probability of preferring trajectory A over B is proportional to the exponential of the difference between the return of trajectory A minus B. In other words, the difference in returns forms a logit for a binary classification problem, and accordingly the reward function is trained using a cross-entropy loss to predict the preference comparison. Christiano et al. (2023).
- Soft Q Imitation Learning (SQIL) : Soft Q Imitation learning learns to imitate a policy from demonstrations by using the DQN algorithm with modified rewards. During each policy update, half of the batch is sampled from the demonstrations and half is sampled from the environment. Expert demonstrations are assigned a reward of 1, and the environment is assigned a reward of 0. This encourages the policy to imitate the demonstrations, and to simultaneously avoid states not seen in the demonstrations Reddy et al. (2020).
- GLM : A Generalized Linear Model (GLM) is a statistical framework that extends linear regression to response variables with non-Gaussian distributions. In our setting, the regret trajectory $R(\pi, 1, T)$ is modeled as a function of time, $R(\pi, 1, T) \sim f(t)$, where f is linked to a linear predictor through a canonical link function. A Poisson GLM is employed when the noise structure is count-like, while a Gamma GLM is used to capture multiplicative noise. This allows us to statistically frame the evolution of regret as a stochastic process, while accounting for heterogeneous variability across agents. Nelder & Wedderburn (1972).
- Contextual GLM : The contextual variant incorporates side information (e.g., environmental or experimental conditions) into the predictor, enabling the model to capture how context modulates regret dynamics. Then $R(\pi, 1, T) \sim f(t, x_t)$ McCullagh & Nelder (1989).

8.1 MAE COMPARISON OF METHODS

Table 10: MAE comparison of methods across the five datasets. Values are reported as mean ± standard deviation. We fix $\tau = 7$ for all MAYA variant

Dataset	GAIL	BC	AIRL	Dagger	DBR	MCE	Pref-Comp	SQIL	GLM (no ctx)	GLM (ctx)	MAYA-KL	MAYA-Wass	MAYA-DTW
1	3.75± 2.5	**1.61 ± 0.79**	**0 ± 0**	2.9 ± 2.8	4.3± 3.8	10.38 ± 1.60	8.35± 3.25	3.71±1	**1.4 ± 0.3**	**1.4 ± 0.3**	1.5 ± 0.7	**1.2 ± 0.5**	1.9 ± 1
2	3.69±1.8	**1.24± 0.72**	**0 ± 0**	1.93 ± 1.7	2.72 ± 1.89	6.04 ± 1.0	3.7 ± 1.9	2.18±0.9	**0.8 ± 0.5**	**0.8 ± 0.5**	1.4 ±0.6	1.5 ±0.5	2.1 ± 1
3	3.62±2.4	**1.79 ± 0.98**	**0 ± 0**	2.6 ± 3.1	3.4 ± 4.1	8.13 ± 1.10	9.76 ± 1.75	3.2±1	**1.4 ± 0.4**	**1.4 ± 0.4**	3.7 ± 3	**2.6 ± 1**	1.8 ± 0.8
4	3.1±2.8	**1.65 ± 0.86**	**0 ± 0**	3.0 ± 2.7	4.60 ± 4.8	10± 1.6	9.7 ± 1.7	3.2±1	**2.1 ± 1**	**2.1 ± 1**	1.4± 0.6	1.5 ± 0.5	2.1 ± 1
5	4.9 ± 2.8	**3.23 ± 3**	**0 ± 0**	6.5± 5.1	5.5 ± 7.8	15.0 ± 7.6	14.3 ± 6.92	4.52±2	8.0 ± 8	**2.2 ± 1**	**1.2 ± 0.9**	1.3 ± 0.7	2.1 ± 1

9 FINETUNING IMITATION LEARNING

We present ablations over the fine-tuning budget of the IRL methods. As the tuning knobs differ across methods, we use the unified notation b for the method-specific budget (see Tab 11). The best results are summarized in the main text.

$b^{(\text{GAIL})}$	$b^{(\text{BC})}$	$b^{(\text{Dagger})}$	$b^{(\text{DBR})}$	$b^{(\text{MCE})}$	$b^{(\text{PrefComp})}$	$b^{(\text{PrefComp})}$
epochs	*epochs*	*env. steps*	*epochs*	*epochs*	*# envs*	*eval episodes*

Table 11: Hyperparameters of each comparative methods.

	MSE (b=1)	MAE (b=1)	MSE (b=10)	MAE (b=10)	MSE (b=50)	MAE (b=50)
GAIL	29.6 +/- 41	3.75+/-2.5	29.6 +/- 41	3.75+/-2.5	29.6 +/- 41	3.75+/-2.5
BC	23.2 +/- 30.8	3.26 +/- 2.74	19.8 +/- 26.5	3.1+/-2.3	5.16+/-3.94	1.61+/-0.79
AIRL	0 +/- 0	0 +/- 0	0 +/- 0	0 +/- 0	0 +/- 0	0 +/- 0
Dagger	22.8+/- 32.9	2.9+/-2.8	36.9+/-52.0	3.7 +/- 3.8	32.5 +/- 50.6	3.7+/- 3.3
Density based reward	43.1 +/- 54.81	4.3+/-3.8	43.1 +/- 54.8	4.3+/-3.8	43.1 +/- 54.8	4.3+/-3.8
MCE	148.83 +/- 38.47	10.38 +/- 1.60	148.83 +/- 38.47	10.38 +/- 1.60	148.83 +/- 38.47	10.38 +/- 1.60
Pref-Comp	120.25 +/- 52.1	9.17 +/- 2.99	114 +/- 53	8.9 +/- 2.9	104.5 +/- 57	8.35 +/- 3.25
SQIL	26.2 +/-19	3.75 +/- 1	26.2 +/-19	3.75 +/- 1	26.2 +/-19	3.75 +/- 1

Table 12: Dataset 1 (Cold weather, France)

	MSE (b=1)	MAE (b=1)	MSE (b=10)	MAE (b=10)	MSE (b=50)	MAE (b=50)
GAIL	23.2 +/- 17	3.69 +/- 1.8	23.2 +/- 17	3.69 +/- 1.8	23.2 +/- 17	3.69 +/- 1.8
BC	12.1+/-12.1	2.54+/-1.74	7.3+/-7.7	1.99+/-1.3	2.86 +/- 2.95	1.24 +/- 0.72
AIRL	0	0 +/- 0	0 +/- 0	0 +/- 0	0 +/- 0	0 +/- 0
Dagger	15.63 +/- 19.2	2.54+/-2.2	11.8 +/- 16.5	2.1+/-2.0	9.67+/- 12.6	1.93 +/-1.7
Density based reward	15.26 +/- 16.43	2.72 +/- 1.89	15.26 +/- 16.43	2.72 +/- 1.89	15.26 +/- 16.43	2.72 +/- 1.89
MCE	49.5 +/- 14.2	6.04 +/- 1.0	49.5 +/- 14.2	6.04 +/- 1.0	49.5 +/- 14.2	6.04 +/- 1.0
Pref-Comp	24.54+/-18.3	3.7 +/-1.9	30.15 +/-17.3	4.49 +/- 1.53	28.84 +/- 16.13	4.46 +/- 1.30
SQIL	9.80 +/-6	2.18+/-0.9	9.80 +/-6	2.18+/-0.9	9.80 +/-6	2.18+/-0.9

