Ethical issues in multi-objective reinforcement learning

Marius Le Chapelier

Encadrants :

Aurélie Beynier Nicolas Maudet Paolo Viappiani

Objectif et direction du stage

Problématique :
Comment intégrer l'éthique dans l'IA ? Comment répondre à des problèmes éthiques avec des systèmes de décisions autonomes ?

Objectif :
Développer un modèle d'apprentissage par renforcement multi-objectif (MORL) pour répondre à ces problématiques liées à l'éthique.

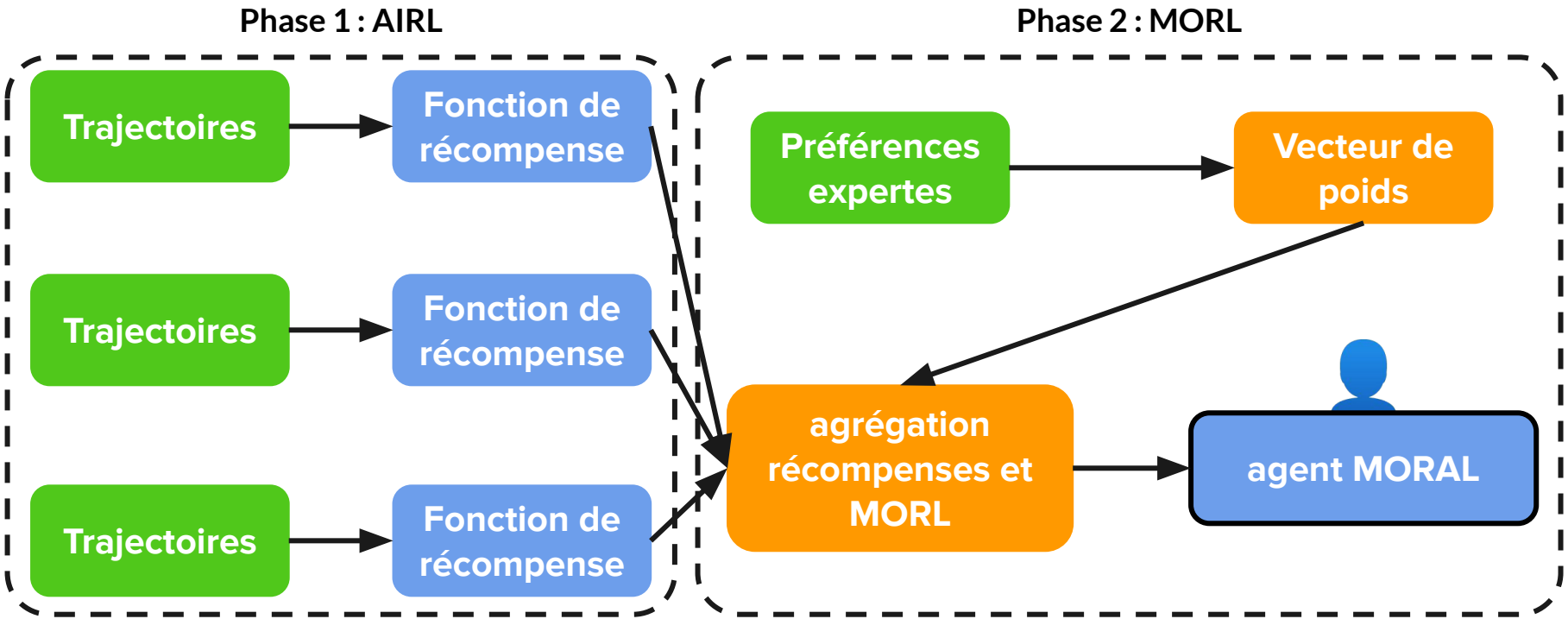
Direction :
Utiliser de l'apprentissage par préférences pour capturer le caractère éthique des actions (récompenses MORL)

