

---

# Problèmes éthiques dans l’Apprentissage par Renforcement

Rapport de stage - M2 ANDROIDE

---

Autrice :

Alexandra DUFOUR

Tuteur·rice·s :

Aurélie BEYNIER  
Nicolas MAUDET  
Paolo VIAPPIANI

Référent :

Thibaut LUST

# Table des matières

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Etat de l'art</b>	<b>3</b>
<b>3</b>	<b>Travaux effectués</b>	<b>4</b>
3.1	Reward Shaping pour l'apprentissage de politiques éthiques . . . . .	4
3.2	Nouveau scénario . . . . .	6
3.2.1	Résultats . . . . .	8
3.3	Complexification des principes éthiques . . . . .	10
3.4	Ajout de l'ambulance dans notre environnement . . . . .	11
3.4.1	Apprentissage via la politique humaine . . . . .	12
3.4.2	Ajout d'une loi stricte . . . . .	13
3.5	Réinterprétation du problème dans un contexte d'apprentissage par renforcement multi-objectif . . . . .	16
3.5.1	Politique unique . . . . .	17
3.5.2	Politique multiple . . . . .	18
<b>4</b>	<b>Critique</b>	<b>20</b>
<b>5</b>	<b>Conclusion</b>	<b>21</b>
<b>A</b>	<b>Annexes</b>	<b>22</b>
A.1	Glossaire . . . . .	22
A.2	Tableaux . . . . .	23
A.3	Figures . . . . .	25
A.3.1	Performances pour la loi stricte sur l'ambulance . . . . .	25

# 1 Introduction

## Problématique

Alors que nous vivons dans une époque où l'intelligence artificielle se voit devenir le nouvel assistant personnel de l'humain, il semble que ses performances soient évaluées uniquement d'après le degré d'optimalité des comportements du système. Cependant être performant ne garantit pas que la décision prise par cette intelligence suit un comportement éthique [2]. On observe d'ailleurs en général une certaine opacité quant au raisonnement éthique que celle-ci suit, avec des structures qui offrent peu de transparence sur les justifications d'une décision [1].

Il est alors possible de tomber dans un dilemme éthique [23], c'est-à-dire une situation dans laquelle être performant correspond à rompre des règles éthiques, et adopter un comportement éthique mène à l'insatisfaction de l'utilisateur (résultat sous-optimal à ses attentes). On peut prendre l'exemple de la conduite automatique : on souhaite arriver le plus vite à destination tout en évitant les accidents qui peuvent être engendrés sur le trajet et en adoptant une conduite agréable pour les passagers. Une conduite confortable va éviter de changer de ligne trop fréquemment et adopter une vitesse modérée, ce qui entre en contradiction avec l'objectif initial qui est d'arriver le plus rapidement à destination. De plus les possibles ralentissements liés à l'évitement de collisions et de potentiels piétons sont une nouvelle entrave à cet objectif. Mais ces objectifs secondaires liés à la moralité de la conduite ne sont pas négligeables.

On retrouve ces dilemmes dans d'autres domaines, tels que la médecine où un équilibre entre efficacité d'un traitement et poids des effets secondaires se font souvent opposition. Les robots de compagnie [20] peuvent aussi rencontrer ce genre de situation en collectant des données sur leurs utilisateurs qui doivent rester privées mais peuvent devenir utiles si un accident se produit. Or il est nécessaire que l'utilisateur soit consentant dans la divulgation de ces informations.

C'est sur l'étude d'un compromis entre performance et respect des règles éthiques que porte le sujet de ce stage. Les modèles de Processus de Décision Markovien (MDP) et d'apprentissage par renforcement (RL) ont su prouver leur efficacité dans le domaine de l'intelligence artificielle pour concevoir des agents autonomes capables de prendre une suite de décisions dans des environnements qui peuvent être incertains (modélisés grâce à la fonction de transition). C'est pourquoi l'étude des dilemmes éthiques sera centrée sur l'emploi de ces deux modèles. Ce choix est aussi motivé par l'idée d'une certaine autonomie d'apprentissage fournie par l'apprentissage par renforcement car ces dilemmes peuvent se produire dans diverses situations qui sont difficilement toutes représentables par l'esprit humain.

Grâce à la construction de bases de données capturant un certain comportement [22] ou aux retours fournis directement par un expert [17], on a pu par apprentissage transférer à un agent les préférences exprimées [9]. La politique de l'agent résultante est donc obtenue sans construction d'une fonction de récompense en amont qui a pour avantage d'expliquer les intentions de l'agent, mais par des intentions exprimées implicitement à travers les données. Il est donc difficile dans ce cas de s'assurer que le processus de décision qui a été construit et adopté par l'agent est bien fondé sur des principes éthiques car il est rarement interprétable.

Ces méthodes offrent de nouvelles pistes pour simplifier l'apprentissage d'une politique, mais offrent aussi une grande liberté qui peut mener à des comportements peu éthiques qui doivent être évités.

## Objectifs

L'objectif principal de ce stage est d'étudier l'emploi de MDP et du RL afin d'obtenir des politiques respectant des principes éthiques. Nous allons aussi centrer nos recherches sur l'acquisition d'une fonction d'utilité éthique à partir d'interactions avec un expert, une procédure qui a été le sujet de papiers récents [10] [8] [11] [21].

Le but est de construire une politique tout autant efficace qu'éthique en s'appuyant sur les interactions avec un expert. Cette méthode repose grandement sur la qualité du retour fourni par l'expert, qu'il est nécessaire de décrire à travers la confiance qu'on peut lui accorder et la fréquence à laquelle nous pouvons le solliciter. Si les préférences de notre expert sont fournies à travers des données, il faut alors être capable de mesurer la qualité de celle-ci puisqu'elles impactent directement les résultats obtenus.

Il est donc nécessaire d'établir une méthode qui nous permette d'obtenir des garanties quant à la qualité éthique de la solution construite. Cette méthode doit être aussi accessible aux développeurs qui possèdent peu de compétences dans le domaine de l'éthique tout en garantissant des performances proches des méthodes plus exigeantes.

Travailler sur ce sujet est aussi l'occasion de gagner en expérience sur la modélisation de MDP, et d'en apprendre d'avantage sur les extensions de ce modèle, et de leurs méthodes d'apprentissage. C'est aussi la possibilité de travailler sur la modélisation de principes éthiques principalement fondés sur nos règles sociales dans un langage informatique, et sur la manière d'employer ces principes.

## 2 Etat de l'art

L'article [23] offre une récente et intéressante taxonomie sur la construction d'éthique dans l'intelligence artificielle. Quatre branches sont mises en lumière : les dilemmes éthiques, les décisions éthiques individuelles, celles collectives, et les interactions humain-IA. Notre sujet se concentre particulièrement sur les première, deuxième et quatrième branches en étudiant la politique développée par un seul agent lorsqu'il fait face à des dilemmes éthiques et qu'il est conseillé par un expert.

