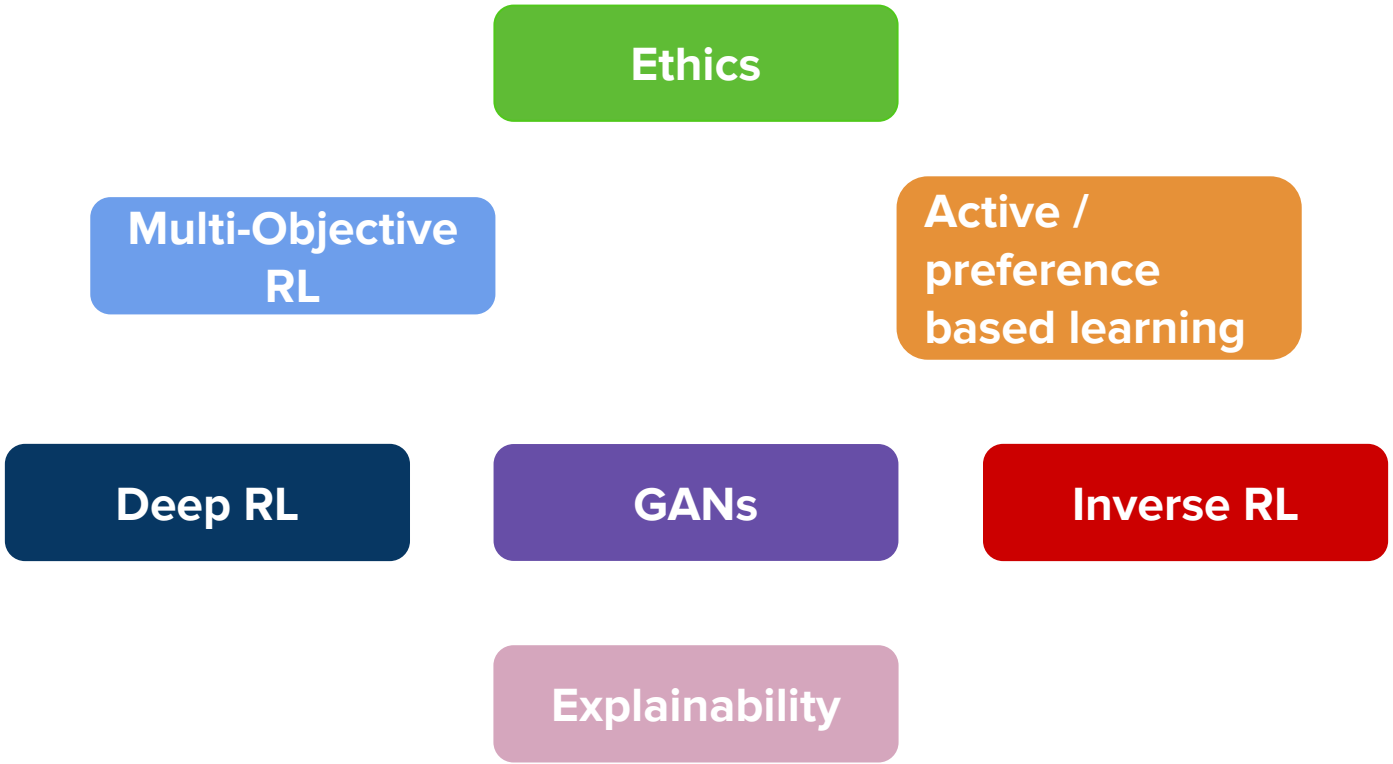
Ethical issues in multi-objective reinforcement learning

Marius Le Chapelier

Recap articles Ius



Ethics

Survey	Rule-based	Data-driven
Rossi and Mattei, 2018	Abel et al., 2016	Wu and Lin, 2017
Yu et al., 2018	Svegliato et al., 2021	Noothigattu et al., 2019
Russel et al., 2015	Rodriguez- Soto et al., 2021	Peschl et al., 2021
Eckersley, 2019	Neufeld, 2022	Glazier et al., 2022
Ventura, Gates, 2018	Ecoffet & Lehman, 2021	Hendrycks et al., 2021

Ethics

Survey

Rossi and Mattei, 2018 : discussion générale sur l'intégration de l'éthique dans l'IA

Yu et al., 2018 : ,,

Russel et al., 2015 : ,,

Ventura, Gates, 2018 : computational creativity et meta-heurisitique pour système éthique normatif

Eckersley, 2019 : impossibility and uncertainty theorems. Discussion sur le fait de valuer (ordre total) l'éthique dans l'IA et quels problèmes sont engendrés, comment les contourner : ajouter de l'incertain, travailler avec des ordres partiels.

Ethics

Rule-based

Abel et al., 2016 : résolution de POMDP éthiques créés à la main

Svegliato et al., 2021 : résolution de MDP avec 3 systèmes différents, inspirés de visions philosophiques différentes de l'éthique (DCT, PFD et VE) & expériences intéressantes

Rodriguez-Soto et al., 2021 : transforme un MOMDP en un MDP contraint éthique, en restreignant les poids de la combinaison linéaires des fonctions de récompenses du MOMDP vis-à-vis de la solution éthique optimale.

Neufeld, 2022 : RL avec un superviseur normatif (NGRL), traduction logique - MDP

Ecoffet & Lehman, 2021 : moral uncertainty & nash voting system (fait des compromis entre les différentes visions philosophiques de l'éthique : utilitarisme et déontologie)

Ethics

Data-driven

Wu and Lin, 2017 : ethics shaping dans la fonction de récompense

Noothigattu et al., 2019 : policy orchestration avec un algo multi-armed bandit

Balakrishnan, Avinash et al., 2018 : ,,

Peschl et al., 2021 : approximation de fonctions de récompenses avec de l'IRL puis calcul des meilleurs compromis avec de l'active learning / preference learning

Glazier et al., 2022 : résolution de soft constrained MDP avec de l'IRL

Hendrycks et al., 2021 : NLP éthique sur des données réelles selon plusieurs paradigmes éthiques (Justice, Virtue, Deontology, Utilitarianism, Commonsense) & création d'une BD éthique labélisée.

Active / preference based learning

Akrour et al., 2012 : Présente un modèle pour faire du DRL à partir de questions (sur les résultats des politiques) posées à un expert (PPL).

Cheng et al., 2011 : Présente 3 modèles pour faire du DRL à partir de questions (sur les actions) posées à un expert, dont certains qui permettent d'utiliser des ordres partiels et non complets (partial PBPI).

Tsochantardis et al., 2005 : comprendre les SVMs (support vector machines) (Cf Akrour)

Joachims, 2005 : comprendre les SVMs (support vector machines) (Cf Akrour)

Chrisitiano et al., 2017 : Présente un modèle de Preference-based DRL (sur les trajectoires) qui se base sur Akrour et des environnements Gym, compare également l'utilisation de données réelles et générées.

***Peschl et al., 2021 :** approximation de fonctions de récompenses avec de l'IRL puis calcul des meilleurs compromis avec de l'active learning / preference learning

Multi-Objective RL

Hayes et al., 2021 : A Practical guide to Multi-Objective Reinforcement Learning and Planning

***Noothigattu et al., 2019 :** policy orchestration avec un algo multi-armed bandit

***Rodriguez-Soto et al., 2021 :** transforme un MOMDP en un MDP contraint éthique, en restreignant les poids de la combinaison linéaires des fonctions de récompenses du MOMDP vis-à-vis de la solution éthique optimale.

***Neufeld, 2022 :** RL avec un superviseur normatif (NGRL)

***Ecoffet & Lehman, 2021 :** moral uncertainty & nash voting system (fait des compromis entre les différentes visions philosophiques de l'éthique : utilitarisme et déontologie)

***Peschl et al., 2021 :** approximation de fonctions de récompenses avec de l'IRL puis calcul des meilleurs compromis avec de l'active learning / preference learning

Inverse RL

Ziebart et al., 2008 : Maximum Entropy Inverse Reinforcement Learning

Malik et al., 2021 : Inverse Constrained Reinforcement Learning

Jarboui 2021 : Offline IRL & GANs

Finn et al., 2016 : Connection GANs et IRL

Fu et al., 2018 : Connection GANs et IRL : AIRL

***Rodriguez-Soto et al., 2021 :**

***Ecoffet & Lehman, 2021 :**

***Peschl et al., 2021 :**

***Noothigattu et al., 2019 :**

***Glazier et al., 2022 :**

GANs

Goodfellow et al., 2014 : Introduit les Generative Adversarial Nets (GANs), système qui permet d'approximer une fonction de coût/récompense en apprenant à la fois un générateur et un discriminant.

***Jarboui 2021 :** Offline IRL & GANs

***Finn et al., 2016 :** Connection GANs et IRL

***Fu et al., 2018 :** Connection GANs et IRL, Adversarial IRL

***Peschl et al., 2021 :**

Deep RL

***Akrour et al., 2012 :** Présente un modèle pour faire du DRL à partir de questions (sur les résultats des politiques) posées à un expert (PPL).

***Cheng et al., 2011 :** Présente 3 modèles pour faire du DRL à partir de questions (sur les actions) posées à un expert, dont certains qui permettent d'utiliser des ordres partiels et non complets (partial PBPI).

***Malik et al., 2021 : Inverse Constrained Reinforcement Learning**

***Hayes et al., 2021 : A Practical guide to Multi-Objective Reinforcement Learning and Planning**

***Christiano et al., 2017 :** Présente un modèle de Preference-based DRL (sur les trajectoires) qui se base sur Akrour et des environnements Gym, compare également l'utilisation de données réelles et générées.

***Peschl et al., 2021 :** approximation de fonctions de récompenses avec de l'IRL puis calcul des meilleurs compromis avec de l'active learning / preference learning

Explainability

Meyes et al., 2020 : explicabilité du Deep RL à travers une étude empirique des comportements d'agents DRL, ablations de groupes de neurones, étude de l'activation, etc.