Article constituant la base du stage

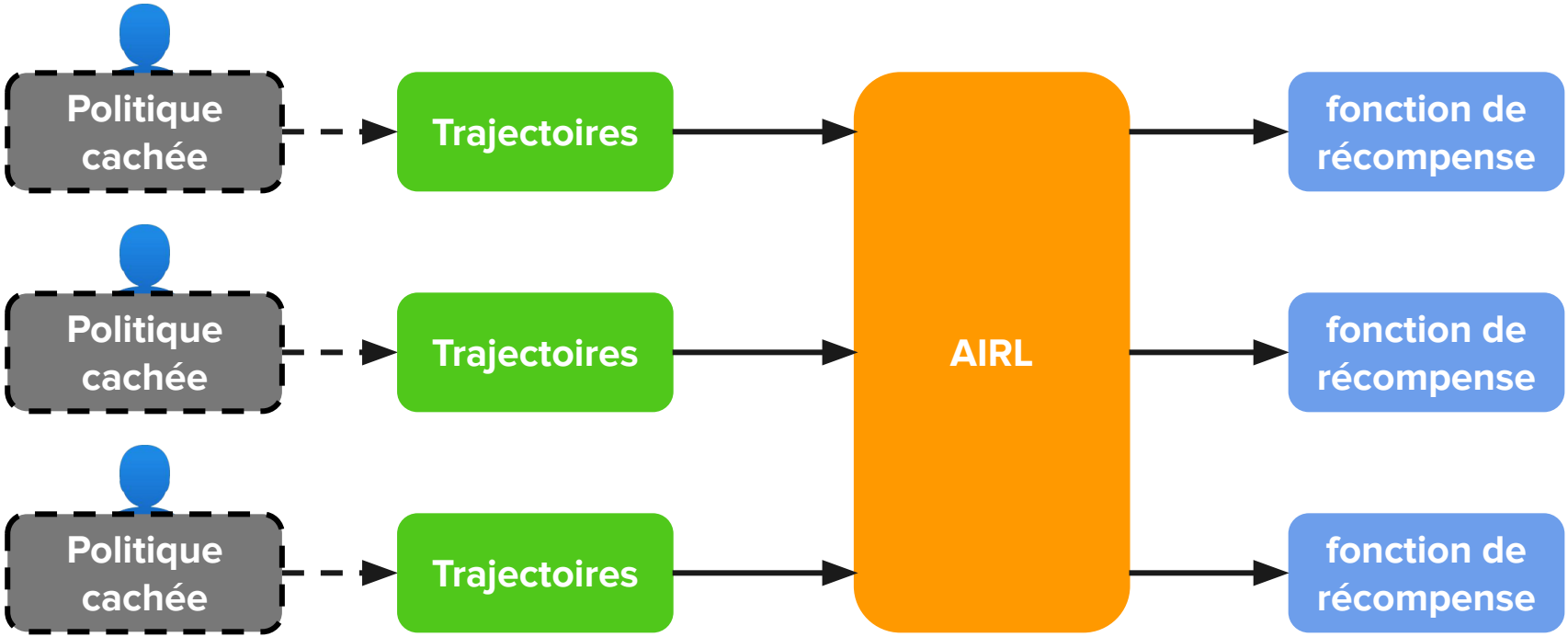
Markus Peschl et al. “MORAL: Aligning AI with Human Norms through Multi-Objective Reinforced Active Learning”. In: 21st International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2022, Auckland, New Zealand, May 9-13, 2022. Ed. by Piotr Faliszewski et al. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), 2022, pp. 1038–1046. url: https://www.ifaamas.org/Proceedings/aamas2022/pdfs/p1038.pdf.

Objectif :
Apprendre à extraire et imiter plusieurs comportements éthiques induits par un ensemble de données.

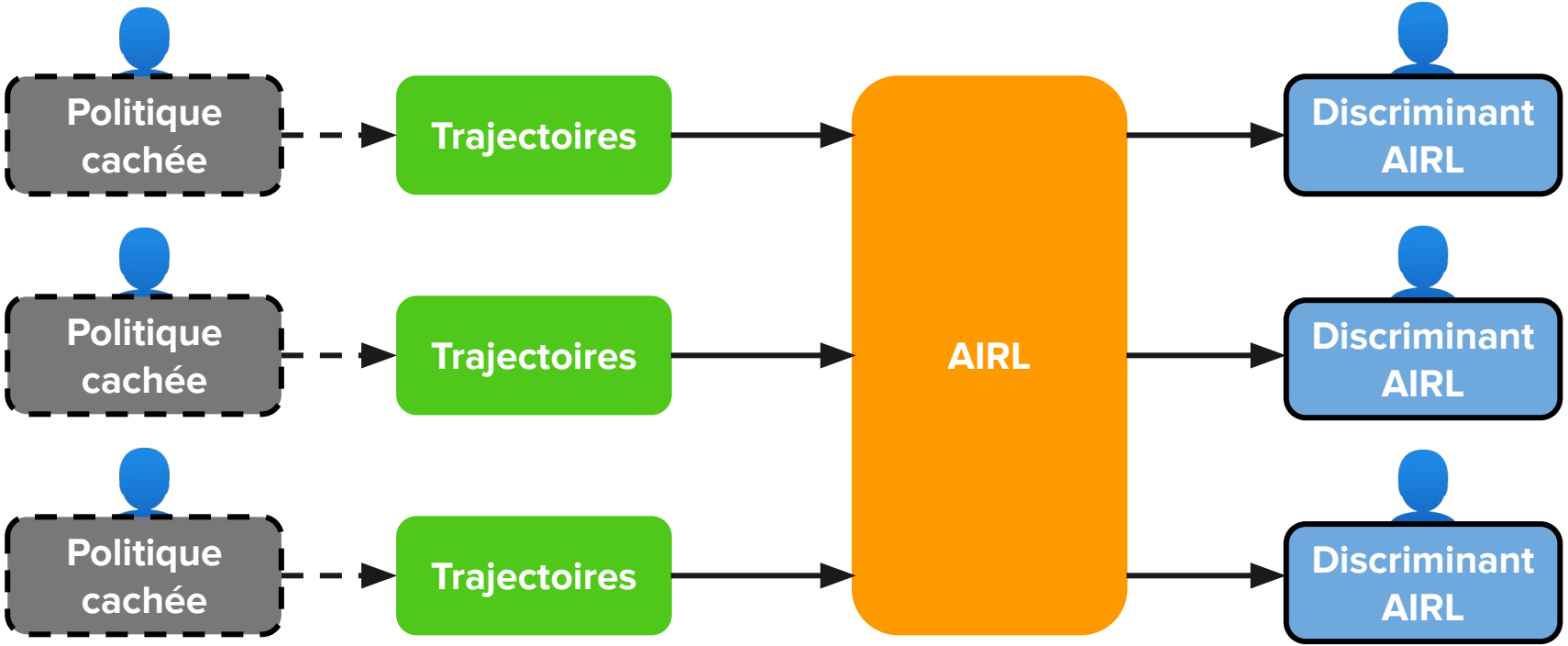
MORAL


Phase 1 : AIRL
Phase 2 : MORL
Trajectoires
Trajectoires
Trajectoires
Fonction de récompense
Fonction de récompense
Fonction de récompense
Préférences expertes
Vecteur de poids
agrégation récompenses et MORL
agent MORAL