Table 13: Dataset 2 (Hot weather, France)

	MSE (b=1)	MAE (b=1)	MSE (b=10)	MAE (b=10)	MSE (b=50)	MAE (b=50)
GAIL	27.5 +/- 40	3.62 +/-2.5	27.5 +/- 40	3.62 +/-2.5	27.5 +/- 40	3.62 +/-2.5
BC	15.9+/-24	2.67 +/- 2.26	22.0+/-25	3.55+/-2.1	5.5+/-4.1	1.79+/-0.98
AIRL	0 +/- 0	0 +/- 0	0 +/- 0	0 +/- 0	0 +/- 0	0 +/- 0
Dagger	35.4+/- 61.8	3.3 +/-3.7	34.5+/-48.2	3.5 +/- 3.4	21.6 +/-46.0	2.6 +/-3.1
Density based reward	41.38 +/- 51.1	3.4+/-4.1	41.38 +/- 51.1	3.4+/-4.1	41.38 +/- 51.1	3.4+/-4.1
MCE	140.3 +/-34.7	8.13 +/-1.10	140.3 +/-34.7	8.13 +/-1.10	140.3 +/-34.7	8.13 +/-1.10
Pref-Comp	130.98 +/-44.7	9.98 +/-1.98	134.12+/-37	10.12 +/-1.39	125.70 +/- 44.1	9.76 +/- 1.75
SQIL	22.65+/-15	3.2+/-1	22.65+/-15	3.2+/-1	22.65+/-15	3.2+/-1

Table 14: Dataset 3 (Moderate weather, France)

	MSE (b=1)	MAE (b=1)	MSE (b=10)	MAE (b=10)	MSE (b=50)	MAE (b=50)
GAIL	25.3 +/-39	3.1 +/- 2.8	25.3 +/-39	3.1 +/- 2.8	25.3 +/-39	3.1 +/- 2.8
BC	23.2 +/- 28.6	3.4+/-2.4	22.3 +/- 26.1	3.5 +/-2.2	5.35+/-4.17	1.65 +/-0.86
AIRL	0+/-0	0+/-0	0+/-0	0+/-0	0+/-0	0+/-0
Dagger	22.9 +/-34.0	3.0 +/- 2.7	45.3 +/- 52.8	4.6 +/- 3.6	24.4 +/- 24.2	3.2 +/- 2.7
Density based reward	46.06 +/-55	4.60+/-4.8	46.06 +/-55	4.60+/-4.8	46.06 +/-55	4.60+/-4.8
MCE	148.2 +/- 39.6	10.3 +/-1.6	148.2 +/- 39.6	10.3 +/-1.6	148.2 +/- 39.6	10.3 +/-1.6
Pref-Comp	124.1 +/-52	9.4 +/- 2.78	128.29 +/- 42.7	9.86 +/- 1.68	125.68 +/- 44.19	9.7 +/- 1.7
SQIL	25.3 +/-20	3.2 +/- 1	25.3 +/-20	3.2 +/- 1	25.3 +/-20	3.2 +/- 1

Table 15: Dataset 4 (Cold weather, Australia)

	MSE (b=1)	MAE (b=1)	MSE (b=10)	MAE (b=10)	MSE (b=50)	MAE (b=50)
GAIL	45.71 +/- 45.7	4.9 +/- 2.8	45.71 +/- 45.7	4.9 +/- 2.8	45.71 +/- 45.7	4.9 +/- 2.8
BC	124.4 +/- 186.46	6.94 +/- 7.05	39.7+/- 70	3.91+/-3	26.7+/-42.7	3.23 +/- 3.17
AIRL	0+/-0	0+/-0	0+/-0	0+/-0	0+/-0	0+/-0
Dagger	113.4 +/-247.5	6.0+/-7.1	93.2 +/-115.9	6.5 +/- 5.1	25.8 +/- 47.1	6.5 +/- 5.1
Density based reward	115.7 +/- 242.51	5.5 +/-7.8	115.7 +/- 242.51	5.5 +/-7.8	115.7 +/- 242.51	5.5 +/-7.8
MCE	374 +/-311.9	15.0+/-7.6	374 +/-311.9	15.0+/-7.6	374 +/-311.9	15.0+/-7.6
Pref-Comp	284 +/-254	12.9 +/- 7	335.6 +/-271	14.5 +/-	332.8 +/- 272.29	14.3 +/-6.92
SQIL	25 +/- 16	4.52+/-2	25 +/- 16	4.52+/-2	25 +/- 16	4.52+/-2

Table 16: Dataset 5 (Hot weather, Australia)

10 CLUSTERING : OTHER VARIANTS

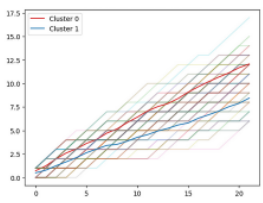


Figure 29: Bees trajectories

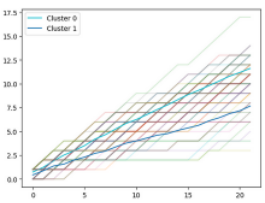


Figure 30: MAYA-Wass

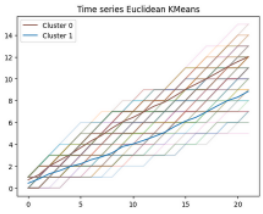


Figure 31: MAYA-KL

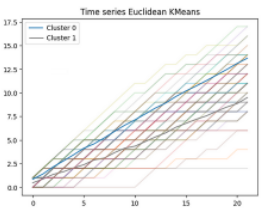


Figure 32: MAYA-DTW

Figure 33: Centroïdes of two clustering of 80 bees trajectories (in Fig29) and 80 MAYA-variant (Fig30, Fig31 and Fig32) simulated trajectories (with $\tau = 7$). Clustering are done with Euclidean method (Clustering I).