L'article aborde le sujet des interactions humain-IA tel que l'agent va tenter d'influencer son interlocuteur, alors que dans notre cas c'est un expert qui va apporter des informations dans le but de modifier la politique de notre agent. En plus de contrôler la qualité du comportement de l'agent, il est aussi nécessaire de contrôler celle des conseils fournis par l'expert. Une méthode pour reconstituer une politique à partir de retours humains en prenant en compte leur consistance et leurs fréquences a été présentée dans l'article [9], et nous fournit une base pour la représentation de notre expert.

Lorsqu'on parle d'éthique dans un comportement, on peut le traduire comme le fait de suivre des règles en accord avec des principes définis (conséquentialiste, déontologique, de vertu, ...). On retrouve alors deux modélisations de ces principes [18] :

- **Rule-based systems** utilisés dans les approches *Top-Down* : Les règles sont alors des lois fixes définies par des expressions logiques ou des limites en amont de la mise en circulation de l'agent. Elles ont l'avantage d'être généralement interprétables puisqu'elles ont dû être définies avant d'être modélisées avec des règles d'inférence. Cependant elles perdent en expressivité car elles se basent uniquement sur la vision du concepteur, et doivent être exprimables. Elles demandent aussi beaucoup plus de travail de représentation de l'éthique de la part du développeur.  
L'article [13] présente trois types d'implémentation de lois éthiques se basant sur des états interdits, des devoirs ou des vertus.
- **Data-driven systems** utilisés dans les approches *Bottom-Up* : les règles sont extraites d'exemples présentés à l'apprenant. Dans ce cas là, aucune règle n'est préintégrée. A partir des exemples et de leur contexte, les comportement à favoriser et à éviter sont distingués et cette tendance est réintroduite dans la construction de la politique. A l'inverse du système précédent, on se retrouve avec des lois difficilement interprétables mais qui gagnent en expressivité puisqu'elles ne sont pas limitées à la créativité du concepteur.

Une combinaison de ces deux systèmes peut aussi être envisagée afin de combiner les avantages de chacunes [19]. Dans le cadre de ce sujet, c'est principalement le deuxième système basé sur les données, déjà récoltées, ou fournies en temps réel, qui retient notre intérêt. Nous nous intéressons particulièrement à l'acquisition par un agent d'une fonction d'utilité éthique à partir d'exemples fournis par un expert, et ainsi libérer la tâche de définition des règles éthiques du concepteur.

Différentes manières d'introduire l'éthique dans l'intelligence artificielle grâce à l'apprentissage par renforcement ont par ailleurs été proposées :

- celle des **multi-armed bandit** où un agent va apprendre à **alterner entre plusieurs politiques** qui optimisent chacune un comportement particulier. Pour l'application d'un comportement éthique, l'agent va alors alterner entre une politique optimisant ses performances et une autre dite "éthique" qui sélectionne des actions qui vont suivre des principes éthiques [16]. L'agent sera entraîné à choisir en fonction de son état, qui sera représentatif du contexte dans lequel il se trouve [6] [5].

- celle qui va **modifier la politique** de l'agent pour qu'elle prenne en compte à la fois les objectifs définissant sa performance et ceux qui la limiteront pour atteindre un comportement éthique. Cette combinaison de deux politiques est atteinte par modification de la fonction de récompense, c'est le cas du **Reward Shaping** [22], ou par modification directe des valeurs de la politique, qu'on appelle alors **Policy Shaping** [9].
- et celle présentée dans le papier [13] qui va chercher le **meilleur compromis** entre la politique optimale pour la tâche qui lui est confiée, mais qui est "amorale", et toutes les politiques qui sont acceptables d'après des lois éthiques définies, donc qui sont "morales". Le but est d'éliminer les politiques qui ne rentrent pas dans les limites éthiques que nous aurions définies. Cette méthode demande d'avantage d'effort au développeur qui est responsable de la représentation du contexte éthique et des principes moraux impliqués choisis.

Le fait que l'on souhaite particulièrement se pencher sur le cas où un agent est influencé par l'avis d'un expert, nous amène d'avantage à explorer et étendre les méthodes qui vont **modifier la politique**.

L'idée de concilier de bonnes performances et un comportement éthique souligne la présence de plusieurs objectifs qui peuvent entrer en contradiction. Elle étend ce champs de recherche à l'étude des modèles markoviens multi-objectifs [3] (MOMDP). Ce modèle représente distinctement les divers objectifs en distinguant les fonctions de récompense propres à chaque objectif. Le signal de récompense se transforme alors en vecteur au lieu d'un simple scalaire recouvrant les différents objectifs. Les MOMDPs nous offre donc la possibilité de choisir plus explicitement les objectifs qu'on souhaite prioriser.

Ce type de modèle peut être résolu par apprentissage par renforcement multi-objectif (MORL). Deux formes de résolution sont proposées par [15]. La première se concentre sur la construction d'une politique unique en scalarisant les éléments du vecteur de récompense avec un vecteur de poids qui donne son importance à chaque objectif. Diverses méthodes quant à la sélection d'une solution désirée sont présentées telles que la norme de *Chebyshev* (distance maximale parmi les objectifs avec le point idéal majoré, à minimiser) et le calcul de l'hypervolume (volume entre un point de référence, proche du point Nadir ou pire cas, et un ensemble de solution, propre à une action dans notre cas, à maximiser). Ensuite dans le cas de la seconde, on cherche à explorer le front de pareto avec un algorithme en ligne, c'est-à-dire qui se construit grâce aux interactions de notre agent. Ils ont renommé cette méthode "Pareto Q-learning" car dans ce cas aucune scalarisation n'est appliquée, et c'est uniquement la dominance entre les diverses solutions trouvées qui est étudiée.

Cependant ce dernier modèle demande, contrairement au classique MDP, de définir en amont les différents objectifs qui seront pris en compte par notre agent. Nous pouvons aussi le voir comme un avantage car elle offre une certaine traçabilité des solutions choisies et ainsi du processus de décision qui est très souvent abstrait dans le cas d'un simple MDP. Or comme le domaine de l'éthique est assez vaste et non borné, il peut être très vite compliqué de pouvoir définir les objectifs bien distinctement.

### 3 Travaux effectués

Les travaux qui seront présentés dans cette partie reposent sur l'étude d'un agent modélisé sous la forme d'un Processus Décisionnel Markovien(MDP). Nous rappelons qu'un MDP est usuellement défini par un tuple  $\langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma \rangle$  tel que  $\mathcal{S}$  est l'ensemble des états possibles,  $\mathcal{A}$  est l'ensemble des actions,  $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S}$  est la fonction de transition qui donne par  $\mathcal{T}(s, a, s')$  la probabilité que notre agent a de finir dans l'état  $s'$  en exécutant l'action  $a$  dans l'état  $s$ ,  $\mathcal{R} : \mathcal{S} \times \mathcal{A} \times \mathcal{S}$  est la fonction de récompense, et  $\gamma$  est le facteur d'actualisation pour le calcul de notre récompense cumulée espérée lors de l'apprentissage.