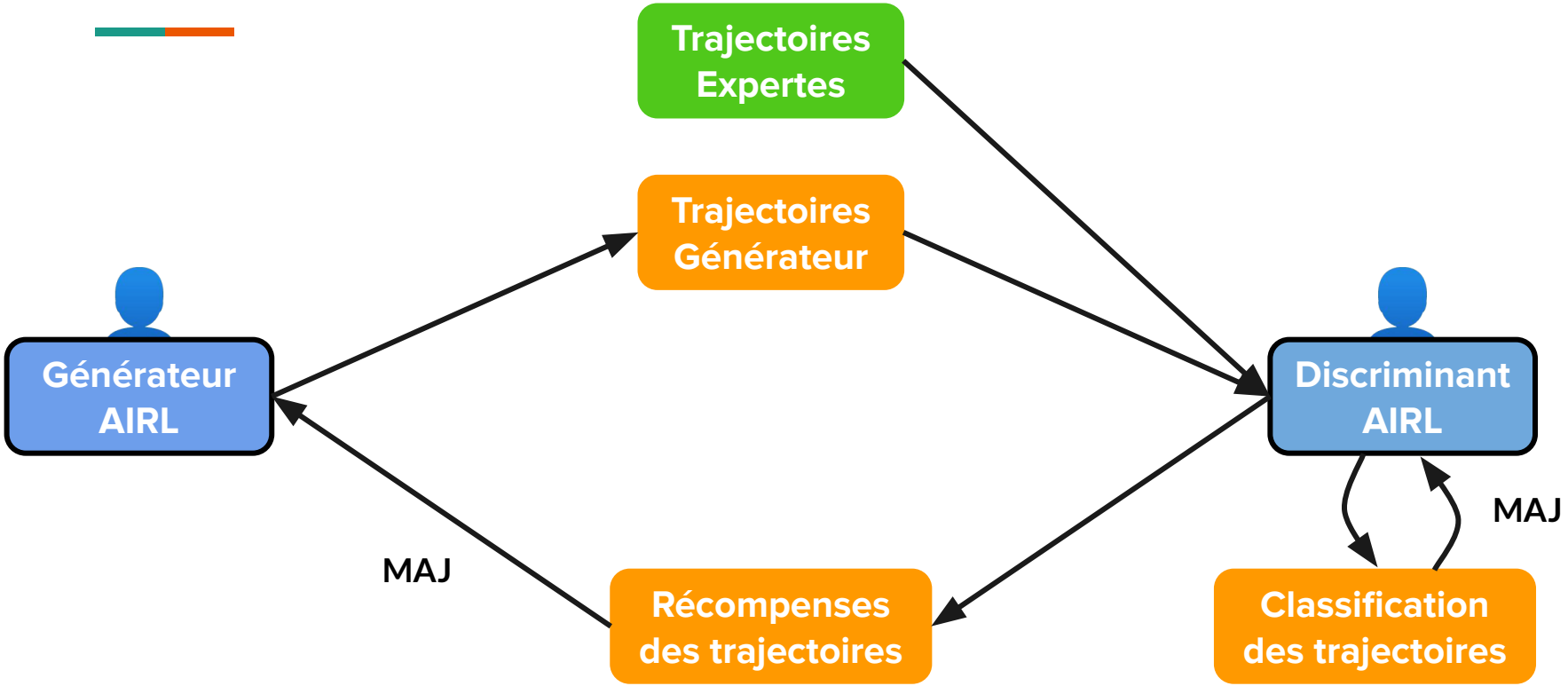
MORAL Phase 1 : AIRL



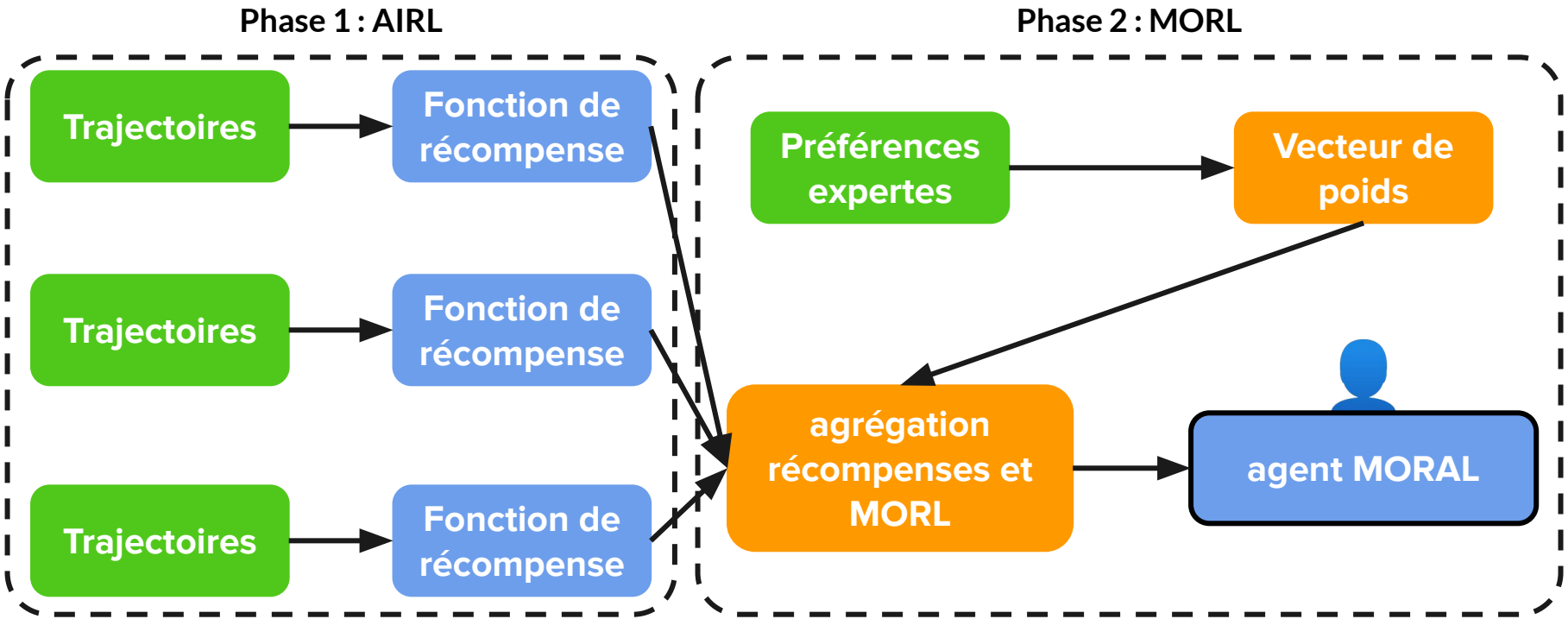
MORAL Phase 1 : AIRL



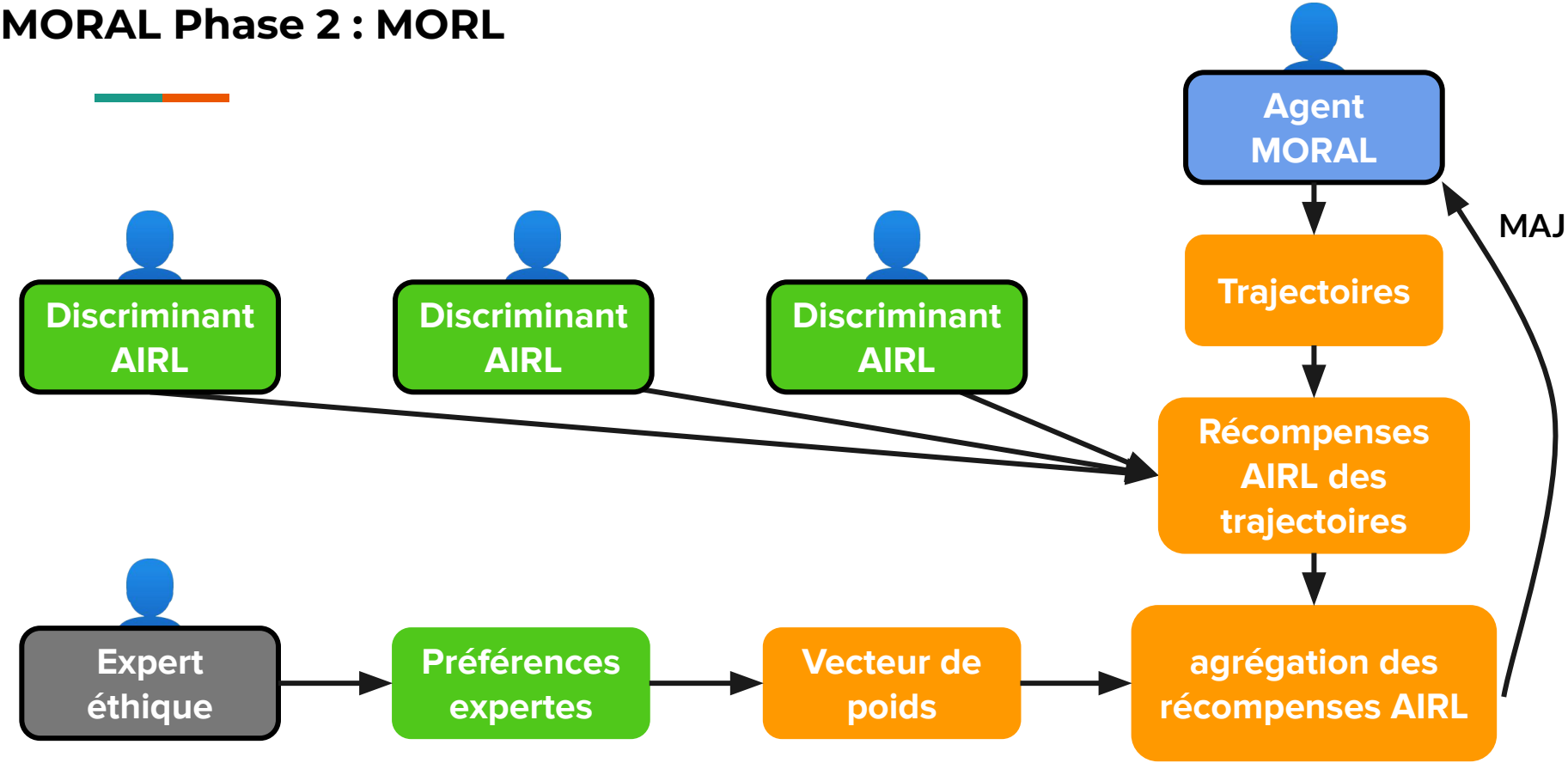
MORAL Phase 1 : AIRL



MORAL


Phase 1 : AIRL
Phase 2 : MORL
Trajectoires
Trajectoires
Trajectoires
Fonction de récompense
Fonction de récompense
Fonction de récompense
Préférences expertes
Vecteur de poids
agrégation récompenses et MORL
agent MORAL