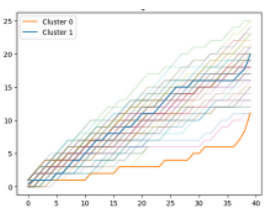


Figure 34: Bees trajectories

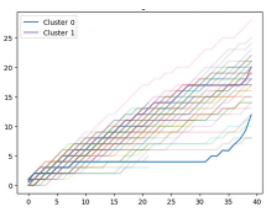


Figure 35: MAYA-Wass

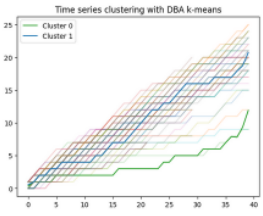


Figure 36: MAYA-KL

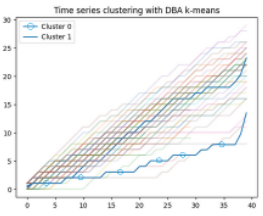


Figure 37: MAYA-DTW

Figure 38: Centroïdes of two clustering of 80 bees trajectories (in Fig34) and 80 MAYA-variant (Fig35, Fig36 and Fig37) simulated trajectories (with $\tau = 7$). Clustering are done with DBA method (Clustering II).

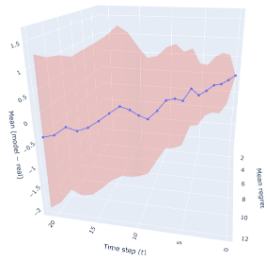


Figure 39: Cluster 0 (I)

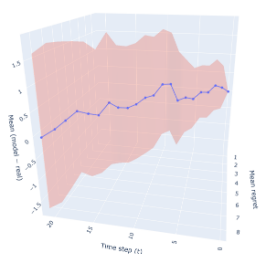


Figure 40: Cluster 1 (I)

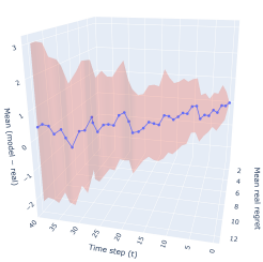


Figure 41: Cluster 0 (II)

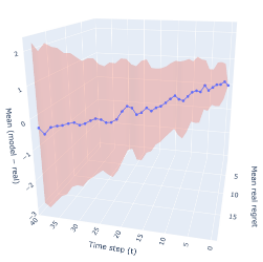


Figure 42: Cluster 1 (II)

Figure 43: Average difference between MAYA-Wass ($\tau = 7$) predictions and real trajectories ($R(\pi_{\texttt{MAYA}}, 1, t) - R(\pi_{\text{bee}}, 1, t)$) (z-axis) for Euclidean (I) and DBA (II) Clustering according 0 and 1 Cluster. Red range correspond to $\pm\sigma$ (standard deviation).

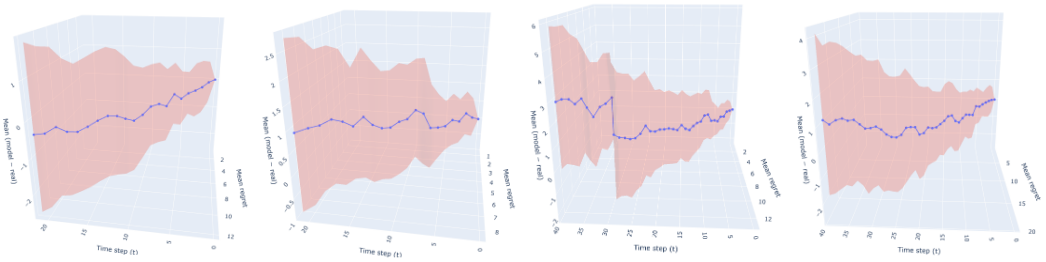


Figure 44: Cluster 0 (I)

Figure 45: Cluster 1 (I)

Figure 46: Cluster 0 (II)

Figure 47: Cluster 1 (II)

Figure 48: Average difference between MAYA-KL ($\tau = 7$) predictions and real trajectories ($R(\pi_{\texttt{MAYA}}, 1, t) - R(\pi_{\text{bee}}, 1, t)$) (z-axis) for Euclidean (I) and DBA (II) Clustering according 0 and 1 Cluster. Red range correspond to $\pm\sigma$ (standard deviation).

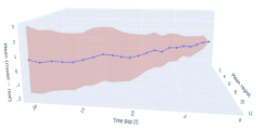


Figure 49: Cluster 0 (I)

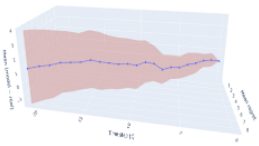


Figure 50: Cluster 1 (I)

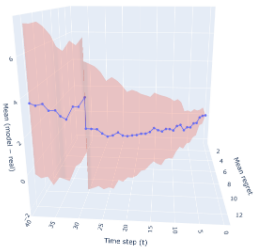


Figure 51: Cluster 0 (II)

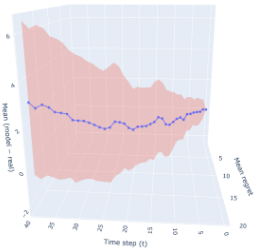


Figure 52: Cluster 1 (II)

Figure 53: Average difference between MAYA-DTW ($\tau = 7$) predictions and real trajectories ($R(\pi_{\texttt{MAYA}}, 1, t) - R(\pi_{\text{bee}}, 1, t)$) (z-axis) for Euclidean (I) and DBA (II) Clustering according 0 and 1 Cluster. Red range correspond to $\pm\sigma$ (standard deviation).

11 MAYA ALGORITHM

Algorithm 1 MAYA : Multi Agent Y-maze Allocation

Require: Logged bee regret trajectory $R(\pi_{\text{bee}}, 1, T)$
Require: Set $\mathcal{P}$ of N bandit policies $\{\pi_1, \ldots, \pi_N\}$
Require: Window size τ such that $t \geq \tau$
Require: A similarity metric δ
$\xi = ()_{t=1}^{T}$
Init π_θ
for $t \in \{2, \ldots, \tau - 1\}$ **do**
    Observe $R(\pi_{\text{bee}}, 1, t-1)$
    Observe a context information x_t
    for $i = 1$ to N **do**
        Simulate policy agent $\pi_i(s_{t-1}|x_t)$
        Compute cumulative regret $R(\pi_i, 1, t-1)$
    end for
    $\xi_t = \text{argmin}_{\pi \in \mathcal{P}}\ \delta(\pi_{\text{bee}}, \pi, t)$
    $\pi_\theta(a_t|s_{t-1}) \leftarrow \pi_\xi(a_t|s_{t-1})$
    Select $A_t \sim \pi_\theta(a_t|s_{t-1})$
    Receive reward r_t
    Update $\pi_i \quad \forall \pi_i \in \mathcal{P}$
    $\xi[t] \leftarrow \xi_t$
end for
for $t \in \{\tau, \ldots, T\}$ **do**
    Observe $R(\pi_{\text{bee}}, \tau, 1, t-1)$
    Observe a context information x_t
    for $i = 1$ to N **do**
        Simulate policy agent $\pi_i(s_{t-1}|x_t)$
        Compute cumulative regret $R(\pi_i, \tau, 1, t-1)$
    end for
    $\xi_t = \text{argmin}_{\pi \in \mathcal{P}}\ \delta(\pi_{\text{bee}}, \pi, \tau, t)$
    $\pi_\theta(a_t|s_{t-1}) \leftarrow \pi_\xi(a_t|s_{t-1})$
    Select $A_t \sim \pi_\theta(a_t|s_{t-1})$
    Receive reward r_t
    Update $\pi_i \quad \forall \pi_i \in \mathcal{P}$
    $\xi[t] \leftarrow \xi_t$
end for
return π_θ