Afin de prendre en main le sujet, il a été nécessaire d'étudier l'état de l'art du domaine de l'éthique dans l'univers de l'intelligence artificielle, et particulièrement les méthodes utilisées pour l'introduire au sein de l'apprentissage par renforcement. Les divers articles étudiés ont alors mis en lumière diverses méthodes qui ont été abordées dans la section précédents. Cependant c'est un article en particulier qui a retenu notre attention et c'est sur lui que se base le travail effectué pendant ce stage.

#### 3.1 Reward Shaping pour l'apprentissage de politiques éthiques

Après avoir étudié les divers papiers recommandés ([23] [18] [1] [22] [16]), c'est la méthode présentée dans le papier [22] qui a particulièrement retenu notre attention. Le concept présenté est basé sur l'utilisation du *Reward*

*Shaping* pour apporter de l'éthique dans le comportement de l'agent. En partant du problème de décision d'un agent modelisé sous la forme d'un MDP qui voit sa fonction de récompense uniquement centrée sur ses performances, elle est alors augmentée d'une valeur négative ou positive selon si l'action est défavorisée ou favorisée par une politique "humaine" considérée éthique. La politique humaine est supposée avoir été construite en amont par des retours humains grâce à la méthode présentée dans l'article [9].

Cette représentation est idéale pour soulager le concepteur de la prise en compte de l'éthique de son agent, puisqu'il y est supposé que l'acquisition d'une politique humaine éthique pour le corriger peut être faite à partir de cas généraux même assez éloignés du cas traité par notre agent tant que le comportement exprimé est éthique. De part l'utilisation qu'ils font du *Reward Shaping*, cette méthode a été renommée **Ethics Shaping**.

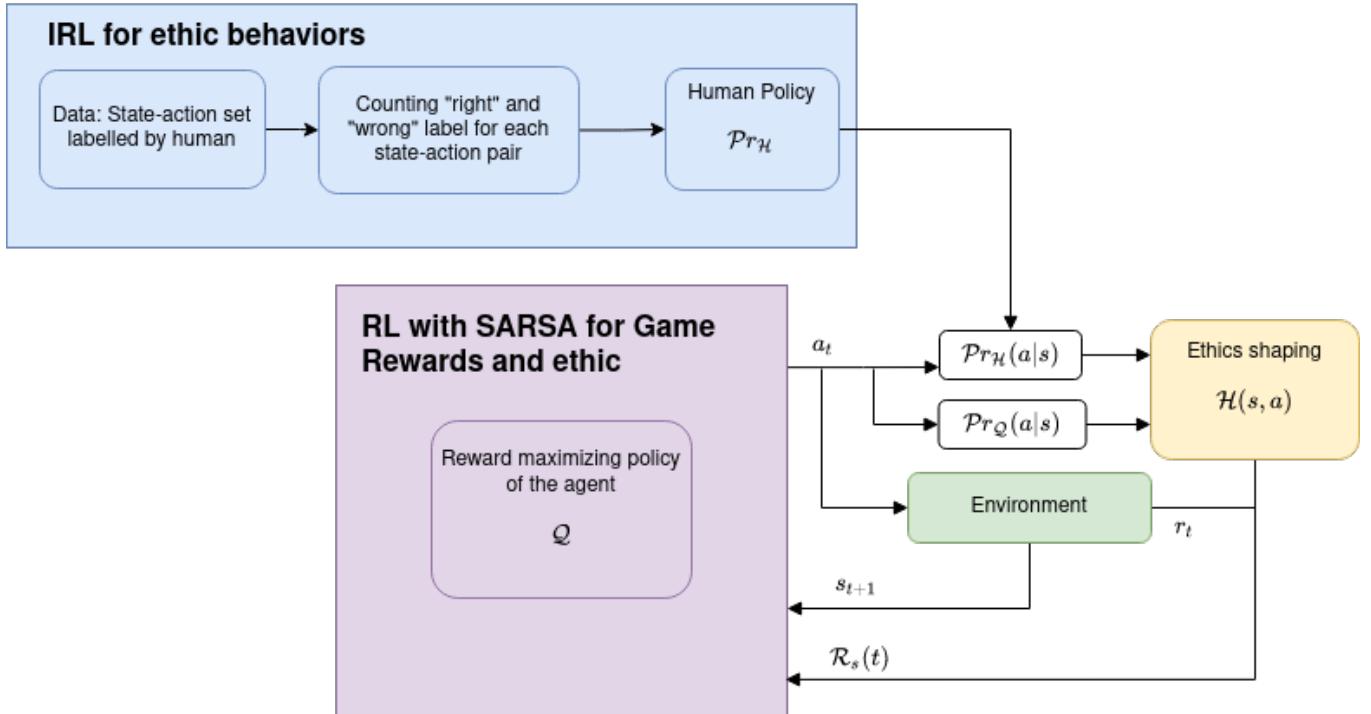


FIGURE 1 – Représentation de la méthode par **Ethics Shaping**

A partir du problème décisionnel de notre agent représenté sous la forme d'un MDP, la valeur de la récompense modifiée par le *Reward Shaping* prise en compte par l'algorithme SARSA pour son apprentissage est définie telle que :

$$\mathcal{R}_s(s_t, a_t, s_{t+1}) = \mathcal{R}(s_t, a_t, s_{t+1}) + \mathcal{H}(s_t, a_t, s_{t+1})$$

avec :

- $\mathcal{R}_s$  : la récompense agrégée reçue par l'apprenant
- $\mathcal{R}$  : la récompense de l'environnement
- $\mathcal{H}$  : la récompense du *Reward Shaping*

La valeur générée par le *Reward Shaping* est ajoutée à la récompense de l'environnement lorsque nous faisons face à deux situations :

- celle d'une **décision négativement éthique**, c'est-à-dire que la probabilité de réaliser l'action sélectionnée  $a$  dans un état  $s$  par la politique de l'agent est supérieure à celle de la politique humaine éthique en réalisant la même action  $a$  dans le même état  $s$ , dont la valeur est inférieure à un seuil  $T_n$ . Le seuil étant proche de 0, réaliser l'action  $a$  dans l'état  $s$  correspond à prendre une décision que notre politique éthique souhaite éviter. La valeur du *Reward Shaping* est alors **négative**.
- et celle d'une **décision positivement éthique**, c'est-à-dire que cette fois-ci, la probabilité de réaliser l'action sélectionnée  $a$  dans un état  $s$  par la politique de l'agent est inférieure à celle de la politique humaine éthique en réalisant la même action  $a$  dans le même état  $s$ , dont la valeur est supérieure à un seuil  $T_p$ . Le seuil étant

choisi le plus grand possible, réaliser l'action  $a$  dans l'état  $s$  est favorisé par notre politique. La valeur du *Reward Shaping* est alors **positive**.