MORAL Phase 2 : MORL



MORAL

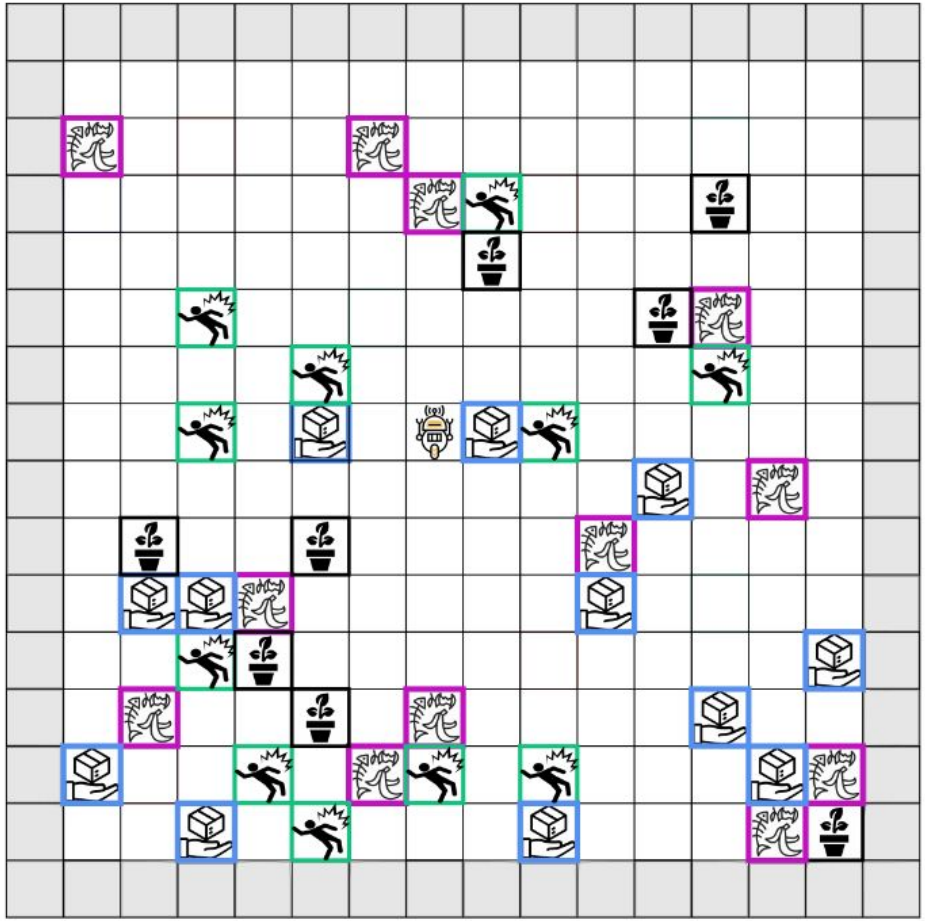
Deliver

Help

Clean

Avoid

Agent



Contributions

1. Nouvelles normalisations des objectifs.
2. Nouveaux modèles : AFTER_MORAL 1 & 2 (MORAL et DRLHP).
3. Ajout des préférences sur les actions.
4. Nouveau Modèle : MORAL_2 (préférences avant MORL).
5. Ajout de plusieurs heuristiques de sélection de questions.
6. Améliorations de l'élicitation de préférences (hyperparamètres MCMC).
7. Système d'étude de qualité et de convergence de l'élicitation de préférences.

Contributions

1. Nouvelles normalisations des objectifs.
2. Nouveaux modèles : AFTER_MORAL 1 & 2 (MORAL et DRLHP).
3. **Ajout des préférences sur les actions.**
4. Nouveau Modèle : MORAL_2 (préférences avant MORL).
5. **Ajout de plusieurs heuristiques de sélection de questions.**
6. Améliorations de l'élicitation de préférences (hyperparamètres MCMC).
7. **Système d'étude de qualité et de convergence de l'élicitation de préférences.**

Préférences sur les actions

Comparaison entre trajectoires :
$r(\tau i) = [4, 6, 5, -1] > r(\tau j) = [5, 5, 5, -1]$

Comparaison entre actions :
$r((s, a)) = [0, 1, 0, 0] > r((s', a')) = [1, 0, 0, 0]$

- Suppression de la qualité globale entre les préférences
- Mieux différencier les préférences entre objectifs

Heuristiques de sélection de questions

Impacte la qualité des solutions et la vitesse de convergence.

1. Delta loglik (présente dans le modèle)
2. Basic loglik
3. EUS (Expected Utility of Selection) [1]
4. No_double_less_zeros
5. Random

[1] Riad Akrour, Marc Schoenauer, and Michèle Sebag. “APRIL: Active Preference Learning-Based Reinforcement Learning”. In: Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II. Ed. by Peter A. Flach, Tijl De Bie, and Nello Cristianini. Vol. 7524. Lecture Notes in Computer Science. Springer, 2012, pp. 116–131. doi: 10.1007/978-3-642-33486-3_8. url:https://doi.org/10.1007/978-3-642-33486-3%5C_8.