12 MICE DATASET EXPERIMENT

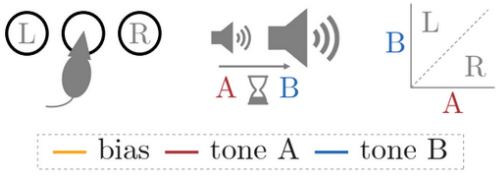
Dataset and setup. We use the dataset of Ashwood et al. (2020a), which reports trial-by-trial changes in mice policy and decomposes those updates into a learning component and a noise component (see Fig. 54a). Unlike their original analysis, which simulates an average trajectory across individuals, our method (MAYA) simulates one trajectory *per* individual. The dataset contains 19 rats with between 1500 and 6000 trials each. To control the computational cost of DTW and to align with our bee experiments, we reduce the number of individual at 100.

Selecting the memory horizon τ. According with Tab 17, Fig 54b shows MAE and MSE as a function of the memory window τ. MAYA-KL clearly identifies an optimal range around $\tau \in [6, 7]$, whereas MAYA-Wass suggests $\tau \in [8, 10]$ when balancing MAE and MSE. For consistency with previous experiments, we set $\tau = 7$ in all subsequent analyses.

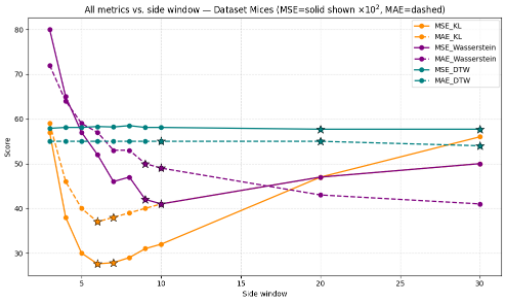
Explanations and performance. With $\tau = 7$, Fig. 63 and Fig. 59 provides MAYA explanations for the rats with the lowest and highest cumulative regret (see Fig. 55). For slow learners, all MAYA variants behave similarly (Fig. 65); for fast learners, MAYA-KL achieves the best fit, capturing rapid policy changes better than MAYA-Wass (Fig. 64). A plausible explanation is that, under KL similarity, MAYA acts more often from LinUCB-like behavior than with Wasserstein similarity (see Tab18b). As in previous datasets, MAYA-DTW tends to act more like Epsilon-Greedy, likely due to DTW's alignment properties. Overall, all MAYA variants outperform GLM baselines (Table 18a).

side_window	MSE MAYA-KL		MAE MAYA-KL		MSE MAYA-Wass		MAE MAYA-Wass		MSE MAYA-DTW		MAE MAYA-DTW	
	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
3	5760	3894	59	24	8083	5012	72	25	5790	5683	55	29
4	3868	3493	46	25	6547	3672	64	23	5815	5770	55	30
5	3046	3307	40	24	5724	3803	59	23	5819	5788	55	29
6	2763	3090	37	23	5276	3511	57	21	5830	5758	55	29
7	2786	3161	38	23	4640	3382	53	22	5822	5747	55	29
8	2974	3197	39	23	4728	3722	53	23	5851	5777	55	29
9	3114	3424	40	24	4231	3403	50	22	5819	5740	55	29
10	3223	3378	41	25	4197	3576	49	24	5810	5701	54	29
20	4710	6689	47	33	3491	3515	43	25	5771	5725	54	29
30	5618	8543	50	38	3453	3896	41	27	5760	5724	54	29

Table 17: MSE and MAE of MAYA as a function of the window size τ for Mice Dataset.



(a) According Ashwood et al. (2020a), on each trial, a sinusoidal grating (with contrast values between 0 and 100%) appears on either the left or right side of a screen. Mice must report the side of the grating by turning a wheel (left or right) in order to receive a water reward.



(b) Comparative study of the best window size τ by average MSE and MAE. $\star$ symbol refers as best performance according standard deviation and average reward (see Tab.17 for the full results). MSE is displayed as $\times 10^2$.

Figure 54: Left : experimental description of the Mice Dataset. Right : Comparative study of the best window size τ for Mice Dataset.

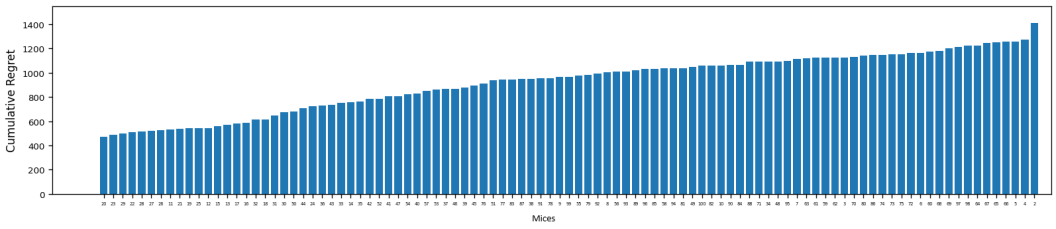


Figure 55: Proportion of cumulative regret for the Mice dataset, per mice

	MSE		MAE	
	Mean	Std	Mean	Std
MAYA KL	2786	3161	38	23
MAYA-Wass	4640	3382	53	22
MAYA-DTW	5822	5777	55	29
GLM	6427	4137	63	21
GLM Contextual	6416	4133	63	21

(a)

	Epsilon-Greedy	Lin-UCB	UCB	Uniform
MAYA-KL	30%±2.5	2%±1.1	29%±1.3	36%±2.2
MAYA-W	27%±1.8	10%±1	28%±1	33%±1.5
MAYA-DTW	28%±3	0.5%±1	56%±4	15%±3

(b)

Table 18: **Left** : MSE and MAE comparison of MAYA (with $\tau = 7$) and GLM variants. **Right** : MAYA explainability for all MAYA choices ($\tau = 7$)

	MAYA-KL	MAYA-Wass	MAYA-DTW
ClusterAcc (Euclidean, Max L = 1400)	90%	85%	75%
ClusterAcc (DBA, Max L = 6000)	80%	75%	65%

Table 19: ClusterAcc (%) for Mice Datset)

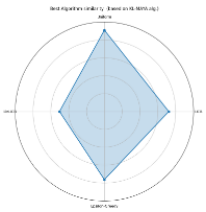


Figure 56: MAYA-KL

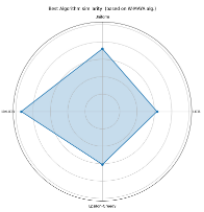


Figure 57: MAYA-Wass

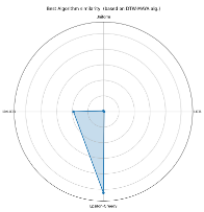


Figure 58: MAYA-DTW

Figure 59: MAYA explainability for mouse 20 (fast learner, low regret) from Mice dataset. We report choice interpretability for MAYA-variants ($\tau = 7$).