***Ecoffet & Lehman, 2021 :** moral uncertainty & nash voting system (fait des compromis entre les différentes visions philosophiques de l'éthique : utilitarisme et déontologie)

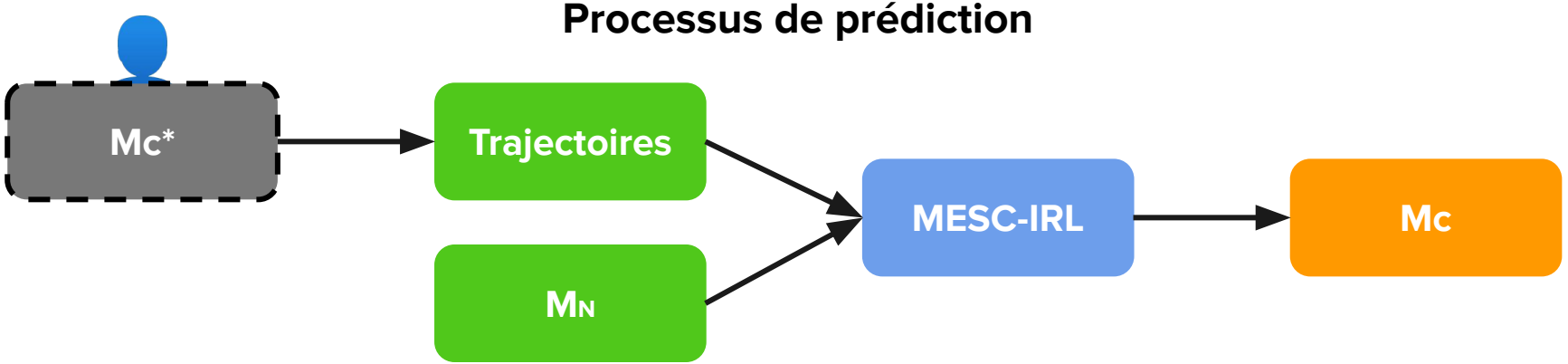
Article 1

Glazier, A., Loreggia, A., Mattei, N., Rahgooy, T., Rossi, F., & Venable, B. (2022). ***Learning Behavioral Soft Constraints from Demonstrations***.
https://doi.org/10.48550/arXiv.2202.10407

We propose and evaluate a novel method, MESC-IRL, that is able to learn both hard and soft constraints over states, actions, and state features in both deterministic and non-deterministic MDPs from a set of demonstrations. This method strictly generalizes existing methods in the literature and achieves state of the art performance in our testing in gridworld domains. Our method is also decomposable into features of the environment, which supports transferring learned constraints between environments.

Objectif :
Prédire les contraintes cachées d'un environnement à partir de trajectoires d'un agent expert, grâce à un algorithme d'IRL. Plus précisément, on veut créer un MDP contraint Mc, qui est un MDP nominal Mn auquel on a ajouté, une fonction de coût C, qui simule les pénalités des "mauvaises" actions et un budget α, définissant si les contraintes sont fortes ou faibles.

Idée générale



Détails : MESC-IRL

Définition du MDP cible Mc

Definition 1 *Given* $\mathcal{M}^{\mathcal{N}} = \langle \mathcal{S}, \mathcal{A}, P, \mu, \phi, R^{\mathcal{N}} \rangle$ *we define* soft-constrained MDP $\mathcal{M}^{\mathcal{C}} = \langle \mathcal{S}, \mathcal{A}, P, \mu, \phi, R^{\mathcal{C}} \rangle$ *where* $R^{\mathcal{C}} = R^{\mathcal{N}} - R^{\mathcal{R}}$.

$$R^{\mathcal{C}}(s_t, a_t, s_{t+1}) = \omega^{\mathcal{C}} \phi(s_t, a_t, s_{t+1})$$

$$\omega^{\mathcal{C}} = \omega^{\mathcal{N}} - \omega^{\mathcal{R}}$$

Max Entropy IRL

$$\omega^* = \underset{\omega}{\operatorname{argmax}} \sum_{\tau \in \mathcal{D}} \log P(\tau|\omega) \qquad P(\tau|\omega) \approx \frac{e^{\omega^T \phi(\tau)}}{Z(\omega)} \prod_{(s_t, a_t, s_{t+1}) \in \tau} P(s_{t+1}|s_t, a_t)$$

$$\nabla_{\omega^{\mathcal{C}}} \mathcal{L}(\mathcal{D}) = \mathbb{E}_{\mathcal{D}}[\phi(\tau)] - \sum_{(s_t, a_t, s_{t+1})} D_{s_t, a_t, s_{t+1}} \phi(s_t, a_t, s_{t+1}).$$

Le calcule des poids optimaux w* se fait par une descente de gradient.
Différence entre features extraits des trajectoires expertes et features attendues de nos poids actuels w.
Ds est une estimation de la fréquence d'utilisation des transitions, calculé avec une technique similaire au backward/forward des algos de RL.

$$\nabla_{\omega^c} \mathcal{L}(\mathcal{D}) = \mathbb{E}_{\mathcal{D}}[\phi(\tau)] - \sum_{(s_t, a_t, s_{t+1})} D_{s_t, a_t, s_{t+1}} \phi(s_t, a_t, s_{t+1}).$$

$$\nabla L(\theta) = \tilde{\mathbf{f}} - \sum_{\zeta} P(\zeta|\theta, T)\mathbf{f}_{\zeta} = \tilde{\mathbf{f}} - \sum_{s_i} D_{s_i}\mathbf{f}_{s_i} \quad (6)$$

Détails : Calcul de Ds

Idée algo :
Back/forward
des algos de RL

1. revient d'un état terminal possible
2. calcule les probabilités de chaque branche de retour possible, grâce à la fonction de partition
3. chaque branche peut retourner une probabilité locale d'action
4. calcule de l'estimation de la fréquence d'occupation pour chaque pas de temps
5. „
6. calcule de l'estimation de la fréquence d'occupation totale

Z = fonction de partition de physique statistique et thermodynamique

Algorithm 1 Expected Edge Frequency Calculation

Backward pass

1. Set $Z_{s_i,0} = 1$
2. Recursively compute for N iterations

$$Z_{a_{i,j}} = \sum_k P(s_k|s_i, a_{i,j}) e^{\text{reward}(s_i|\theta)} Z_{s_k}$$

$$Z_{s_i} = \sum_{a_{i,j}} Z_{a_{i,j}}$$

Local action probability computation

3. $P(a_{i,j}|s_i) = \dfrac{Z_{a_{i,j}}}{Z_{s_i}}$

Forward pass

4. Set $D_{s_i,t} = P(s_i = s_{\text{initial}})$
5. Recursively compute for $t = 1$ to N

$$D_{s_i,t+1} = \sum_{a_{i,j}} \sum_k D_{s_k,t} P(a_{i,j}|s_i) P(s_k|a_{i,j}, s_i)$$

Summing frequencies

6. $D_{s_i} = \sum_t D_{s_i,t}$

Détails : Transformation en probabilités pour évaluation

Probabilité

Ils estiment qu'une pénalité dépends d'une loi logistique :

$$\mathbb{C} \sim logistic(\sigma_{pooled}, \sigma_{pooled}) \qquad \sigma_{pooled} = \sqrt{(\sigma_{\mathcal{N}}^2 + \sigma_{\mathcal{C}}^2)/2}$$

Probabilité que la transition st, at, st+1 soit une contrainte en fonction de la valeur de la pénalité :

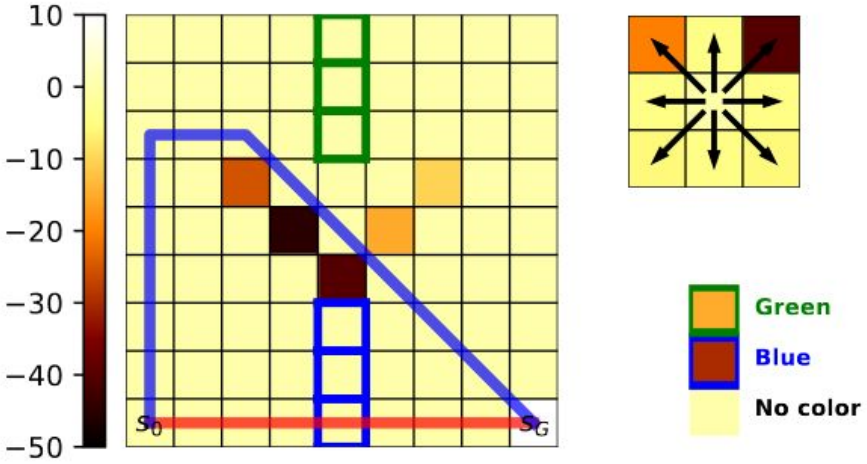
$$\zeta \equiv P\left(\mathbb{C} \leq R^R(s_t, a_t, s_{t+1})\right) = sigmoid\left(R^R(s_t, a_t, s_{t+1}) - \sigma_{pooled} / \sigma_{pooled}\right)$$

Evaluation

- Contraintes fortes :
  $\zeta > \alpha$ ⇒ contrainte prédite. (α un seuil ex : 0.6)
  Un faux positif si contrainte prédite alors que la vraie valeur ne dépasse pas le seuil.
- Contraintes faibles :
  $|\zeta^* - \zeta| < \chi$ ⇒ faux positif. Avec ζ^* la valeur cachée et χ un seuil fixé.

$$fp = \frac{\left|\left\{x \in \mathcal{C} | \mathcal{C}^*(x) = 0 \wedge (\zeta_{\mathcal{C}}(x) - \zeta_{\mathcal{C}^*}(x) > \chi)\right\}\right|}{\text{Num. Constraints}}$$

Environnement



ϕ : vecteur de features de taille 92

- 81 cases
- 8 actions possibles (directions)
- 3 couleurs

reward function du MDP nominal Mn :

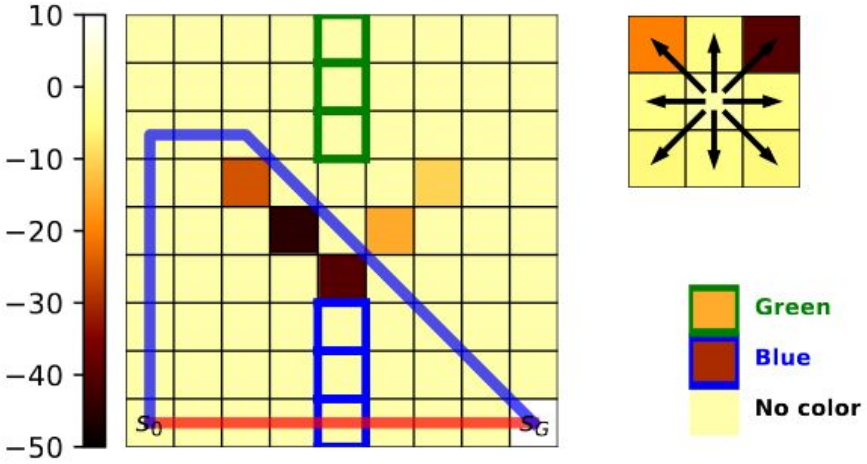
- -4 : directions cardinales
- $-4 \times \sqrt{2}$: diagonales
- +10 : état but

reward function du MDP caché Mc* :

- [-50 : 0] : contraintes (-50 = contrainte forte)

Environnement non déterministe : 10% de chance d'échec (action aléatoire)