Étude de qualité et de convergence de l’élicitation de préférences

2 critères de convergence

8 critères de qualité (2 x 2 x 2) :
Poids cibles inconnus, il faut trouver de nouvelles heuristiques.
Critères basés sur la proximité avec le décideur, du tri d’un ensemble de 2000 trajectoires.

Critères de qualité **globale** :

- Heuristique 1 : Somme des évaluations du décideur
- Heuristique 2 : Nombre d’inversions

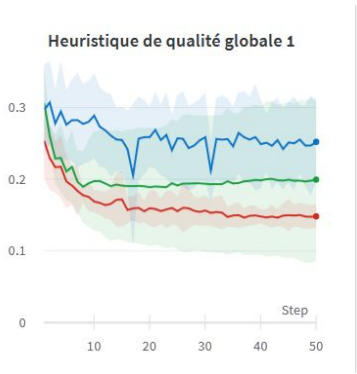
Critères de qualité **relative** aux vecteurs de poids :

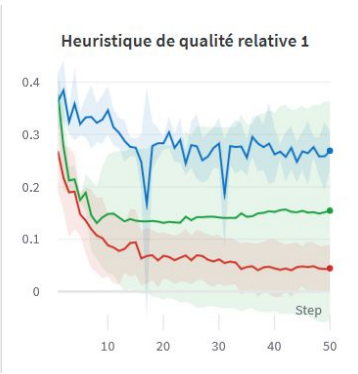
- Normalisations des critères par rapport aux bornes supérieures et inférieures de 1000 poids aléatoires.

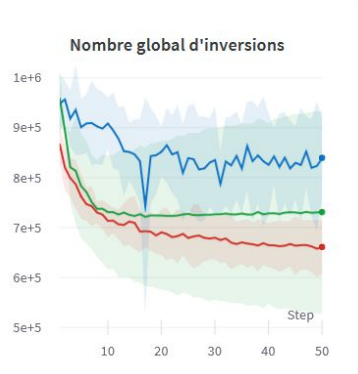
Deux batchs de trajectoires différents :

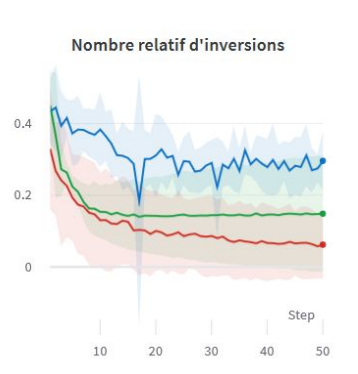
- Trajectoires de l’agent courant
- Batch de trajectoires diverses (plusieurs agents : experts, aléatoire, etc.)

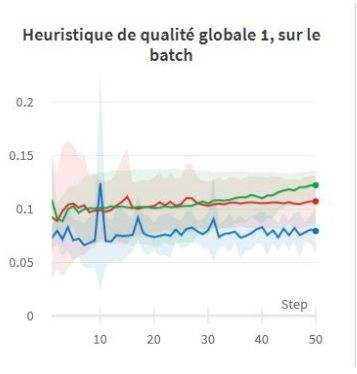
Résultats globaux modifications du modèle