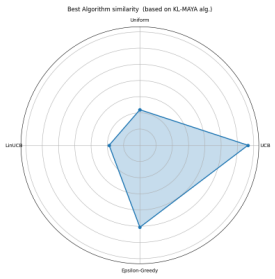


Figure 60: MAYA-KL

Figure 61: MAYA-Wass

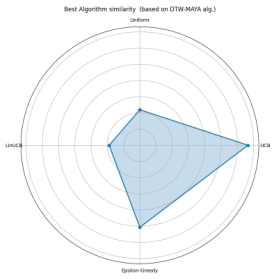


Figure 62: MAYA-DTW

Figure 63: MAYA explainability for mouse 2 (slow learner, high regret) from Mice dataset. We report choice interpretability for MAYA-variants ($\tau = 7$).

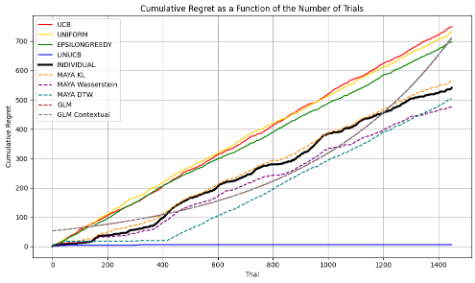


Figure 64: Mouse 20

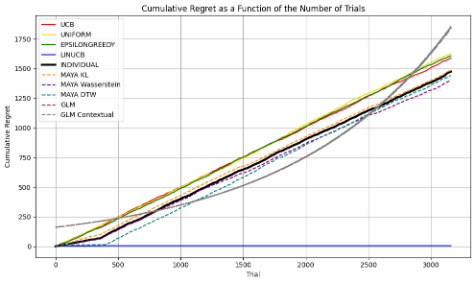


Figure 65: Mouse 2

Figure 66: Regret modelization for mouse 20 (best) and mice 2 (worst) from Mice 2, with $\tau = 7$

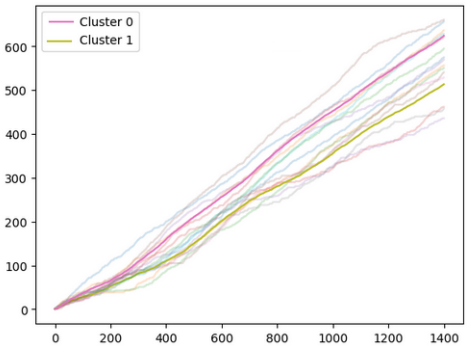


Figure 67: Mouse' trajectories

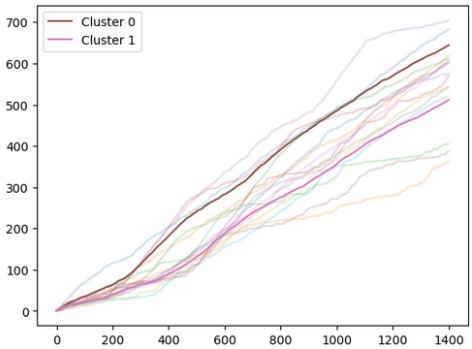


Figure 68: MAYA-KL trajectories

Figure 69: Centroides of Clustering (I) of 100 mice' (**Left**) and MAYA-KL ($\tau = 7$) (**Right**) trajectories.

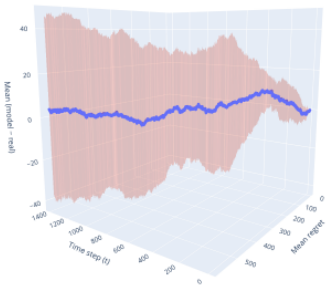


Figure 70: Cluster 0

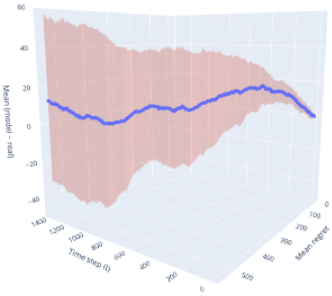


Figure 71: Cluster 1

Figure 72: Average difference between MAYA-KL ($\tau = 7$) predictions and real trajectories ($R(\pi_{\texttt{MAYA}}, 1, t) - R(\pi_{\text{mice}}1, t)$) (z-axis) for Euclidean (I) Clustering according 0 and 1 Cluster. Red range correspond to $\pm\sigma$ (standard deviation).

13 COMPLEMENTARY INFORMATION ABOUT THE BIOLOGY INTEREST

We share with other vertebrates a basic ability for abstract number representation, the *number sense* Dehaene (2011). As early as two days postnatally Izard et al. (2009), this ability enables us to evaluate numbers as concepts: three books are perceived as similar to three cups, even though they differ completely in their visual features (i.e., sensory information). To evaluate quantity, both numerical and sensory information can be used. For example, when visually comparing two quantities, the larger set will often contain more items (i.e., numerosity), but may also exhibit greater density, a larger total surface area, or a wider convex hull encompassing all elements. Neuronal encoding of sensory information occurs early in the primary cortex, whereas numbers are computed in higher integrative areas by what Nieder et al. identified as *number neurons* Nieder (2016).

Quantity discrimination is necessary in contexts as diverse as evaluating food patches, regulating social attraction, or competing for resources Nieder (2020). From sharks to mammals, all major vertebrate clades appear capable of discriminating between different quantities, either spontaneously or in learning tasks Vila Pouca et al. (2019). By carefully designing protocols that control for sensory cues, researchers have demonstrated that several non-human species are capable of performing quantity discrimination based on the abstract evaluation of numbers Cantlon & Brannon (2006). Among them is an insect: the honeybee (*Apis mellifera*). Beyond discriminating numerosities of up to eight items, these insects, with brains of fewer than one million neurons, can also manipulate numbers, performing simple addition, subtraction, and symbolic tasks Dacke & Srinivasan (2008); Gross et al. (2009); Howard et al. (2018; 2019); Giurfa et al. (2022).