Hors de ces deux situations, aucun *Reward Shaping* n'a besoin d'être appliqué, et l'agent suit uniquement les objectifs qui lui sont définis initialement. Un diagramme représentant le processus qui vient d'être décrit est présenté en figure 1.

La valeur de la pénalité ou de la récompense, ajoutée à la récompense de l'environnement de notre agent par le *Reward Shaping*, est proportionnelle à la divergence de Kullback-Leibler entre les distributions d'actions des deux politiques pour un même état. Il est aussi possible de contrôler leur importance grâce à des coefficients définis en amont, indépendamment pour les pénalités ( $c_n$ ) et les récompenses ( $c_p$ ). Notre fonction  $\mathcal{H}$  prend donc la forme suivante :

$$\mathcal{H}(s, a) = \begin{cases} -c_n \cdot D_{KL}(Pr_Q(a|s) || Pr_H(a|s)) & \text{if } Pr_Q(a|s) > Pr_H(a|s) \\ & \text{and } Pr_H(a|s) < T_n \\ & \text{Décisions négativement éthiques} \\ c_p \cdot D_{KL}(Pr_Q(a|s) || Pr_H(a|s)) & \text{if } Pr_Q(a|s) < Pr_H(a|s) \\ & \text{and } Pr_H(a|s) > T_p \\ & \text{Décisions positivement éthiques} \\ 0 & \text{otherwise} \end{cases}$$

tel que la *divergence de Kullback-Leibler* est

$$D_{KL}(Pr_Q(s) || Pr_H(s)) = \sum_{a \in \mathcal{A}} Pr_Q(a|s) \log \frac{Pr_Q(a|s)}{Pr_H(a|s)}$$

et que  $Pr_Q(a|s)$  et  $Pr_H(a|s)$  correspondent à la probabilité de sélectionner l'action  $a$  dans l'état  $s$  avec la politique de notre agent et avec la politique humaine respectivement.

Cette approche est particulièrement intéressante car la politique humaine va venir guider la politique de notre agent et non la modifier directement. On peut s'attendre grâce à cette méthode à obtenir une politique éthique qui maintient une bonne optimisation sur notre objectif initial tout en s'articulant autour pour prendre en compte les objectifs transmis par les données.

### 3.2 Nouveau scénario

Il s'est avéré après avoir récupéré et analysé le code fourni par les auteurs<sup>1</sup>, que les politiques humaines étaient modélisées par des MDPs qui étaient eux-même entraînés par apprentissage par renforcement. Ce qui nous a ouvert la possibilité de créer nos propres politiques humaines éthiques, en modifiant la fonction de récompense pour capturer le comportement attendu, et parfois la dimension de son espace d'état.

Cette méthode d'*Ethics Shaping* a été testée et valorisée sur trois expériences, mais ce sont les deux dernières qui ont retenu notre intérêt de part l'extensivité de leurs scénarios. Elle partage toutes les deux le même environnement qui est une route à cinq voies où d'autres voitures peuvent circuler comme l'illustre la figure 2.

Dans un premier temps, nous avons étudié l'expérience *Driving and Avoiding* qui ajoute dans l'environnement en plus de la présence des autres véhicules, la présence de chats blessés. A partir d'une politique humaine entraînée à éviter les chats blessés immobiles, on cherche à inférer à l'agent cette capacité par *Ethics Shaping* alors que le seul objectif défini dans sa fonction de récompense est d'éviter les collisions avec les autres voitures.

Puis dans un deuxième temps, *Driving and Rescuing* représente l'expérience opposée. Au lieu de chats blessés, ce sont des personnes âgées qui se trouvent coincées dans le trafic. La politique humaine éthique n'est alors plus entraînée à éviter ses cibles, mais à les traverser ce qui correspond à les prendre à bord et les sauver. L'action ne prend pas de temps de supplémentaire et est réalisée lorsque l'on rentre en collision avec l'individu. On souhaite désormais que l'agent, dont le seul objectif est toujours d'éviter la collision avec les autres voitures présentent sur le trafic, sauve simultanément les personnes âgées en détresse qui sont immobiles.