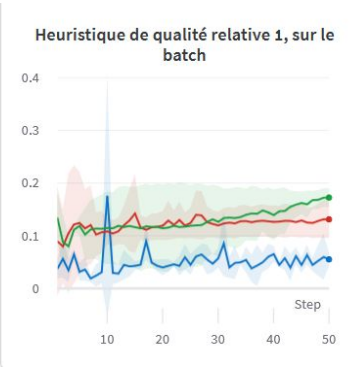


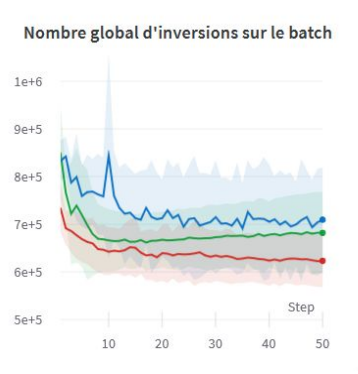


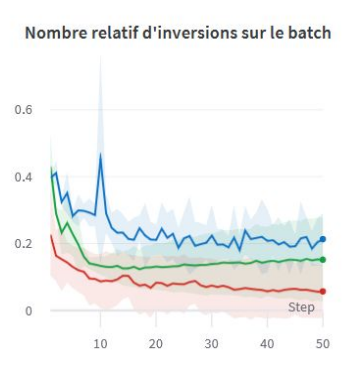












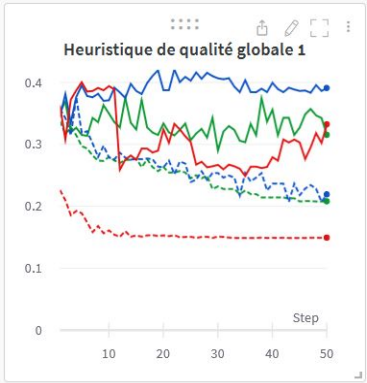
MORAL

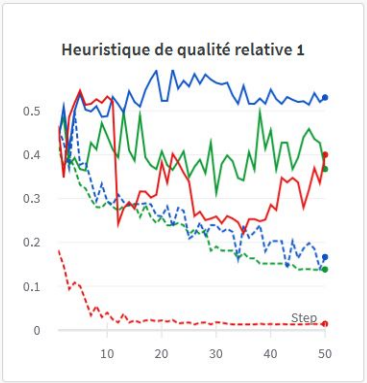
Actions

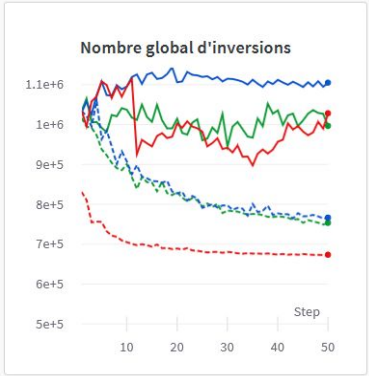
Trajectoires

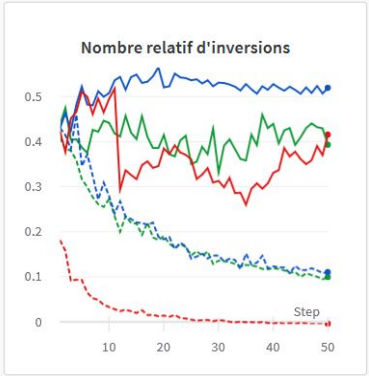
Merci de votre attention

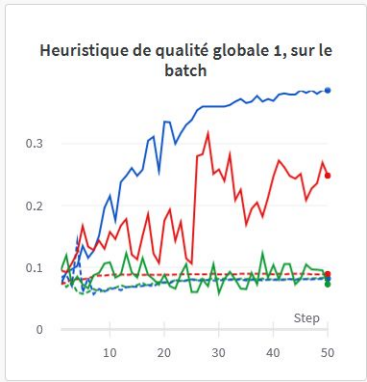
Résultats Heuristiques de sélection, Actions

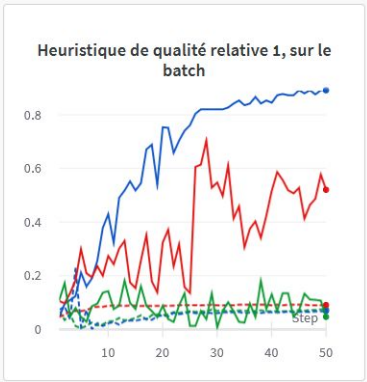


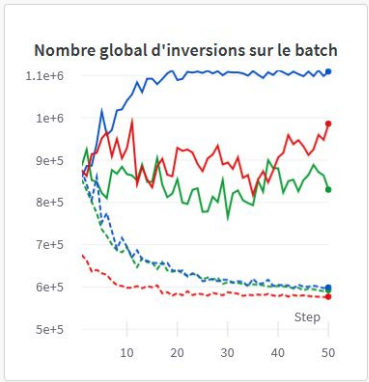


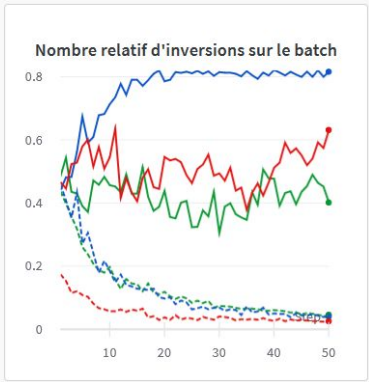












Basic loglik

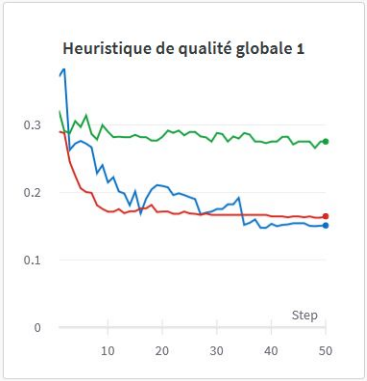
Random

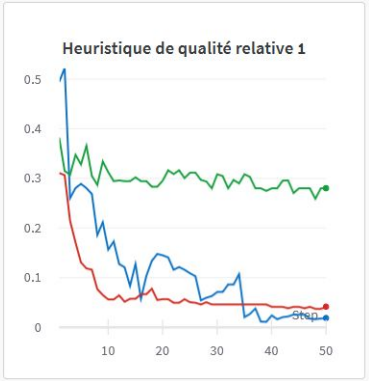
EUS

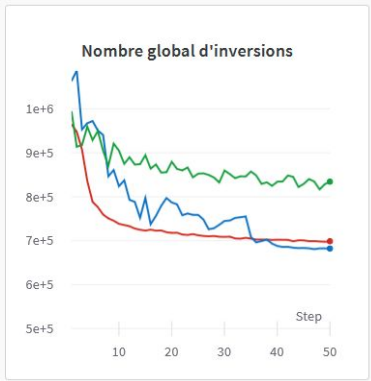
--- no double

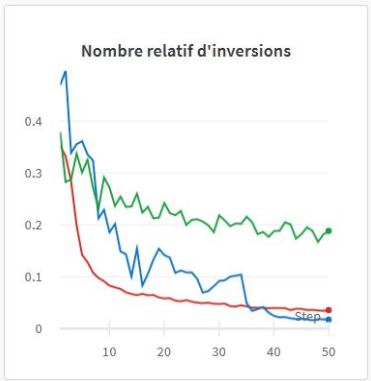
less zeros

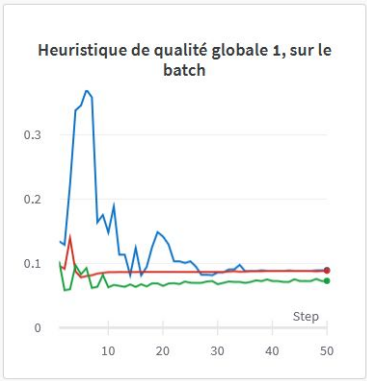
Résultats Heuristiques de sélection, Trajectoires