Later experiments required a Y-maze: a three-armed apparatus shaped like the letter *Y*, commonly used to study memory, learning, and decision-making in rodents Kraeuter et al. (2018) (see Fig. 73). These mazes required bees to inhibit their spatial memory Menzel et al. (2005) (e.g., recalling that the last reward was in the left arm) and to focus instead on the visual stimuli displayed at the end of each arm. The balance between exploring new options and exploiting previously rewarded ones is key to their foraging behavior and likely plays a crucial role in their learning performance within these devices Kembro et al. (2019); Lochner et al. (2024).

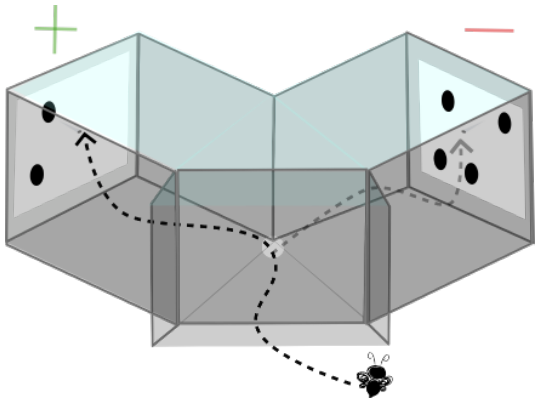


Figure 73: Y-maze for bees experiments

14 MATHEMATICAL PROOF OF MAYA ACCORDING τ

Stationary case (1) : upper bound of MAYA error Consider the case of two policies π_1 that achieves the highest regret i.e. $R(\pi_1, 1, T) = T$ and π_0 that achieves a zero regret i.e. $R(\pi_0, 1, T)$.In this case

$$\Delta_{\pi_1,t} - \Delta_{\pi_0,t} \leq 1 \quad \forall t$$

as the reward is in $\{0, 1\}$. The maximal bound of $R(\pi_{\text{MAYA}}, 1, T) - R(\pi_{\text{bee}}, 1, T)$ corresponds to the case where $R(\pi_{\text{bee}}, 1, T)$ is always centered between $R(\pi_1, 1, T)$ and $R(\pi_0, 1, T)$ (see Fig74a). Let's define ε_t^* the agent who act the closest of the bee at t and ε_t the agent chosen by MAYA at t. Then

$$\mathbb{P}[\varepsilon_t = \varepsilon_t^*] = 0.5 \quad \forall t$$

as no best agent are better from the other one. This case corresponds to an equality between the two possible agent (with extreme regret values) and leads to the worst scenario of a stationary case when the similarity distance $d()$ are when define. Then the maximal cumulative gap between MAYA-regret and Bee-regret in stationary case are :

$$\begin{aligned}
\sum_{t=1}^{T} |\Delta_{\text{MAYA},t} - \Delta_{\text{Bee},t}| &\leq \frac{1}{2}\sum_{t=1}^{T} |\Delta_{\pi_1,t} - \Delta_{\text{Bee},t}| + \frac{1}{2}\sum_{t=1}^{T} |\Delta_{\pi_0,t} - \Delta_{\text{Bee},t}| \\
&\leq \sum_{t=1}^{T} \frac{t}{2} \\
&\leq \frac{\frac{T}{2}(\frac{T}{2}+1)}{2} \\
&\leq \frac{1}{8}(T(T+2)) \qquad (1)
\end{aligned}$$

Stationary case (2) : upper bound of the worst policy Consider the case where π_{MAYA} always chose like π_1 and π_{bee} always chose like π_0 (see Fig 74b). Then the similarity distance $d()$ fails to provide a correct measure and MAYA chose the agent with the largest regret gap relative to the bee's regret. Then for all t

$$\mathbb{P}[\varepsilon_t \neq \varepsilon_t^*] = 1.$$

Then the maximal cumulative gap between MAYA-regret and Bee-regret in the worst policy in stationary case are :

$$\begin{aligned}
\sum_{t=1}^{T} |\Delta_{\text{MAYA},t} - \Delta_{\text{Bee},t}| &\leq \sum_{t=1}^{T} |\Delta_{\pi_1,t} - \Delta_{\pi_0,t}| \\
&\leq \frac{T \cdot (T+1)}{2} \qquad (2)
\end{aligned}$$

The alternative case where π_{MAYA} always chooses as π_0 and π_{bee} always chooses as π_1 is equivalent.

Cyclic case : upper bound of MAYA error with no windows ($\tau = T$) policy Consider that after S trials the bee moves from π_1 to π_0 (alternative cases are equivalent, see Fig 75a). Consider that the distances are well defined, as in the stationary case (1). Then :

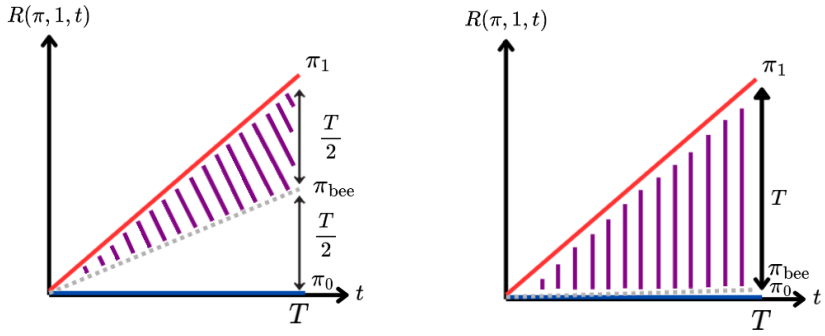
$$\sum_{t=1}^{S} |\Delta_{\text{MAYA},t} - \Delta_{\text{Bee},t}| \leq \frac{1}{8}(S \times (S+2)) \qquad (3)$$

The time required for MAYA to act like π_0 is $2S + 1$ but at $t = 2S + 1$, the bee changes from π_0 to π_1 and MAYA continues to act like π_1 (see Fig.75a). Recursively, MAYA always act like π_1 from $t = 1$ until $t = T$. Then

$$\mathbb{P}[\varepsilon_t = \pi_1] = 1 \quad \forall t$$

and

$$\mathbb{P}[\varepsilon_t = \varepsilon_t^*] = \frac{N_*(T)}{T}, \quad \forall t$$



(a) Distance $d(\cdot)$ provide a correct measure, $R(\pi_1, 1, T)$, and $R(\pi_2, 1, T)$ has the maximal distance from $R(\pi_{\text{bee}}, 1, T)$.

(b) Distance $d(\cdot)$ fails to provide a correct measure. MAYA alawys selects actions as the agent whose behavior is farthest from that of the bee.