---

1. <https://github.com/kristery/EthicsShaping>

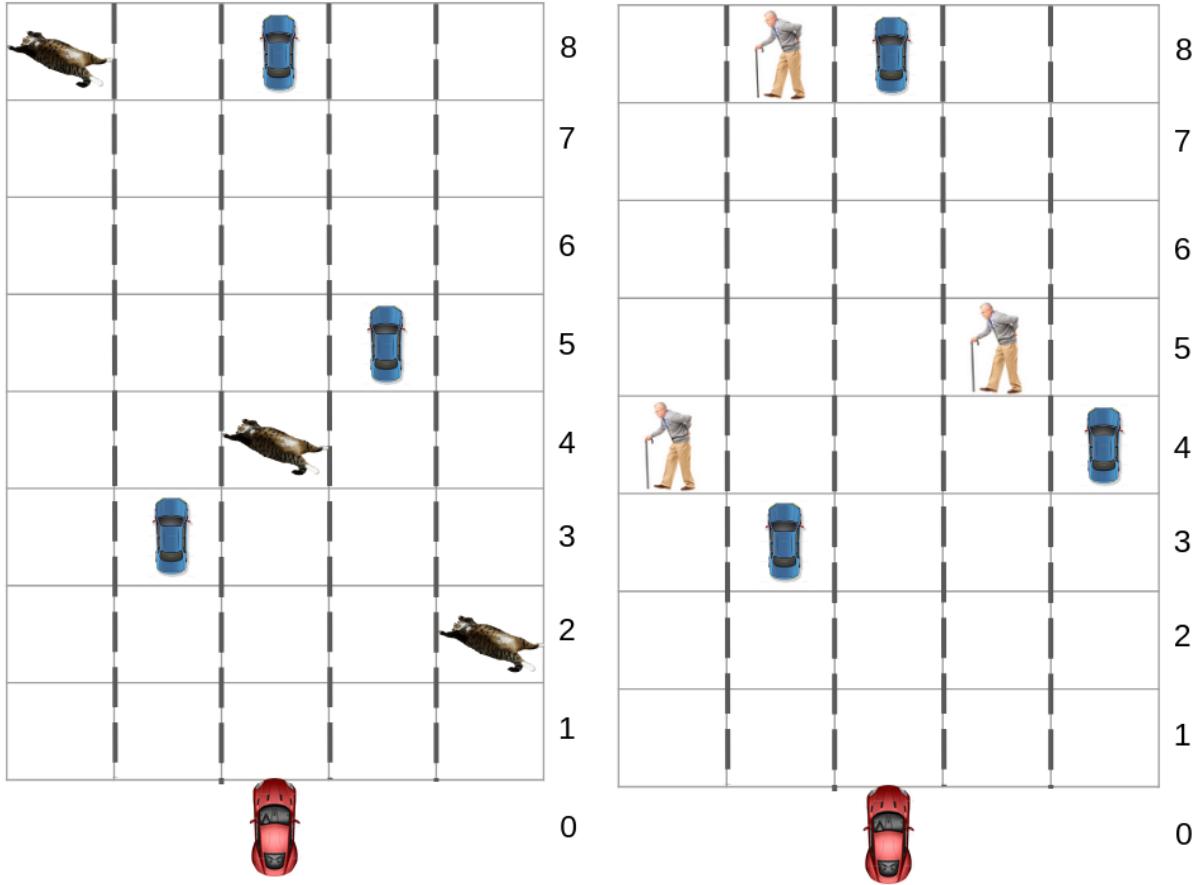


FIGURE 2 – Représentation visuelle de l’expérience *Driving and Avoiding* (gauche) et de l’expérience *Driving and Rescuing* (droite).

A partir de ces deux expériences, nous avons défini une troisième expérience qui combinerait les deux premières, *Driving and Avoiding and Rescuing*. Cette fois, on retrouve des chats blessés et des personnes agées qui sont coincés dans notre trafic. Le but de cette expérience est de vérifier s’il est possible de transférer plusieurs comportements éthiques à notre agent tout en maintenant de bonnes performances. Notre politique humaine est entraînée en combinant les récompenses et pénalités attribuées dans les fonctions de récompense des deux politiques humaines précédentes pour modéliser le comportement attendu. Elle prend aussi en compte l’évitement des autres véhicules même de manière moindre que l’agent, car on suppose que si des personnes se trouvent à bord, on ne peut négliger que des collisions leur seraient très défavorables. Elle est utilisée dans le processus d’*Ethics Shaping* pour transférer ce comportement à notre agent qui ne se concentre que sur ses performances liées à la collision. On retrouve les performances de notre agent obtenues dans les figures 3 et 4.

Dans notre modèle, nous distinguons deux types d’état : l’état **général** et celui **éthique**. L’état **général** est utilisé pour l’apprentissage, il fournit les informations sur notre environnement et correspond à l’ensemble  $\mathcal{S}$  du MDP de notre agent. L’état **éthique** lui est utilisé par l’*Ethics Shaping* pour appeler la politique humaine. Il renferme donc le contexte éthique sous lequel la décision de sélectionner une certaine action a été réalisée.

L’étape la plus importante dans la construction de la politique humaine est le choix de cet état *éthique* sous lequel elle sera sauvegardée. Le choix de sa forme traduit les informations que l’on souhaite fournir à notre agent qui ont un rôle lors de la prise de décision d’un point de vue éthique. Ainsi plus l’état *éthique* est proche de l’état **général**, moins les lois éthiques exprimées sont généralisables et distinctes de l’objectif initial amoral à optimiser. On a pu voir que le choix de l’auteur pour ces deux premiers scénarios a été d’utiliser la ligne de notre agent couplée respectivement à la position des chats blessés ou des personnes agées sur les lignes présente et adjacentes. Les positions des véhicules adjacents ne sont alors pas sauvegardées, et donc pas représentées dans la prise de décision de la politique humaine. Il a fait le choix d’un état éthique réduit qui est déconnecté de l’objectif initial d’éviter les collisions afin de prouver que même des données générales où l’attitude représentée ne prend pas en compte cette objectif peuvent nous permettre d’obtenir les performances souhaitées. De notre côté, de part la combinaison des deux tâches, nous avons remarqué que les meilleurs résultats étaient obtenus lorsque l’état éthique était égal à

l'état général, et que le réduire le menait forcément à un comportement sous optimale (voir le tableau 5 fournis en annexe). Nos deux types d'état sont alors égaux et représentés de la manière suivante dans notre modèle :

$$\text{Etat général} = \text{Etat éthique} = (l, v_0, c_0, e_0, v_1, c_1, e_1, v_2, c_2, e_2)$$

avec :

- $l$  : sa ligne actuelle
- $v_0$  (resp.  $c_0, e_0$ ) : la voiture la plus proche (resp. chat, personne agée) sur la **même ligne**
- $v_1$  (resp.  $c_1, e_1$ ) : la voiture la plus proche (resp. chat, personne agée) sur la **ligne à sa gauche**
- $v_2$  (resp.  $c_2, e_2$ ) : la voiture la plus proche (resp. chat, personne agée) sur la **ligne à sa droite**

Par exemple, dans l'expérience représentée en figure 2 à gauche, notre état prend la valeur  $(3, 8, 4, -1, 3, -1, -1, 5, -1, -1)$ , tel que la valeur  $-1$  est attribuée lorsque un élément n'est pas présent sur la ligne concernée, et  $-2$  si la ligne est inaccessible. A droite, il prends la valeur  $(3, 8, -1, -1, 3, -1, 8, -1, -1, 5)$ .

L'ensemble des actions  $\mathcal{A}$  de notre agent ne contient que trois éléments :  $\{\text{dériver sur la ligne à droite}, \text{dériver sur la ligne à gauche}, \text{aller tout droit}\}$ . Il n'est pas possible pour notre agent de reculer puisque l'environnement n'est plus modélisé au-delà de son horizon et il ne lui est pas possible de contrôler sa vitesse qui est constante et similaire à celle des autres véhicules.

Il est aussi important de préciser les poids utilisés dans la fonction de récompense de notre simulation pour représenter l'importance des différents objectifs. Le tableau 1 regroupe les diverses informations qui définissent cette fonction de récompense reçue par notre agent lors de ses interactions.

Type de la politique	Variable	Domaine	Poids
Politique de l'agent	car_hit go_straight	{0,1} {0,1}	-20 0.5
Politique humaine <i>Driving and Avoiding</i>	car_hit go_straight cat_hit	{0,1} {0,1} {0,1}	-1 0.5 -20
Politique humaine <i>Driving and Rescuing</i>	car_hit go_straight elder_saved	{0,1} {0,1} {0,1}	-1 0.5 20
Politique humaine <i>Driving and Avoiding and Rescuing</i>	car_hit go_straight elder_saved cat_hit	{0,1} {0,1} {0,1} {0,1}	-15 0.5 20 -15

TABLE 1 – Fonction de récompense des différentes politiques. Les variables repérentent les différentes infractions et action qui ont un impact sur la récompense reçue. Si elle est égale à 1 c'est qu'elle est réalisée, sinon elle vaut 0. Les différentes variables sont multipliées par leur poids puis additionnées ensemble pour former la valeur de la récompense.