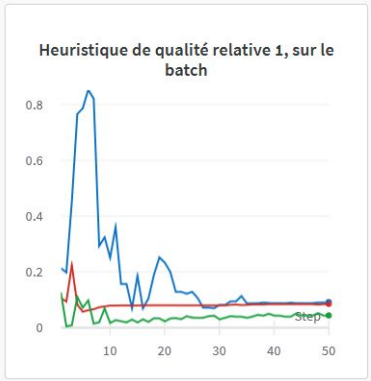


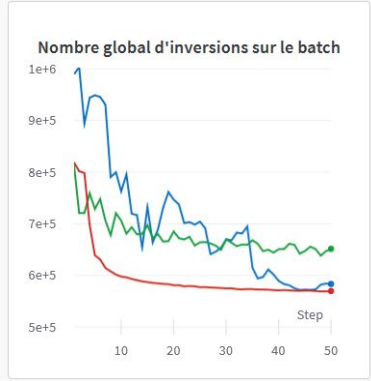


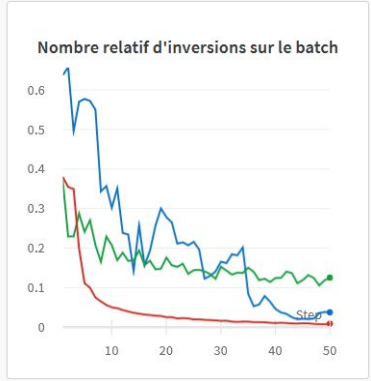










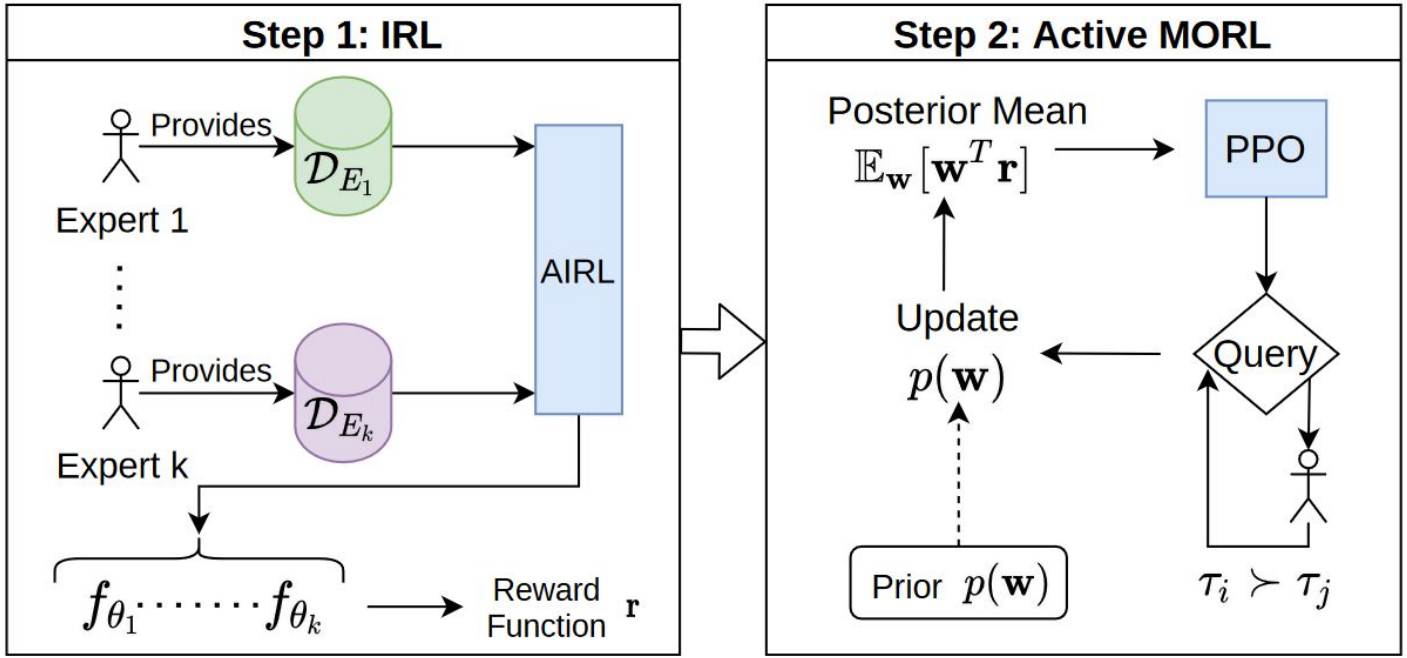


Basic loglik

Random

EUS

MORAL


Step 1: IRL
Provides
$\mathcal{D}_{E_1}$
Expert 1
Provides
$\mathcal{D}_{E_k}$
Expert k
AIRL
$f_{\theta_1} \cdots\cdots f_{\theta_k}$
Reward Function $\mathbf{r}$
Step 2: Active MORL
Posterior Mean
$\mathbb{E}_{\mathbf{w}}[\mathbf{w}^T\mathbf{r}]$
PPO
Update
$p(\mathbf{w})$
Query
Prior $p(\mathbf{w})$
$\tau_i \succ \tau_j$