Figure 74: Maximal cumulative gap between MAYA-regret and Bee-regret in **stationary case** according the distance $d(\cdot)$ abilities to provide a correct measure

Where

$$N_*(T) = qS + \min(S, r),$$
$$q = \left\lfloor \frac{T}{2S} \right\rfloor,$$
$$r = T - 2Sq \in [0, 2S).$$

A minimal bound of N_* are :

$$N_*(T) \geq \frac{T}{2}$$

Then the maximal cumulative gap between MAYA-regret and Bee-regret in a cyclic case with no windows is :

$$\begin{aligned}\sum_{t=1}^{T} |\Delta_{\text{MAYA},t} - \Delta_{\text{Bee},t}| &\leq \frac{N_*(T)}{T}\frac{1}{8}(T.(T+2)) + (1 - \frac{N_*(T)}{T})\frac{T.(T+1)}{2} \\ &\leq \frac{T}{2}\frac{1}{T}\frac{1}{8}(T.(T+2)) + (1 - \frac{T}{2}\frac{1}{T})\frac{T.(T+1)}{2} \\ &= \frac{T(5T+6)}{16} \end{aligned} \tag{4}$$

Cyclic case : upper bound of MAYA error with windows $\tau = S$ Assume that S are even. Consider that after S trials, the bee moves from π_1 to π_0 (alternative cases are equivalent, see Fig75b). Consider that the distance is well define like in the stationary case (1). From time $t = 1$ until S, MAYA act as the best agent :

$$\sum_{t=1}^{S} |\Delta_{\text{MAYA},t} - \Delta_{\text{Bee},t}| \leq \frac{1}{8}(S \times (S+2)) \tag{5}$$

and

$$\mathbb{P}[\varepsilon_t = \varepsilon_t^*] = 1 \quad \forall t \in \{1, \ldots, S\}.$$

From time $S+1$ until $S + \frac{S}{2}$, MAYA acts as the worst policy (start cycle)

$$\sum_{t=S+1}^{S+\frac{S}{2}} |\Delta_{\text{MAYA},t} - \Delta_{\text{Bee},t}| \leq \sum_{t=S+1}^{S+\frac{S}{2}} t \tag{6}$$

$$\leq \frac{S(5S+2)}{8} \tag{7}$$

and

$$\mathbb{P}[\varepsilon_t \neq \varepsilon_t^*] = 1 \quad \forall t \in \{S+1, \ldots, S+\frac{S}{2}\}.$$

And from $t = S + \frac{S}{2} + 1$ until $t = 2S$ MAYA acts with the best policy (end cycle):

$$\begin{aligned}\sum_{t=S+\frac{S}{2}+1}^{2S} |\Delta_{\text{MAYA},t} - \Delta_{\text{Bee},t}| &\leq \sum_{t=S+\frac{S}{2}+1}^{2S} \frac{t}{2} \\ &\leq \frac{S(7S+2)}{16}\end{aligned} \tag{8}$$

and

$$\mathbb{P}[\varepsilon_t = \varepsilon_t^*] = 1 \quad \forall t \in \{S+\frac{S}{2}+1, \ldots, 2S\}.$$

Consider a full cycle, the event $\varepsilon_t = \varepsilon_t^*$ appears $S - \frac{S}{2}$ times. Let's set

$$q = \Big\lfloor \frac{\max(0, T-S)}{S} \Big\rfloor, \qquad r = \max(0, T-S) - qS \in [0, S).$$

Here q is the number of full cycle S in $t > S$, and r is the rest of a potential unfinished tail segment of the started cycle. Let $N_*(T) = \sum_{t=1}^{T} 1_{\varepsilon_t = \varepsilon^*}$ with $N_*(T) \leq T$ equal to

$$N_*(T) = \min(T, S) + q \cdot \frac{S}{2} + \max(0, r - \frac{S}{2})$$

If S is even and $T > S$ then

$$N_*(T) \geq \frac{T}{2} + \frac{S}{4} \tag{9}$$

Proof:
With $T = S + qS + r$:

$$N_*(T) - (\frac{T}{2} + \frac{S}{4}) = \frac{S}{2} - \frac{r}{2} + \max(0, r - \frac{S}{2}) \geq 0,$$

where the minimum are archived with $r = \frac{S}{2}$.

$$\mathbb{P}[\varepsilon_t = \varepsilon_t^*] = \frac{N_*(T)}{T} \geq \frac{1}{2} + \frac{S}{4T} \tag{10}$$

In the cases where S is not not even

$$q = \Big\lfloor \tfrac{T-S}{S} \Big\rfloor, \qquad r = T - S - qS \in [0, S).$$

then

$$N_*(T) = S + \frac{q(S+1)}{2} + \max\Big(0,\ r - \frac{S-1}{2}\Big).$$

As $T = S + qS + r$, we have

$$N_*(T) - \frac{T}{2} = \frac{S}{2} + \frac{q}{2} + \max\Big(0,\ r - \frac{S-1}{2}\Big) - \frac{r}{2}.$$

and for any $r \in [0, S)$,

$$\min_r \Big(\max(0, r - \tfrac{S-1}{2}) - \tfrac{r}{2}\Big) = -\frac{S-1}{4}.$$

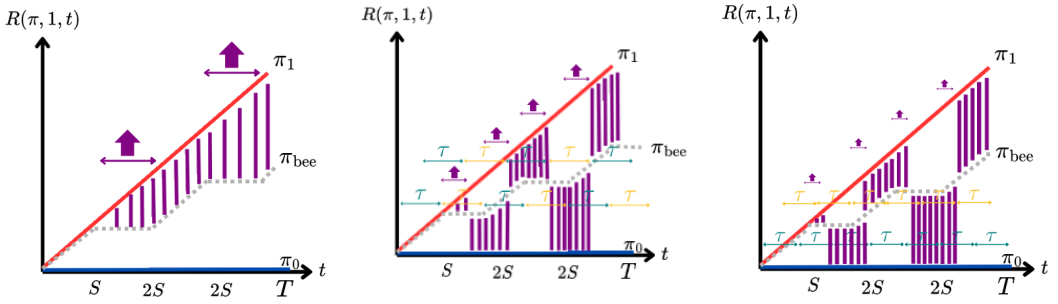
Then

$$N_*(T) \ \geq \ \frac{S}{2} + \frac{q}{2} - \frac{S-1}{4} + \frac{T}{2} = \frac{S+1}{4} + \frac{q}{2} + \frac{T}{2} \ \geq \ \frac{S+1}{4} + \frac{T}{2}.$$

$$N_*(T) \ \geq \ \frac{T}{2} + \frac{S+1}{4} \ \ \geq \ \ \frac{T}{2} + \frac{S}{4}\ . \tag{11}$$