### 3.2.1 Résultats

Avant de présenter les résultats de nos expériences, nous introduisons le code de couleur suivant qui associe une couleur à une politique qui est résumé dans le tableau 6 en annexe :

- en **noir** : la politique de l'agent classique dont le but est d'éviter un maximum les collisions avec les autres voitures. Elle est amorphe et ne prend pas en compte les problématiques éthiques.
- en **bleu** : la politique humaine dont le but principal est d'éviter d'écraser les chats blessés sur la route, propre à l'expérience *Driving and Avoiding*. Elle ne prend que faiblement en compte l'évitement des autres véhicules.
- en **rouge** : la politique résultante de l'apprentissage de la politique classique couplée au Ethics Shaping avec la politique humaine de *Driving and Avoiding*. Elle ne prend que faiblement en compte l'évitement des autres véhicules.
- en **vert** : la politique humaine dont le but principal est de sauver le plus de personnes agées sur la route en les percutant, propre à l'expérience *Driving and Rescuing*.
- en **jaune** : la politique résultante de l'apprentissage de la politique classique couplée au Ethics Shaping avec la politique humaine de *Driving and Rescuing*.
- en **violet** : la politique humaine **mixte** dont le principal est de sauver des personnes agées, et d'éviter d'écraser le plus de chat tout en limitant le nombre de collisions, propre à l'expérience *Driving and Avoiding and Rescuing*.
- en **orange** : la politique résultante de l'apprentissage de la politique classique couplée au Ethics Shaping avec la politique humaine **mixte** de *Driving and Avoiding and Rescuing*.

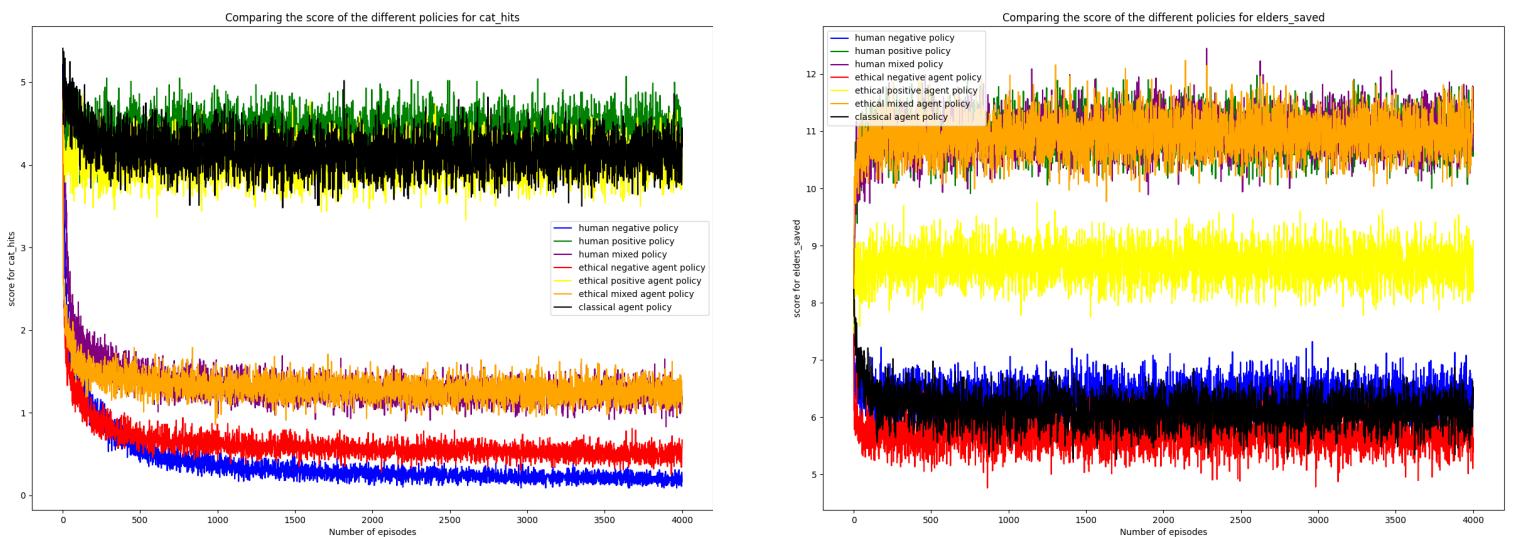


FIGURE 3 – Evolution du nombre de chats blessés percutés (gauche) à minimiser et du nombre de personnes agées démentes sauvées (droite) à maximiser au cours de l'apprentissage pour des expériences de 300 pas. Moyenne réalisée sur 100 runs d'apprentissage, tel que  $c_n = 1.00$ ,  $\mathcal{T}_n = 0.20$ ,  $c_p = 2.00$  et  $\mathcal{T}_p = 0.50$ .

On remarque bien d'après la figure 3 que nous avons su générer une politique éthique mixte (**violet**) capable de prendre en compte l'évitement des chats avec une moyenne inférieure à 2 chats percutés par expérience à la fin de notre apprentissage contre une moyenne au dessus de 4 lorsque l'*Ethics Shaping* n'est pas appliqué. On remarque cette même amélioration dans le sauvetage de personnes agées avec un score de 11 personnes sauvées avec *Ethics Shaping* contre environ 6 sans *Ethics Shaping*, et légèrement en dessous de 9 lorsque l'*Ethics Shaping* est appliqué avec une politique humaine qui ne prend en compte que le sauvetage des personnes agées. C'est cette dernière observation qui nous a fait prendre conscience de l'importance du contexte représenté par un état.

La figure 4 met quant à elle en évidence la capacité nettement supérieure de notre politique éthique mixte par rapport aux autres politiques générées par l'*Ethics Shaping*, à éviter la collision avec les autres véhicules grâce à la qualité de la politique éthique dont elle dépend. Ces résultats lui permettent d'atteindre des performances au moins aussi bonnes que celles de l'agent classique en se reposant uniquement sur le nombre de collisions avec les autres