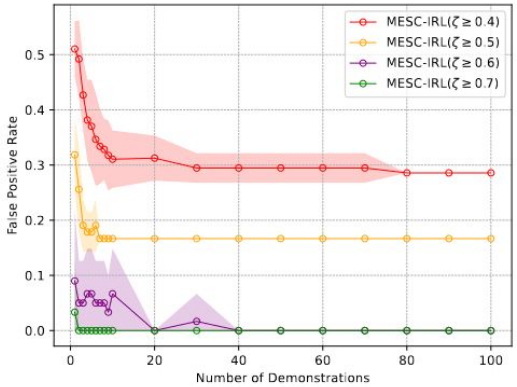
Environnement

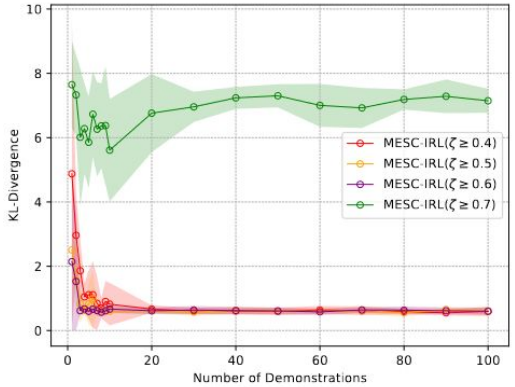


3 Types de contraintes :

- **Action Constraints** : We may not want an agent to ever perform some (set of) action ai.
- **Occupancy Constraints** : We may not want an agent to occupy a (set of) states si.
- **Feature Constraints** : Given a feature mapping transitions to a feature vector φ, we may not want an agent to perform an action (or set of actions) in the presence of specific state features.

Résultats : contraintes fortes





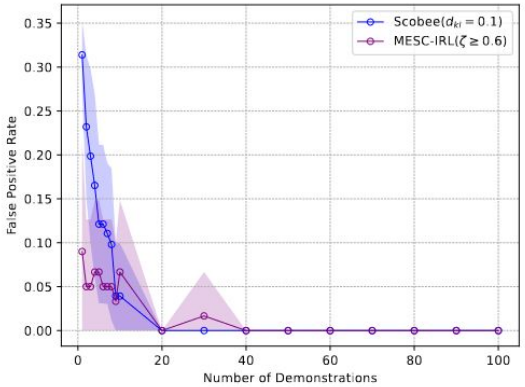


Figure 2: Performance of MESC-IRL for various settings of ζ at recovering hard constraints in a deterministic setting according to false positive rate (left) and KL-Divergence from the demonstrations $\mathcal{D}$ (center), and a comparision with the best performing method of Scobee et al. (right) as we vary the number of demonstrations. Each point is the mean of 10 independent draws.

Résultats : contraintes faibles

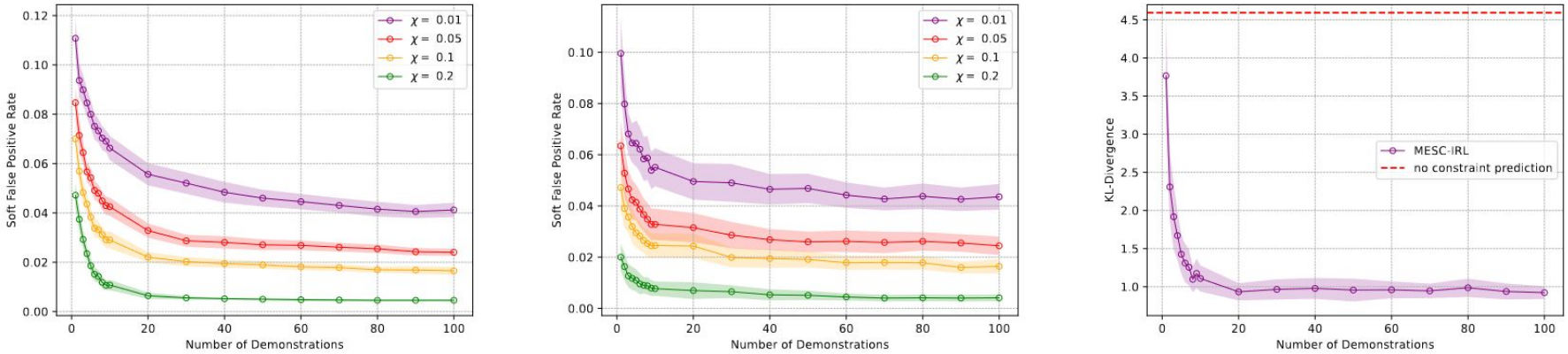
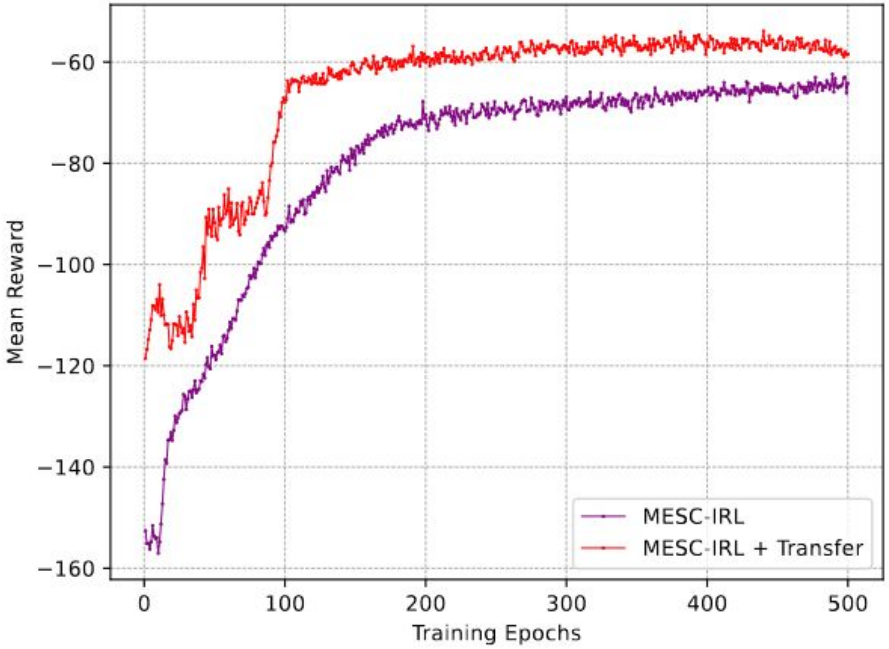
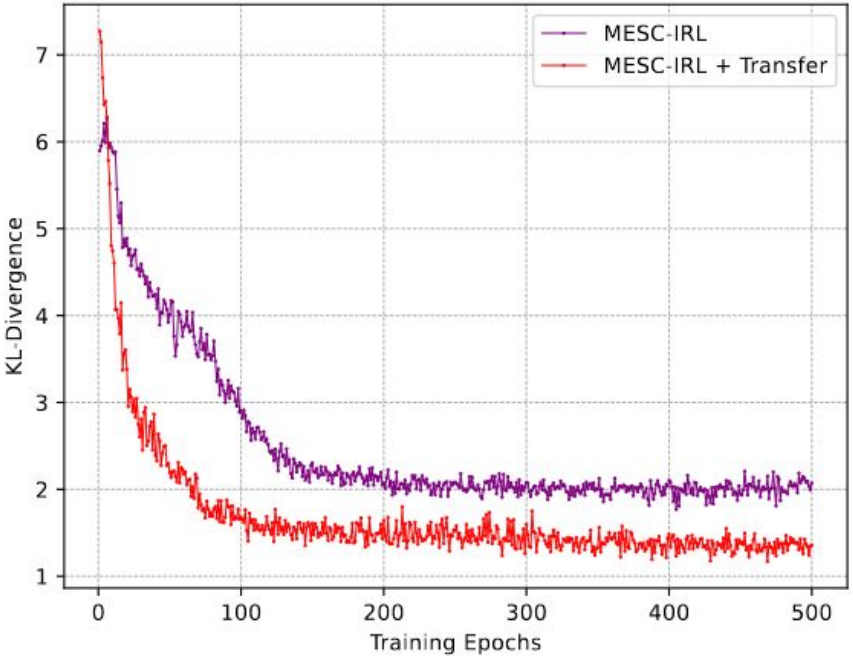


Figure 3: Performance of MESC-IRL on recovering soft constraints in deterministic settings (left) and non-deterministic settings (center) according to false positive rate as well as and KL-Divergence to $\mathcal{D}$ for the non-deterministic setting (right). We see that across all these settings we are able to accurately recover constraints and generate behavior similar to the $\mathcal{D}$ even with few demonstrations.

Résultats : Transfer learning





Inconvénients

Comment faire la différence entre une trajectoire qui a peu de récompense car elle remplit mal la tâche et une fonction de récompense qui remplit bien la tâche mais prend beaucoup de pénalités ?

Comment différencier les pénalités ? - important pour des problèmes éthiques

Comment expliquer et/ou comprendre les résultats à partir de seulement la valeur du coût d'une trajectoire ?