Which are better to the S parity case.
Then the maximal cumulative gap between MAYA-regret and Bee-regret with windows $\tau = S$ is

$$\begin{aligned}\sum_{t=1}^{T} |\Delta_{\text{MAYA},t} - \Delta_{\text{Bee},t}| &\leq \frac{N_*(T)}{T}\frac{T(T+2)}{8} + (1 - \frac{N_*(T)}{T})\frac{T(T+1)}{2} \\ &\leq (\frac{T}{2} + \frac{S}{4}).\frac{1}{T}.\frac{T(T+2)}{8} + (1 - (\frac{T}{2} + \frac{S}{4}).\frac{1}{T})\frac{T(T+1)}{2} \\ &\leq \frac{10T^2 + 12T - 3ST - 2ST}{32} \end{aligned} \tag{12}$$



(a) $\tau = T$ (b) $\tau = S$ (c) $\tau \in \{\frac{S}{2} + 1; \ldots, S - 1\}$

Figure 75: Maximal cumulative gap between MAYA regret and bee regret in a non-stationary case, measured with respect to window τ. The purple arrow highlights the period during which MAYA chooses actions in accordance with the agent whose behavior is most distant from that of the bee.

Cyclic case : upper bound of MAYA error with windows $\tau \in \{\frac{S}{2} + 1; \ldots, S - 1\}$. We consider the case where $\frac{S}{2} + 1 \leq \tau < S$ (see Fig75c). Assume that S are even. From time $t = 1$ until S, MAYA act as the best agent (stationary case 1) :

$$\sum_{t=1}^{S} |\Delta_{\text{MAYA},t} - \Delta_{\text{Bee},t}| \leq \frac{1}{8}(S \times (S + 2)) \tag{13}$$

and

$$\mathbb{P}[\varepsilon_t = \varepsilon^*] = 1 \quad \forall t \in \{1, \ldots, S\}.$$

From time $S + 1$ until $S + \frac{\tau}{2}$, MAYA acts as the worst policy (start cycle)

$$\begin{aligned}\sum_{t=S+1}^{S+\frac{\tau}{2}} |\Delta_{\text{MAYA},t} - \Delta_{\text{Bee},t}| &\leq \sum_{t=S+1}^{S+\frac{\tau}{2}} t \\ &\leq \frac{\tau}{4}(2S + 1 + \frac{\tau}{2}) \\ &\leq \frac{\tau^2}{8} + \frac{S\tau}{2} + \frac{\tau}{4} \end{aligned} \tag{14}$$

and

$$\mathbb{P}[\varepsilon_t \neq \varepsilon^*] = 1 \quad \forall t \in \{S + 1, \ldots, S + \frac{\tau}{2}\}.$$

And from $t = S + \frac{\tau}{2} + 1$ until $t = 2S$, MAYA acts as the best policy (end cycle) with :

$$\begin{aligned}\sum_{t=S+\frac{\tau}{2}+1}^{2S} |\Delta_{\text{MAYA},t} - \Delta_{\text{Bee},t}| &\leq \sum_{t=S+\frac{\tau}{2}+1}^{2S} \frac{t}{2} \\ &\leq \frac{(3S + \frac{\tau}{2} + 1)(S - \frac{\tau}{2})}{4} \end{aligned} \tag{15}$$

and

$$\mathbb{P}[\varepsilon_t = \varepsilon_t^*] = 1 \quad \forall t \in \{S + \frac{\tau}{2} + 1, \dots, 2S\}.$$

Consider a full cycle, the event $\varepsilon_t = \varepsilon_t^*$ appears $S - \frac{\tau}{2}$ times. Let's set

$$q = \lfloor \frac{T-S}{S} \rfloor \qquad r = (T-S) - qS \in [0, S).$$

Let $N_*(T) = \sum_{t=1}^{T} 1_{\varepsilon_t = \varepsilon^*}$ with $N_*(T) \leq T$ equal to

$$N_*(T) = S + q(S - \frac{\tau}{2}) + \max(0, r - \frac{\tau}{2}).$$

and

$$\mathbb{P}[\varepsilon_t = \varepsilon_t^*] = \frac{N_*(T)}{T} \tag{16}$$

The maximal cumulative gap between MAYA-regret and Bee-regret with windows $\tau \in \{\frac{S}{2} + 1; \dots, S-1\}$ with S parity is

$$\begin{aligned}
\sum_{t=1}^{T} |\Delta_{\text{MAYA},t} - \Delta_{\text{Bee},t}| &\leq \frac{N_*(T)}{T} \cdot \frac{T(T+2)}{8} + \Big(1 - \frac{N_*(T)}{T}\Big) \cdot \frac{T(T+1)}{2} \\
&\leq \frac{S + q(S - \frac{\tau}{2}) + \max(0, r - \frac{\tau}{2}).}{T} \cdot \frac{T(T+2)}{8} \\
&+ (1 - \frac{S + q(S - \frac{\tau}{2}) + \max(0, r - \frac{\tau}{2}).}{T}) \cdot \frac{T(T+1)}{2}
\end{aligned}$$

As $N_*(T) \geq T(1 - \frac{\tau}{2S})$ without any condition on S parity, the maximal cumulative gap between the MAYA-regret and the Bee-regret with windows $\tau \in \{\frac{S}{2} + 1; \dots, S-1\}$ is

$$\sum_{t=1}^{T} |\Delta_{\text{MAYA},t} - \Delta_{\text{Bee},t}| \leq \frac{T(T+2)}{8} + \frac{(3T+2)T}{16} \frac{\tau}{S} \tag{17}$$

Cyclic case : upper bound of MAYA with windows $\tau < \frac{S}{2} + 1$ In this case, there is no way to be sure that the distance $d()$ do not fails to identify the best agent. It's equivalent to choose randomly and the worst case corresponds to the upper bound of the worst policy. Then the maximal cumulative gap between MAYA regret and Bee-regret with $\tau < \frac{S}{2} + 1$ in cyclic case are equivalent to Eq. 2.

Cyclic case : upper bound of MAYA with windows $\tau > S$ In this case, the time required to change the policy is over a cycle $S > 1$. Then, the bee switch two times in τ and MAYA allows it to act as the same agent. Then it is equivalent to act as a cyclic case with no windows ($\tau = T$) Then the maximal cumulative gap between MAYA regret and Bee-regret with $\tau > S$ in cyclic case are equivalent to Eq. 4.

15 DISCLOSURE OF LLM USE

Large Language Models (LLMs) were used in a limited capacity during the preparation of this paper. Their use was restricted to (i) spelling and phrasing assistance (to support a dyslexic co-author), and (ii) suggesting improvements to Python scripts for graph generation and visualization. No part of the scientific content, analyses, or conclusions was produced by LLMs.