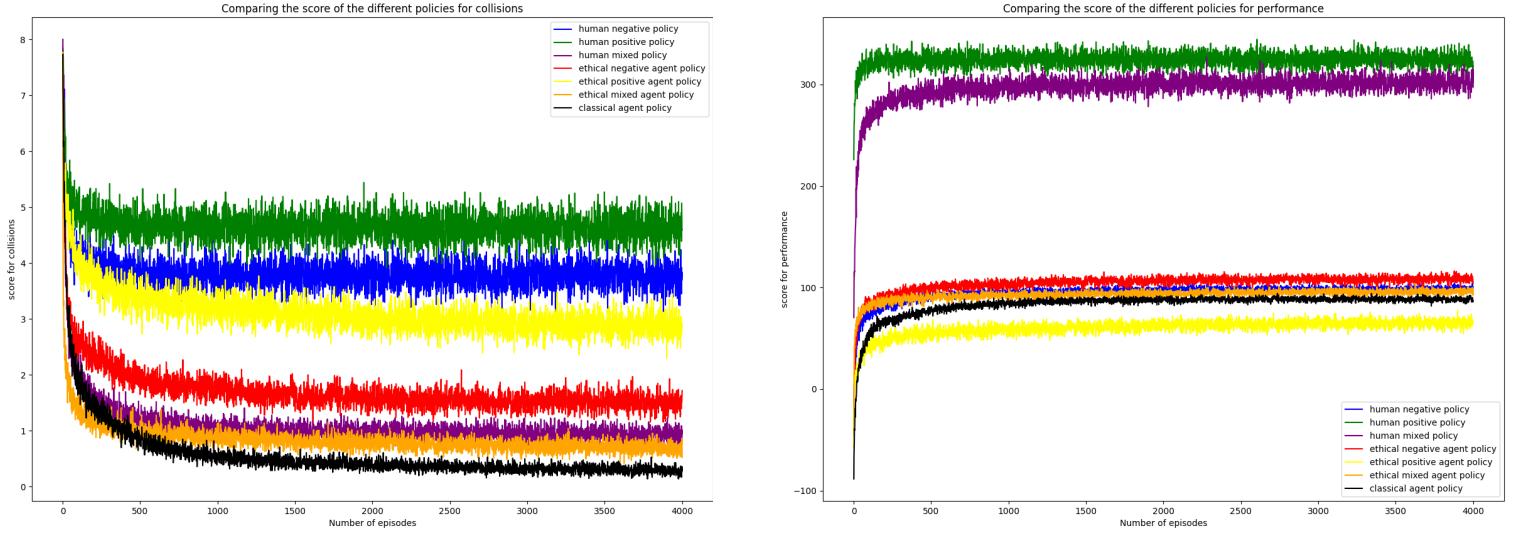


FIGURE 4 – Evolution du nombre de véhicules percutés (gauche) à minimiser et de la récompense cummulée (droite) à maximiser au cours de l'apprentissage pour des expériences de 300 pas. Moyenne réalisée sur 100 runs d'apprentissage, tel que  $c_n = 1.00$ ,  $T_n = 0.20$ ,  $c_p = 2.00$  et  $T_p = 0.50$ .

véhicules et sa capacité à rester sur la même voie. Les performances des politiques humaines ne sont pas à prendre en compte car elles se basent sur la récompense de leur propre modèle.

Notre politique résultante a des performances nettement supérieures aux politiques obtenues avec les expériences précédentes. Cette différence s'explique par l'expression de l'état de la politique humaine qui est très similaire à celle de notre agent dans notre cas, alors qu'un contexte restreint avait été utilisé par l'auteur pour ses deux politiques humaines. Cela remet en cause l'affirmation de [22] disant qu'il est possible d'utiliser des données très générales pour inférer un comportement éthique à notre agent. Ces résultats prouvent qu'utiliser des données humaines qui manquent de précision a un réel impact négatif quant aux performances de l'*Ethics Shaping*.

### 3.3 Complexification des principes éthiques

Nous avons aussi tenter d'inférer des règles temporellement complexes [4] à notre agent. Ce sont des lois qui vont changer au cours du temps et de l'évolution du contexte. Elles dépendent des états passés de notre agent ou du temps, une propriété en opposition avec le principe de Markov qui de part son processus sélectionne une action uniquement à partir de l'état courant de l'agent. Ce principe est souvent contourné en intégrant des informations propres à la temporalité au sein même de l'état ou en définissant de nouveaux modèles de décision plus généraux [7]. Dans cette optique, nous avons implémenté une loi qui fait décroître l'importance de l'écrasement des chats quand nous prenions en charge d'avantage de personnes âgées à bord. Ce scénario avait pour but de prendre en compte les dilemmes éthiques lorsque l'agent doit choisir entre entrer en collision avec un chat blessé ou un autre véhicule, en supposant qu'un choc avec un autre véhicule serait plus préjudiciable aux passagers avec une santé fragile que de subir l'écrasement d'un chat. Il s'est avéré que les objectifs sont suffisamment alignés pour que les résultats montrent juste une augmentation du nombres de chats percutés, suite au relâchement de la fonction de récompense dans le temps, sans voir les performances des autres critères augmenter. Les résultats de cette expérience ne seront donc pas affichés ici mais peuvent être retrouvés dans d'autres rapports<sup>2 3</sup>.

Une partie de la construction de ces nouveaux scénarios s'est aussi concentrée sur la forme de l'état du MDP de notre agent. En effet, quand nous avons souhaité prendre en compte le nombre de passagers dans notre exemple précédent, nous nous sommes rendus compte que cette valeur pouvait aller de 0 à 11 en moyenne, ce qui générait un grand nombre d'états. Pour palier à ce problème, nous avons regroupé les valeurs prises par cette variables en

2. <https://drive.google.com/file/d/1H0qvhWN8L6oMUGml9n0j8KFESs8ZmAom/view?usp=sharing>

3. [https://drive.google.com/file/d/1Fp\\_HAH3E01BFz1o07v6qN0Tn1eN1YSr/view?usp=sharing](https://drive.google.com/file/d/1Fp_HAH3E01BFz1o07v6qN0Tn1eN1YSr/view?usp=sharing)

plusieurs groupes pour qu'elle ne puisse prendre plus que des valeurs comprises entre 0 et 3. Les états général et éthiques étaient donc désormais de la forme suivante :

$$\text{Etat général} = \text{Etat éthique} = (l, v_0, c_0, e_0, v_1, c_1, e_1, v_2, c_2, e_2, n\_passenger)$$

avec

- $l, v_0, c_0, e_0, v_1, c_1, e_1, v_2, c_2, e_2$  : identiques
- $n\_passenger$  : une échelle du nombre de personnes agées déjà sauvées à bord du véhicule telle que :

$$n\_passenger = \begin{cases} 0, & \text{si pas de passagers} \\ 1, & \text{si au moins 1 passager et au plus 3 passagers} \\ 2, & \text{si au moins 4 passagers et au plus 7 passagers} \\ 3, & \text{si au moins 8 passagers ou plus} \end{cases}$$

Cependant les résultats restaient inférieurs ou égaux à ceux que nous venons de présenter. Diminuer l'importance de l'écrasement des chats pour favoriser l'évitement des autres véhicules ne semble donc pas être le meilleur critère pour étudier l'implémentation de lois éthiques temporellement complexes et de compromis compte-tenu de l'alignement suffisant des différents objectifs.