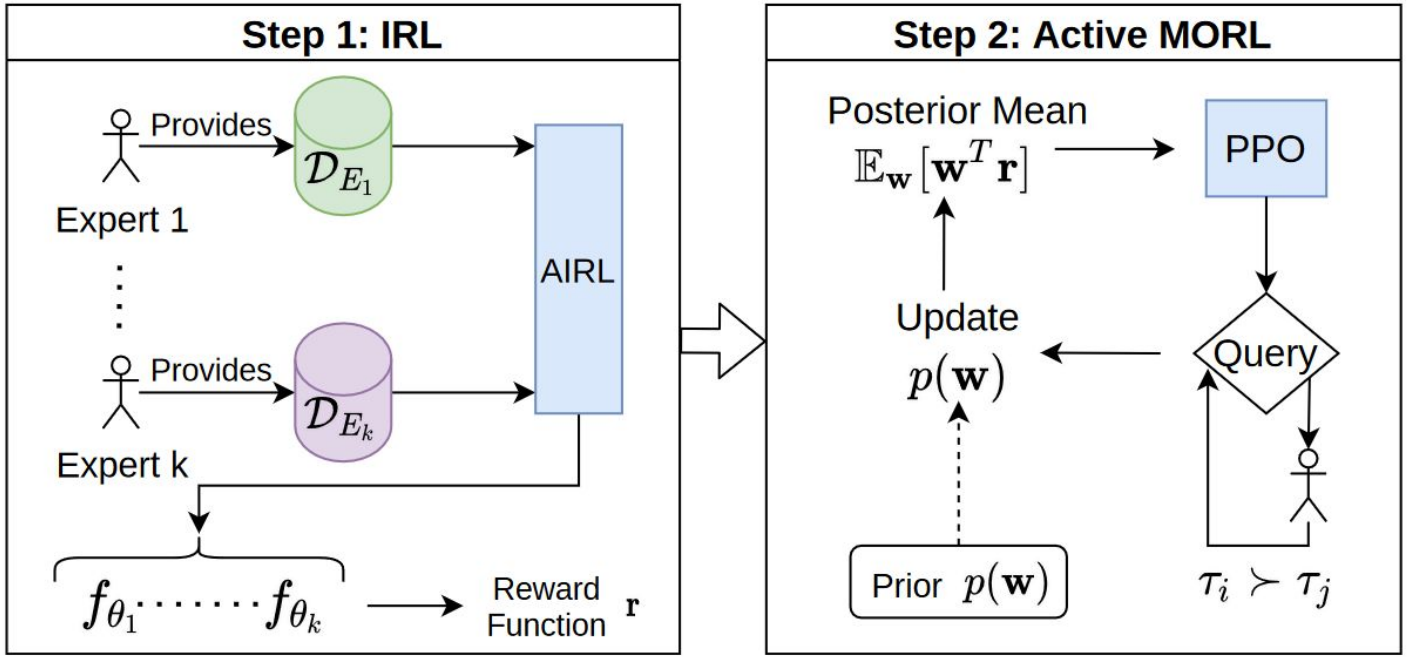
Article 2

Peschl, M., Zgonnikov, A., Oliehoek, F. A., & Siebert, L. C. (2021). ***MORAL: Aligning AI with Human Norms through Multi-Objective Reinforced Active Learning***.

We propose Multi-Objective Reinforced Active Learning (MORAL), a method that combines active preference learning and IRL to interactively learn a policy of social norms from expert demonstrations. MORAL first finds a vector-valued reward function through adversarial IRL, which is subsequently used in an interactive MORL loop. By requesting pairwise preferences over trajectories of on-policy experience from an expert, MORAL learns a probability distribution over linear combinations of reward functions under which the optimal policy most closely matches the desired behavior. We show that our approach directly approximates a Pareto-optimal solution in the space of expert reward functions, without the need of enumerating through a multitude of preference weights. Finally, we demonstrate that MORAL efficiently captures norms in two gridworld scenarios, while being able to adapt the agent's behavior to a variety of preferences.

Objectif :
Approximer les comportements d'un ou plusieurs experts avec une combinaison linéaire de reward functions, calculée avec de l'AIRL, puis du MORL interactif.

Idée générale



Etape 1 : AIRL

hϕ = approximation de la value function
gθ = approximation de la reward function
fθϕ = fonction de récompense prédite
Dθϕ = Discriminant à partir duquel on calcule π

Algorithm 1 Adversarial inverse reinforcement learning

1: Obtain expert trajectories τ_i^E
2: Initialize policy π and discriminator $D_{\theta,\phi}$.
3: **for** step t in $\{1, \ldots, \mathrm{N}\}$ **do**
4: Collect trajectories $\tau_i = (s_0, a_0, ..., s_T, a_T)$ by executing π.
5: Train $D_{\theta,\phi}$ via binary logistic regression to classify expert data τ_i^E from samples τ_i.
6: Update reward $r_{\theta,\phi}(s, a, s') \leftarrow \log D_{\theta,\phi}(s, a, s') - \log(1 - D_{\theta,\phi}(s, a, s'))$
7: Update π with respect to $r_{\theta,\phi}$ using any policy optimization method.
8: **end for**

π = PPO agent

$$D_{\theta,\phi}(s, a, s') = \frac{\exp\{f_{\theta,\phi}(s, a, s')\}}{\exp\{f_{\theta,\phi}(s, a, s')\} + \pi(a|s)}, \qquad f_{\theta,\phi}(s, a, s') = g_\theta(s, a) + \gamma h_\phi(s') - h_\phi(s).$$

Idée générale AIRL/GAN :
on entraîne un générateur (π ici) à générer des données de plus en plus fidèle à un ensemble de données cible. Et en même temps, on entraîne un discriminant (D) à classifier les données (de l'expert ou générées) avec une méthode d'entropie croisée binaire pour estimer si les données sont proches ou non des données cibles.

Etape 2 : MORAL

Objective to maximize :

$$p(\mathbf{w}|q_1, \ldots, q_n) \propto p(\mathbf{w}) \prod_{t=1}^{n} p(q_t|\mathbf{w}).$$

$$p(\tau_i \succ \tau_j|\mathbf{w}) = \frac{\exp(\mathbf{w}^T \mathbf{r}(\tau_i))}{\exp(\mathbf{w}^T \mathbf{r}(\tau_j)) + \exp(\mathbf{w}^T \mathbf{r}(\tau_i))},$$

Tous les calculs d'espérances sur les poids (Ew) se font avec MCMC (Markov Chain Monte Carlo)

La boucle k permet de trouver la comparaison qui va faire la séparation la plus dichotomique possible (proba(a>b) ~ proba(b>a)).

Algorithm 1: Multi-Objective Reinforced Active Learning

Input: Expert demonstrations $\mathcal{D}_E = \{\tau_i\}_{i=1}^N$, prior $p(\mathbf{w})$.
Initialize: Reward function $\mathbf{r} = (f_{\theta_1}, \ldots, f_{\theta_k})$ by running AIRL on $\mathcal{D}_E$, PPO agent π_ϕ.
for $n = 0, 1, 2, \ldots$ **do**
  Approximate $p(\mathbf{w}|q_1, \ldots, q_n)$ through MCMC.
  Get mean reward function $r \leftarrow \mathbb{E}_{\mathbf{w}}[\mathbf{w}^T \mathbf{r}]$.
  $volume \leftarrow -\infty$
  for $k = 0, 1, 2, \ldots, N$ **do**
    Sample trajectories $\mathcal{D} = \{\tau_i\}_{i=1}^m$ using π_ϕ.
    Update ϕ using PPO to maximize $\mathbb{E}_{\pi_\phi}\left[\sum_{t=0}^T \gamma^t r(s_t, a_t)\right]$.
    Sample a pair of trajectories (τ_i, τ_j) from $\mathcal{D}$.
    $next\_volume \leftarrow \min(\mathbb{E}_{\mathbf{w}}[1 - p(\tau_i \succ \tau_j|\mathbf{w})], \mathbb{E}_{\mathbf{w}}[1 - p(\tau_j \succ \tau_i|\mathbf{w})])$.
    if $next\_volume > volume$ **then**
      $next\_query \leftarrow (\tau_i, \tau_j)$
      $volume \leftarrow next\_volume$
  Query expert using $next\_query$ and save answer q_n.

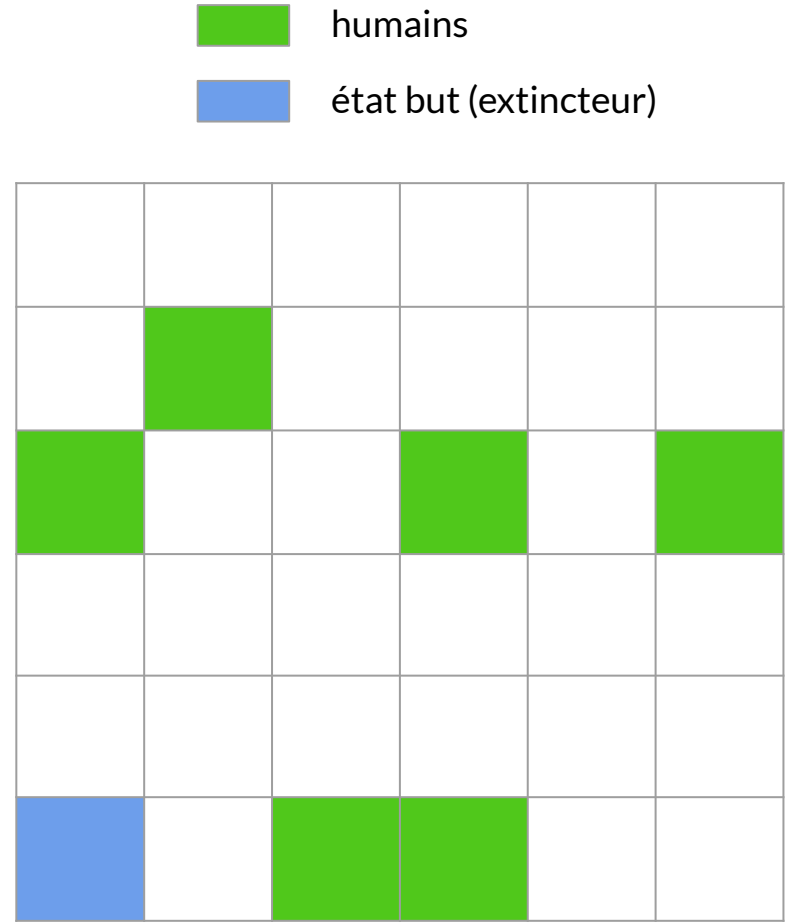
1ère Expérience : Emergency

Objectif :
Cas simple d'optimisation bi-objectif

Environnement :
T = 75
grille 6x6

AIRL :
input : 50 trajectoires d'un agent PPO maximisant le nombre d'humains sauvés.
output : fθ, reward function estimée

MORL :
input : r = (rp, fθ), avec rp la reward function de l'état but.
process : 25 question posées à expert (maximisant fθ avant rp).
output : w, un vecteur de poids qui maximise p(w|qi) (qi = questions)



1ère Expérience : Emergency

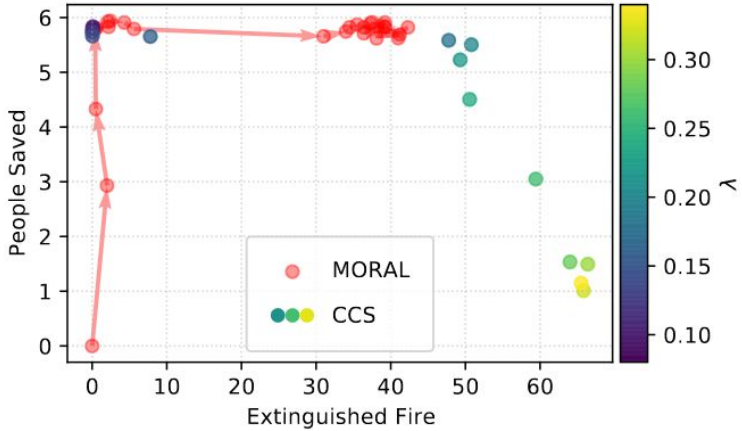


Figure 2: Intermediate policies found during MORAL in the *Emergency* domain, compared to a manually computed CCS. MORAL approximates a Pareto-optimal solution that most closely matches the given preferences.

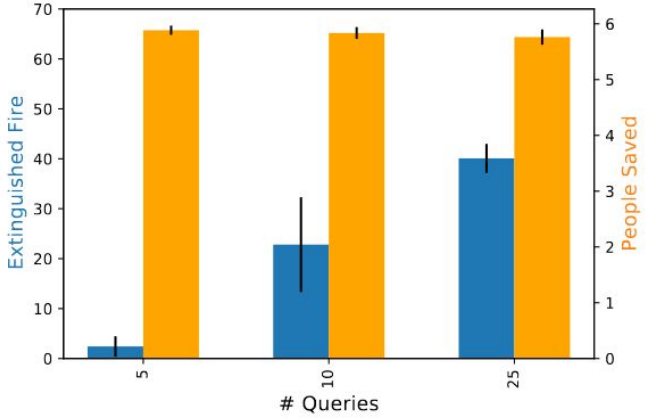
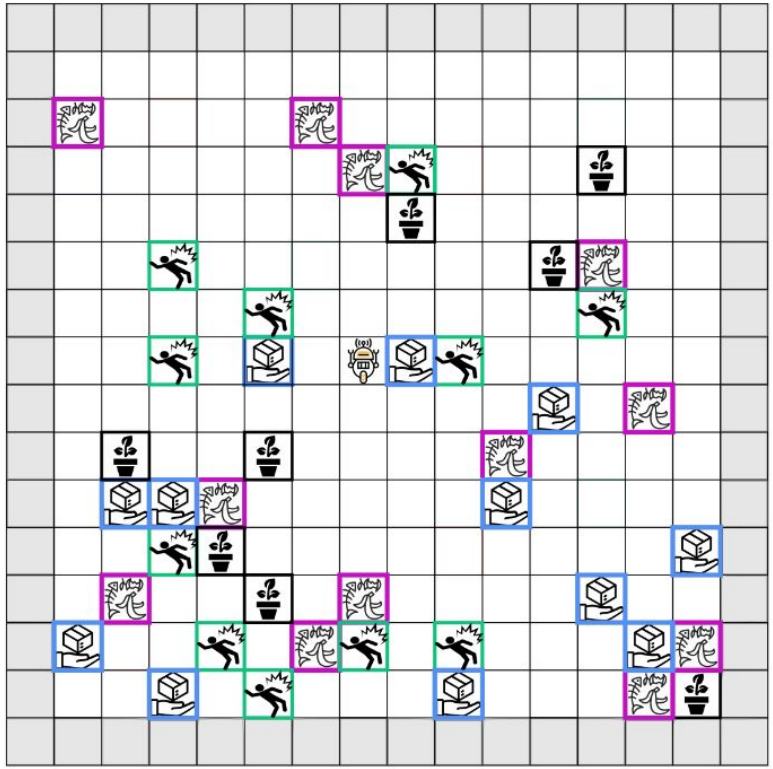


Figure 3: Query efficiency of MORAL for finding a trade-off that matches the given preferences. Averaged over three random seeds.

2ème Expérience : Delivery

Objectif :

Montrer que même si r est composé de fonctions de reward contradictoires, on peut estimer avec précision la reward fonction de l'expert en trouvant un jeu de poids d'agrégation de ces fonctions.
C'est à dire prouver que les résultats du MORL sont robustes qu'importe les résultats de l'AIRL.

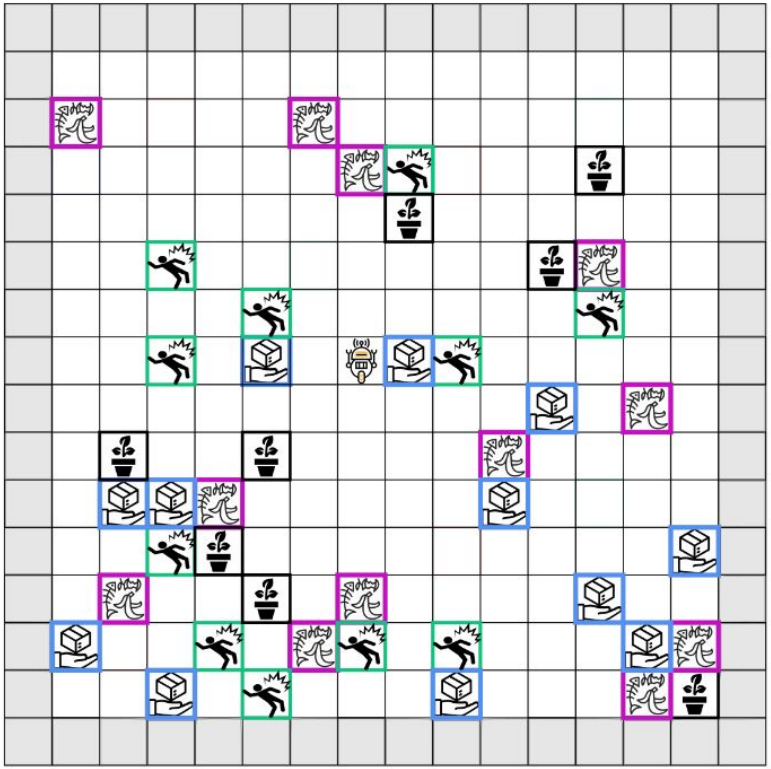


2ème Expérience : Delivery

Environnement :
T = 50
grille 16x16

AIRL :
input : 50 trajectoires de deux agents PPO maximisant respectivement le nombre de gens aidés et de tuiles nettoyées, et tous les deux le nombre de vases évités.
output : fθ1, fθ2, reward functions estimées

MORL :
input : r = (rp, fθ1, fθ2), avec rp la reward function correspondante aux livraisons.
process : 25 questions posées à un expert.
output : w, un vecteur de poids qui maximise p(w|qi) (qi = queries)



2ème Expérience : Delivery

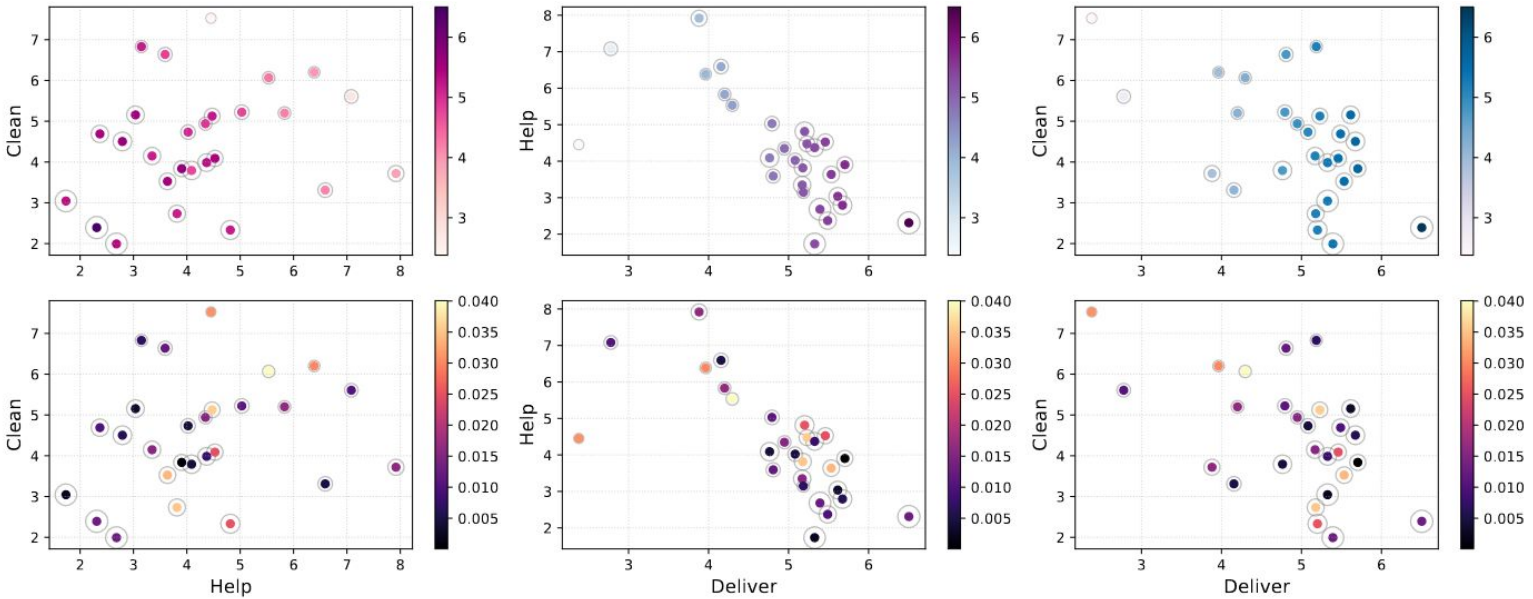


Figure 5: The convex coverage set found by MORAL for three reward dimensions. We plot two-dimensional projections of the attained explicit objectives, with colors indicating the third objective (*top three panels*). The colors in the bottom three panels show the deviation (8) to the respective preference vector m used during training. Gray circles around each policy indicate the relative amount of broken vases.

Inconvénients/Questions sur l'article

Questionnement sur l'explicabilité réelle avec du DRL pendant les deux étapes du processus.

Pourquoi ne pas avoir fait un 4ème objectif avec le nombre de vases brisés dans l'expérience Delivery ?

Dans la 2ème expérience, on a cherché des solutions satisfaisantes pour chaque préférences possibles de l'expert (exploration de tout l'espace). Dans la réalité, comment avoir cet expert éthique qui donne ses préférences entre les trajectoires ?

Peut-on aller plus loin que la recherche d'une solution pareto optimale et comparer toutes ces solutions pour les classer ? Métaheuristique éthique, système de vote entre les n experts pour leurs préférences parmi les solutions pareto optimales etc.

Article 3

Ecoffet, A., & Lehman, J. (2021). **Reinforcement Learning Under Moral Uncertainty.**

Problématique : Comment faire du MORL quand les résultats des fonctions de reward ne sont pas comparables. Par exemple lorsqu'elles représentent des paradigmes éthiques.

Idée : Calcul de compromis entre les différentes visions philosophiques de l'éthique : utilitarisme et déontologie avec un vote de nash entre les différentes théories. Un agent par théorie, son poids de vote est défini par la "credence", la confiance que l'on a en la théorie.

Article 3 : Ecoffet & Lehman, 2021

A chaque pas de temps, chaque théorie vote pour ou contre chaque action possible, l'action avec le plus de votes (indexés sur la credence de chaque théorie) est exécutée.

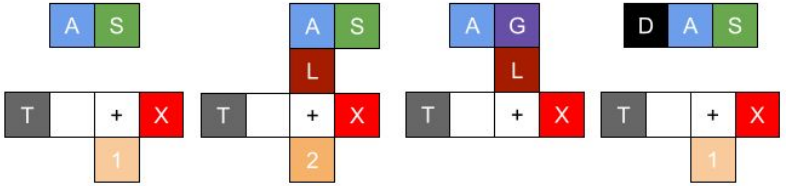
Le vote de Nash possède 2 grands désavantages qu'ils ont appelé "Stakes Insensitivity" (augmenter les enjeux pour une théorie ne va pas augmenter son poids dans le vote) et "No Compromise" (si une action n'est la préférée d'aucune action, elle ne peut être choisie, même si c'est un bon compromis). Ces désavantages proviennent de possibles votes "tactiques" et ils ont décidé d'utiliser un système de vote qui oblige les théories à voter selon leurs vraies préférences : le variance voting

$$\mu_i(s) = \frac{1}{k} \sum_a Q_i(s,a)$$

$$\sigma_i^2 = \mathbf{E}_{s \sim S} \left[\frac{1}{k} \sum_a \left(Q_i(s,a) - \mu_i(s) \right)^2 \right]$$

$$\pi(s) = \arg\max_a \sum_i C_i \frac{Q_i(s,a) - \mu_i(s)}{\sqrt{\sigma_i^2 + \varepsilon}}$$

Article 3 : Ecoffet & Lehman, 2021

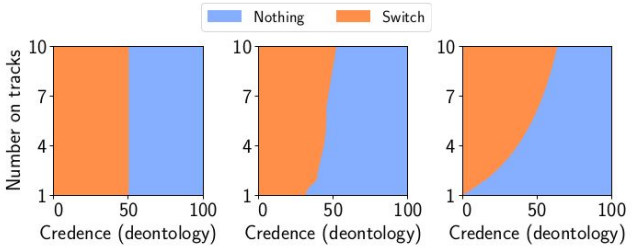


(a) Classic (b) Double (c) Guard (d) Doomsday

Stakes Insensitivity

	Crash into 1	Crash into X
Utilitarianism	-1	-X
Deontology	-1	0

(a) Preferences in the classic trolley problem.

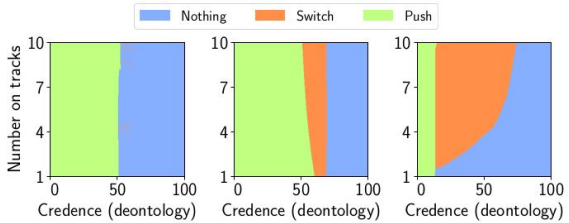


(b) Nash voting (c) Iterated Nash voting (d) Variance voting

No Compromise

	Push L	Crash into 2	Crash into X
Util.	-1	-2	-X
Deont.	-4	-1	0
Altered Deont.	-1	-4	0

(a) Preferences for the double trolley problem. Altered Deont. only used for Nash voting with unknown adversary.

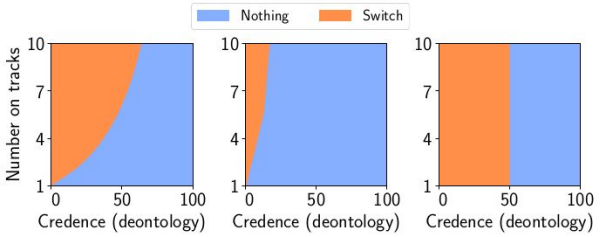


(b) Nash voting (c) Nash voting (unknown adversary) (d) Variance voting

Independence of irrelevant alternatives

	Crash into 1	Crash into X	Doomsday
Util.	-1	-X	-300
Deont.	-1	0	-10

(a) Preferences in the doomsday trolley problem.



(b) Variance voting (non-doomsday) (c) Variance voting (doomsday) (d) Nash voting (both)

Article 4

Eckersley, P. (2019). **Impossibility and Uncertainty Theorems in AI Value Alignment (or why your AGI should not have a utility function).**

Objectif :
Discussion autour de l'utilisation de fonction d'utilité pour l'apprentissage éthique et des contradiction rencontrées avec l'étude de plusieurs théorèmes d'impossibilité et d'incertitude. Présentation de systèmes permettant de faire de l'apprentissage sans aller à l'encontre de ces théorèmes, notamment en introduisant de l'incertitude dans nos choix ou en raisonnant sur des ordres partiels.

Article 4 : Eckersley, P. (2019)

Arrhenius's unpublished book contains a collection of many more uncertainty theorems. Here is the simplest, which is a more compelling version of Parfitt's Mere Addition Paradox. It shows the impossibility of an objective function satisfying the following requirements simultaneously:

The Quality Condition: There is at least one perfectly equal population with very high welfare which is at least as good as any population with very low positive welfare, other things being equal.

The Inequality Aversion Condition: For any triplet of welfare levels A, B, and C, A higher than B, and B higher than C, and for any population A with welfare A, there is a larger population C with welfare C such that a perfectly equal population B of the same size as A∪C and with welfare B is at least as good as A∪C, other things being equal.

The Egalitarian Dominance Condition: If population A is a perfectly equal population of the same size as population B, and every person in A has higher welfare than every person in B, then A is better than B, other things being equal.

The Dominance Addition Condition: An addition of lives with positive welfare and an increase in the welfare in the rest of the population doesn't make a population worse, other things being equal.

Article 4: Eckersley, P. (2019)

has many of these impossibility results. For example, Arrhenius [2] shows that all total orderings of populations must entail one of the following six problematic conclusions, stated informally:

The Repugnant Conclusion For any population of very happy people, there exists a much larger population with lives barely worth living that is better than this very happy population (this affects the "maximise total wellbeing" objective).

The Sadistic Conclusion Suppose we start with a population of very happy people. For any proposed addition of a sufficiently large number of people with positive welfare, there is a small number of horribly tortured people that is a preferable addition.[4]

The Very Anti-Egalitarian Conclusion For any population of two or more people which has uniform happiness, there exists another population of the same size which has lower total and average happiness, and is less equal, but is better.

Anti-Dominance Population B can be better than population A even if A is the same size as population B, and every person in A is happier than their equivalent in B.

Anti-Addition It is sometimes bad to add a group of people B to a population A (where the people in group B are worse of than those in A), but *better* to add a group C that is larger than B, and worse off than B.

Extreme Priority There is no n such that create of n lives of very high positive welfare is sufficient benefit to compensate for the reduction from very low positive welfare to slightly negative welfare for a single person (informally, “the needs of the few outweigh the needs of the many”).

Article 4 : Eckersley, P. (2019)

Les solutions étudiées pour contourner les théorèmes sont les suivantes :
Small-scale evasion : On considère que les tâches des IAs ont des enjeux trop faibles pour que ce soit un problème à grande échelle.
Value learning : On considère que le problème vient du fait de vouloir définir explicitement une fonction d'utilité et que l'on peut contourner ce problème (ex : human-guidance) mais il faut dans ce cas raisonner avec un autres système.
Theory normalization : Essayer de théoriser des façons de faire des compromis entre des objectifs de bien être social (ex : bien être total, moyen et la réduction de la souffrance). L'idée d'un vote entre ces objectif est évoqué. Mais la manière dont les votes sont combinés pourrait entrer en conflit avec un des axiomes du théorème.
Accept one of the axioms : On pourrait, dans un domaine précis, considérer un axiome acceptable et choisir de l'accepter. Cette solution n'est pas applicable de manière générique.
Treat impossibility results as uncertainty results : Dans des domaines à grands enjeux, il est plus approprié d'introduire de l'incertitude dans les fonctions objectifs. En considérant des ordres partiels ou en introduisant directement des probabilités de confiance en les objectifs.

Article 5 :

Cheng, W., Fürnkranz, J., Hüllermeier, E., & Park, S. (2011). **Preference-Based Policy Iteration: Leveraging Preference Learning for Reinforcement Learning.**

Idée : présente 3 modèles pour faire du DRL à partir de questions (sur les actions) posées à un expert, dont certains qui permettent d'utiliser des ordres partiels et non complets (partial PBPI).

Algorithm 2. Multi-class variant of Approx. Policy Iteration with Roll-Outs [11]

Require: generative environment model E, sample states S, discount factor γ, initial (random) policy π_0, number of trajectories/roll-outs K, max. length/horizon of each trajectory T, max number of policy iterations p

$\pi' \leftarrow \pi_0$
repeat
  $\pi \leftarrow \pi'$, $\mathcal{T} \leftarrow \emptyset$
  for each $\boldsymbol{s} \in S$ **do**
    for each $\mathbf{a} \in A$ **do**
      $(\boldsymbol{s}', r) \leftarrow \text{SIMULATE}(E, \boldsymbol{s}, \mathbf{a})$ # do (possibly off-policy) action $\mathbf{a}$
      $\tilde{Q}^{\pi}(\boldsymbol{s}, \mathbf{a}) \leftarrow \text{ROLLOUT}(E, \boldsymbol{s}', \gamma, \pi, K, T) + r$ # estimate state-action value
    end for
    $\mathbf{a}^* \leftarrow \arg\max_{\mathbf{a} \in A} \tilde{Q}^{\pi}(\boldsymbol{s}, \mathbf{a})$
    if $\tilde{Q}^{\pi}(\boldsymbol{s}, \mathbf{a}^*) >_T \tilde{Q}^{\pi}(\boldsymbol{s}, \mathbf{a})$ for all $\mathbf{a} \in A, \mathbf{a} \neq \mathbf{a}^*$ **then**
      $\mathcal{T} \leftarrow \mathcal{T} \cup \{(\boldsymbol{s}, \mathbf{a}^*)\}$
    end if
  end for
  $\pi' \leftarrow \text{LEARN}(\mathcal{T})$
until $\text{STOPPINGCRITERION}(E, \pi, \pi', p)$

Approximate Policy Iteration (API) generates one training example $(\boldsymbol{s}, \mathbf{a}^*)$ if $\mathbf{a}^*$ is the best available action in $\boldsymbol{s}$, i.e., if $\tilde{Q}^{\pi}(s, \mathbf{a}^*) >_T \tilde{Q}^{\pi}(s, \mathbf{a})$ for all $\mathbf{a} \neq \mathbf{a}^*$. If there is no action that is better than all alternatives, no training example is generated for this state.

Pairwise Approximate Policy Iteration (PAPI) works in the same way as API, but the underlying base learning algorithm is replaced with a label ranker. This means that each training example $(\boldsymbol{s}, \mathbf{a}^*)$ of API is transformed into $a - 1$ training examples of the form $(\boldsymbol{s}, \mathbf{a}^* \succ \mathbf{a})$ for all $\mathbf{a} \neq \mathbf{a}^*$.

Preference-Based Policy Iteration (PBPI) is trained on all available pairwise preferences, not only those involving the best action. Thus, whenever $\tilde{Q}^{\pi}(s, \mathbf{a}_k) >_T \tilde{Q}^{\pi}(s, \mathbf{a}_l)$ holds for a pair of actions $(\mathbf{a}_k, \mathbf{a}_l)$, PBPI generates a corresponding training example $(\boldsymbol{s}, \mathbf{a}_k \succ \mathbf{a}_l)$. Note that, contrary to PAPI, $\mathbf{a}_k$ does not need to be the best action. In particular, it is not necessary that there is a clear best action in order to generate training examples. Thus, from the same roll-outs, PBPI will typically generate more training information than PAPI or API.

Article 5 : Cheng et al., 2011

PBPI : ordre total, toutes les actions sont testées par état
PBPI-1 : ordre partiel, même nb états que PBPI (même nb états)
PBPI-2 : ordre partiel, k/2 * nb états de PBPI (même nombre de roll-outs)
PBPI-3 : ordre partiel, k(k-&)/2 * nb états de PBPI (même nombre de préférences)

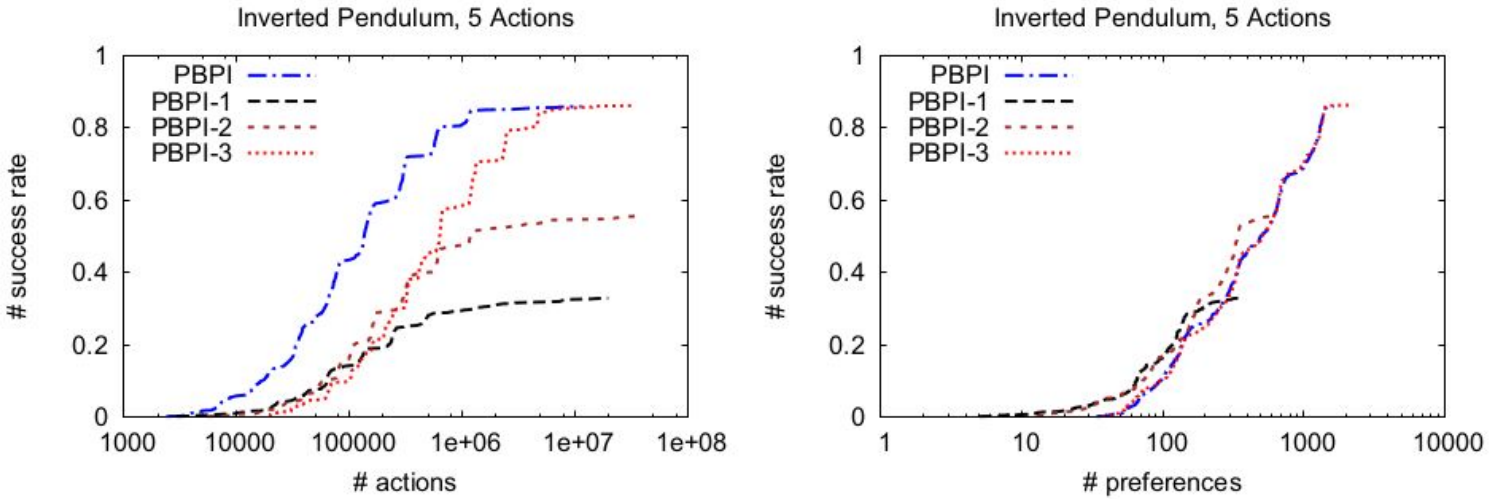


Fig. 2. Comparison of complete state evaluation (PBPI) with partial state evaluation in three variants (PBPI-1, PBPI-2, PBPI-3)

Article 6 : Meyes, R., Schneider, M., & Meisen, T. (2020). **How Do You Act? An Empirical Study to Understand Behavior of Deep Reinforcement Learning Agents.**

Idée : explicabilité du Deep RL à travers une étude empirique des comportements d'agents DRL, ablations de groupes de neurones, étude de l'activation, etc.

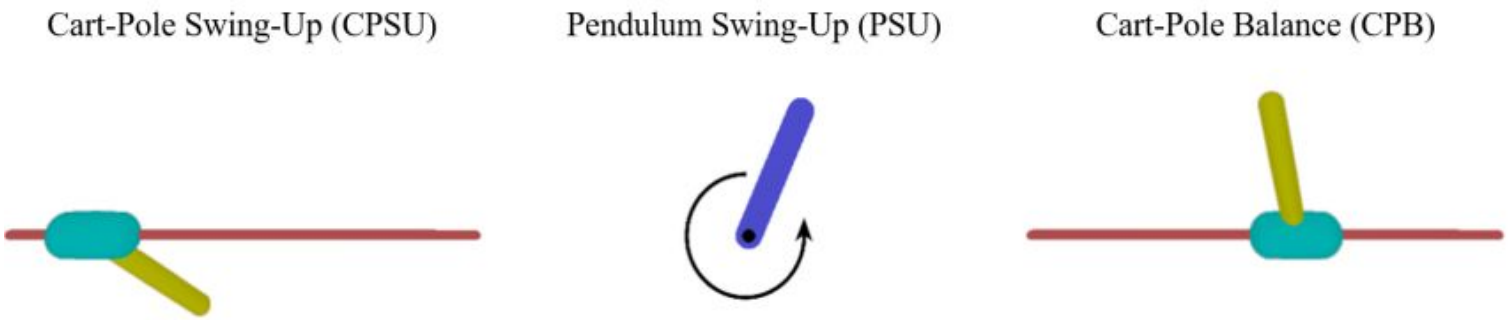


Fig. 1: Three exemplary rendered images of the respective control environments.

Article 6 : Meyes et al., 2020

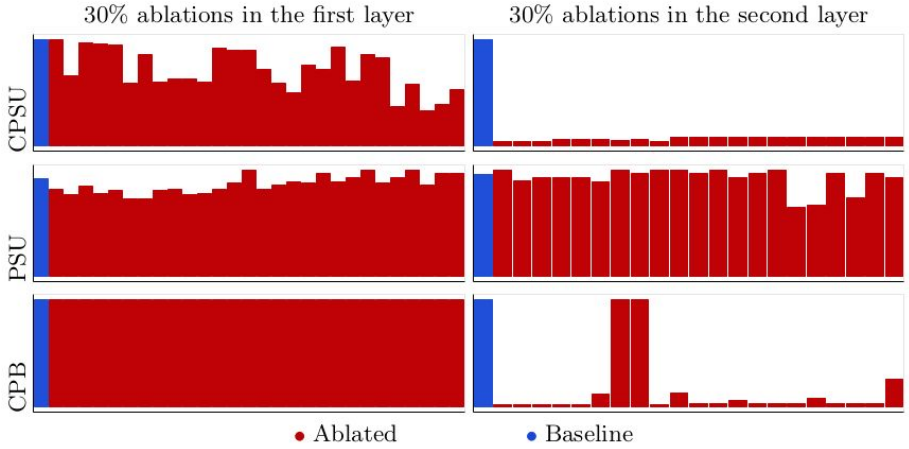


Fig. 2: Comparison of the normalized returns achieved as a result of ablations of 30% of the units (red bars) in to its respective baselines (blue bars).

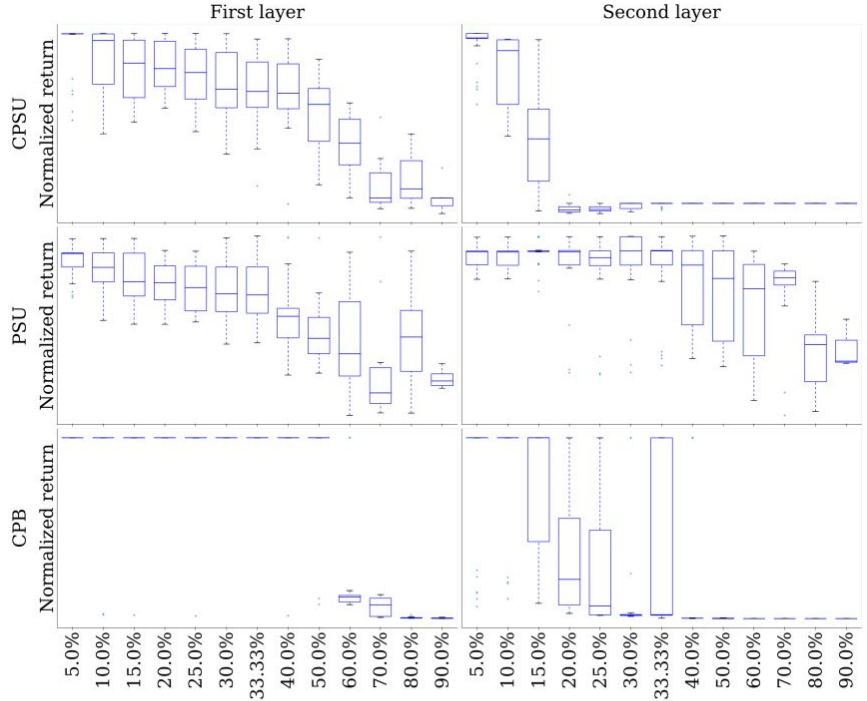


Fig. 3: Distributions of the normalized returns for all ablations performed in the first layer (left side) and second layer (right side).

Article 6 : Meyes et al., 2020

Pearson correlation coefficient : une haute valeur indique la contribution exclusive d'une unit à une action donnée (une direction)

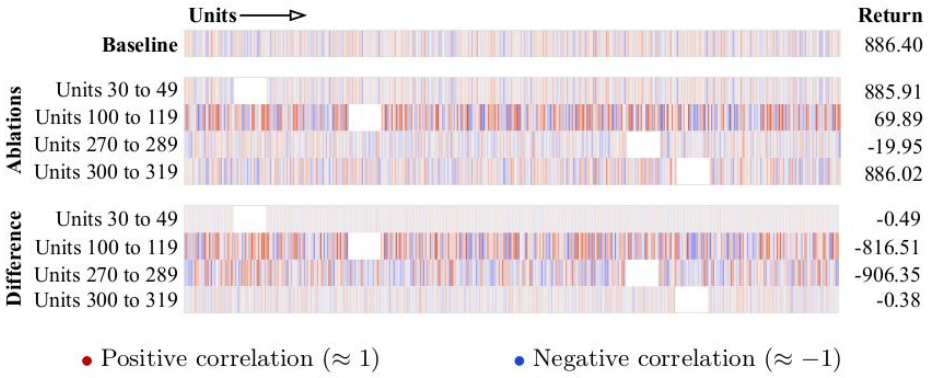


Fig. 4: Correlation pattern of the activations of all 400 units in the first layer during the CPSU task for the healthy agent (baseline) and four exemplary ablations, as well as the change of these patterns compared to the baseline (bottom four rows).

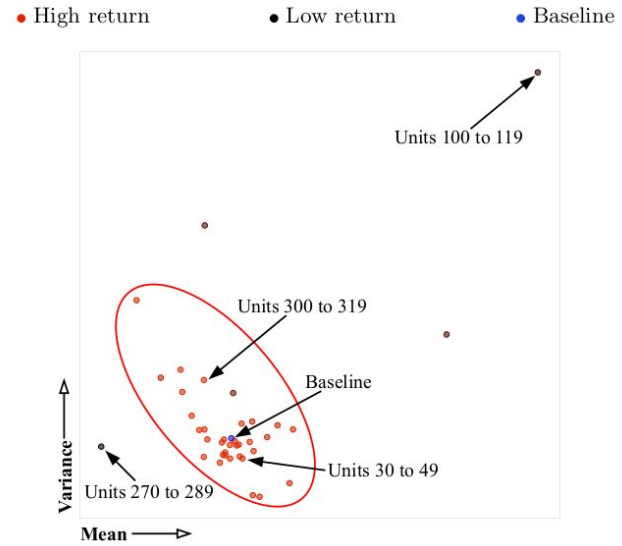
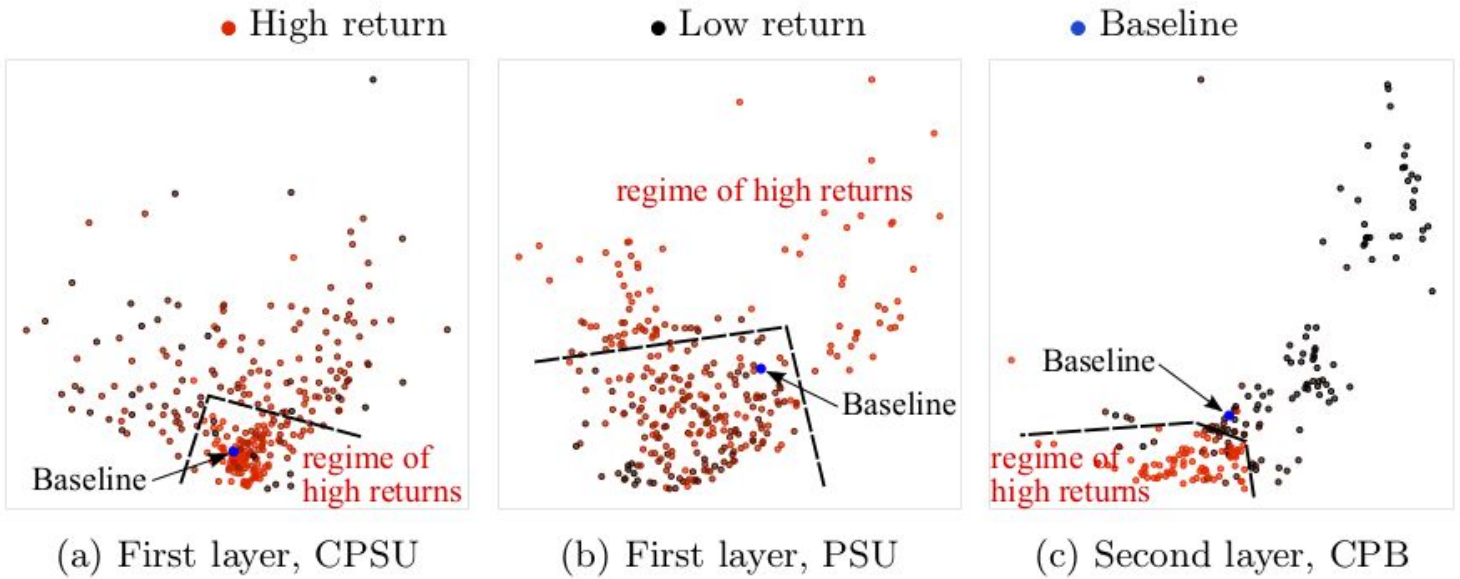


Fig. 5: Scatter plot of the mean and the variance of the correlation patterns for the baseline and all 29 ablations of the size of 5% and their corresponding returns in the CPSU task.

Article 6 : Meyes et al., 2020

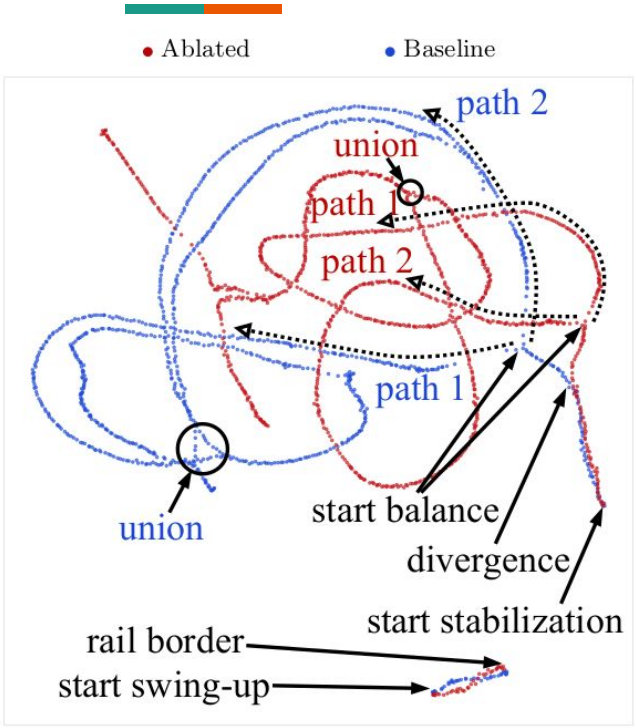


(a) First layer, CPSU (b) First layer, PSU (c) Second layer, CPB

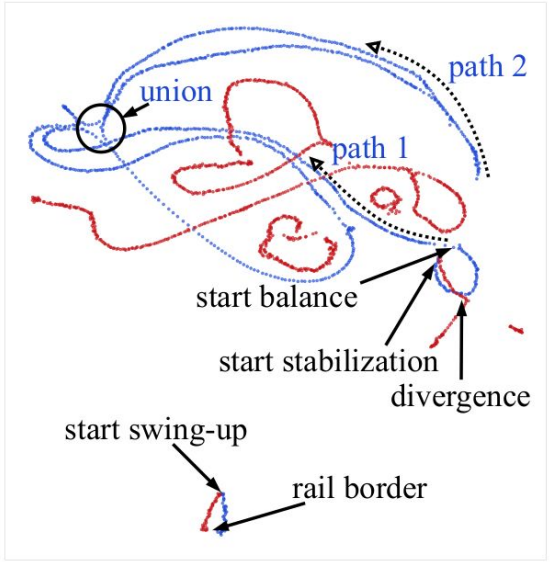
Fig. 6: Scatterplot showing the mean (x-axis) and variance (y-axis) of the correlation coefficients for all ablations of the specified layer.

Article 6 : Meyes et al., 2020

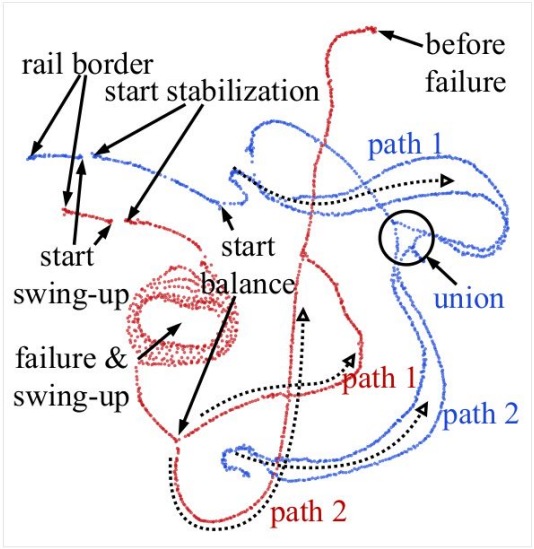
Fig. 7: Comparison of the temporal evolvement of layer activations between the baseline and three exemplary ablation cases for the CPSU task.



(a) Ablation of units 20 to 39 (5%) in the first layer.



(b) Units 110-149 (10%), layer 1.



(c) Units 260-289 (10%), layer 2.

Direction stage

Ordre partiels pour RL éthique : Cheng 2011 (modèle de preference-based RL utilisant des ordres partiels entre les actions), Eckersley 2019 (affirme que raisonner sur des ordres totaux pour faire du RL éthiques enfreint des lois morales et qu'il est obligatoire d'utiliser des ordres partiels à la place).
-> Adapter le modèle de Cheng 2011 pour faire du preference-based sur des politiques et non des actions comme Akrour 2012 mais avec le système d'ordres partiels en se basant sur les théorèmes d'Eckersley 2019.

Uncertainty pour RL éthique : Eckersley 2019 (affirme que valuer des politiques ou calculer des ordres totaux pour faire du RL éthiques enfreint des lois morales et qu'il est obligatoire d'ajouter de l'incertain), Glazier 2022 (transforme des pénalités de choisir les actions en probabilités -> incertain).
-> Adapter le modèle de Glazier pour intégrer de l'incertain dans un modèle de MORL éthique.
-> Adapter le modèle de Peschl 2021 pour intégrer le système de probabilités de Glazier 2021 pour ajouter de l'incertain dans ce MORAL éthique.

Systèmes de vote (de nash) pour décider quel est le meilleur compromis : Ecoffet & Lehman 2021 (vote de nash entre différents paradigmes éthiques) , Peschl 2021 (permet de générer des compromis divers entre plusieurs approximations de fonctions de récompenses d'experts -éthiques- mais manque d'une métaheuristique pour choisir le meilleur compromis).
-> Ajouter une métaheuristique éthique qui se base sur un vote de nash entre les experts et/ou des paradigmes éthiques prédéfinis pour choisir le meilleur compromis entre les solutions pareto-optimales générées par le MORAL de Peschl 2021.

Explicabilité du modèle (surtout pour DRL, l'IRL/GANs) : Meyes 2020 (explicabilité empirique du DL),
-> Beaucoup d'articles parlent de l'importance de l'explicabilité en éthique et de pouvoir assurer que les comportements des agents suivent vraiment les règles éthiques.
-> Faire une étude empirique inspirée de Meyes 2020 sur le modèle que l'on choisit d'implémenter.