### 3.4 Ajout de l'ambulance dans notre environnement

Dans l'optique d'étudier les comportements temporellement complexes, et d'enrichir le type d'obstacle proposé à notre agent, nous avons introduit la présence d'ambulances dans le trafic. Lorsqu'elles sont présentes, elles privent notre agent de l'accès aux deux lignes de gauche sur notre route comportant cinq voies. Ainsi quand une ambulance s'approche, l'agent doit libérer au plus vite ces deux voies. Elles redeviennent accessibles dès que l'ambulance s'éloigne puisque le risque de collision est alors évité compte-tenu que notre simulation tourne sous l'hypothèse que la vitesse de l'ambulance est supérieure à celle de notre agent.

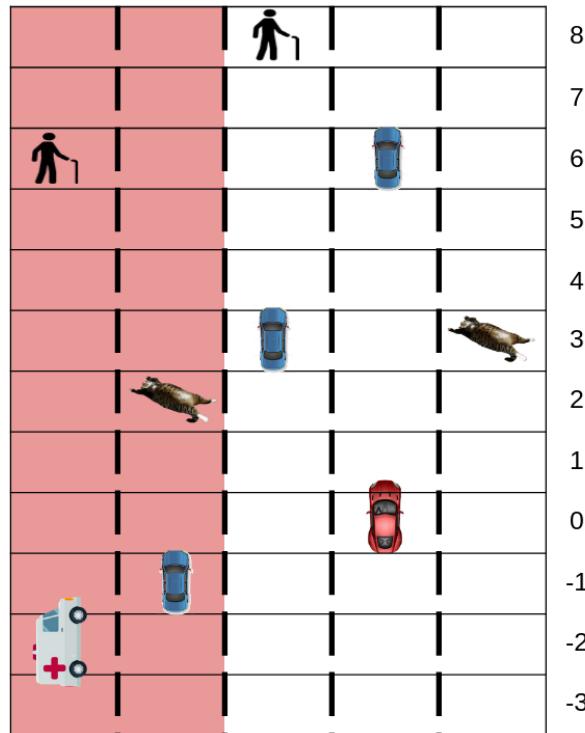


FIGURE 5 – Scénario *Driving and Avoiding and Rescuing* avec intégration de l'ambulance.

Nous avons choisi de modéliser la présence de l'ambulance dans le trafic par l'intensité de sa sirène qui est captée par l'agent et donc représentée dans son référentiel. Elle a l'avantage d'augmenter lorsqu'elle se rapproche de

l'horizon de notre agent et de diminuer lorsqu'elle s'en éloigne. Cependant elle ne nous fournit aucune information sur quelle ligne de gauche l'ambulance se trouve. On suppose pour le moment que cette intensité est totalement observable sans prendre en compte le bruit dont elle pourrait souffrir, et on néglige le déplacement des autres véhicules vers la droite lorsqu'ils sont présents sur les voies de gauches. Ces futurs modifications peuvent réduire les valeurs des performances mais n'impactent pas la manière dont nous tentons de résoudre cette problématique.

Afin de représenter ce nouvel élément dans notre environnement, nous avons dû redéfinir la structure d'un état pour qu'il puisse contenir les informations qui nous intéressent. On le redéfinit alors de la manière suivante :

$$\text{General State} = (l, v_0, c_0, e_0, v_1, c_1, e_1, v_2, c_2, e_2, n\_passenger, I_{ambulance}, I_{ambulance\_deriv})$$

avec

- $l, v_0, c_0, e_0, v_1, c_1, e_1, v_2, c_2, e_2$  : identiques.
- $I_{ambulance}$  : l'intensité de la sirène de l'ambulance, qui peut prendre une valeur comprise entre 0 et 9, ce qui correspond au champ d'observabilité de notre agent qui est choisi identiquement à celui des autres obstacles (8 cases en amont au maximum). Elle est représentée de la manière suivante :

$$I_{ambulance} = 9 - | \text{position de l'ambulance par rapport à l'horizon} |$$

Il n'est pas possible pour l'agent de savoir si elle ne se trouve devant ou derrière lui à partir de cette valeur car cette information ne peut être transmise par la captation simple d'un signal sonore. Cela nous permet aussi de limiter le nombre de valeurs prises par cette variable qui est symétrique en amont et aval de l'horizon.

- $I_{ambulance\_deriv}$  : la dérivée de l'intensité de la sirène de l'ambulance, elle dépend de la valeur perçue au temps  $t$  et de celle observée au pas de temps précédent  $t - 1$ .

Le modèle de l'ambulance est formé de deux paramètres :  $I_{ambulance}$  fait le lien directement avec la distance qui sépare l'ambulance de l'horizon de notre agent, et  $I_{ambulance\_deriv}$  apporte une information sur la vitesse de déplacement de celle-ci. Grâce à ces deux informations, l'agent peut adapter son comportement afin d'éviter l'ambulance avec une vitesse variable. Si la dérivée de l'intensité de l'ambulance n'est pas représentée dans notre état, on suppose alors que la vitesse est constante car notre agent ne peut pas prévoir les sauts d'intensité variables et donc cela le rendra moins performant dans l'évitement de collisions.

En amont, nous avions mené des expériences employant le signe de la dérivée de l'intensité et le nombre de fois cumulées qu'elle est observée pour modéliser l'ambulance dans l'état général et éthique. De très bons résultats étaient obtenus uniquement en fixant un seuil de danger quant aux valeurs prises par ces deux variables, induisant une connaissance de ces limites par le développeur<sup>4</sup><sup>5</sup>. Cette approche était plus contraignante et sa représentation plus abstraite. De plus elle ne prend pas en compte la vitesse de l'ambulance en supposant que celle-ci est constante, une hypothèse qu'il est difficile de conserver en environnement réel. C'est la dépendance temporelle entre les états induite par les valeurs prises par ces deux variables qui nous avait poussé à étudier cette approche qui n'a finalement pas été concluante de part son manque de robustesse.

Deux méthodes ont été étudiées pour répondre à cette nouvelle problématique à partir des deux modélisations principales présentées dans la section 2. La première reprend le fonctionnement des approches **Bottom-Up** jusqu'alors utilisées dans nos expériences en enrichissant la politique humaine introduite précédemment. La seconde quant à elle s'appuie sur l'approche basée sur les lois, aussi appelée **Top-Down**, en forçant l'agent à appliquer une politique qui interdit tout mouvement pouvant conduire à une collision avec l'ambulance. Ces deux méthodes sont présentées dans les sous-sections suivantes.

### 3.4.1 Apprentissage via la politique humaine

Cette expérience reste une extension de celle présentée plus tôt dans ce rapport. Ainsi comme justifié dans la section 3.2, l'état général et l'état éthique sont identiques. L'ensemble des états a donc grandi, et la fonction de récompense de la politique humaine a été rallongée d'une variable comme résumé dans le tableau suivant.

---

4. <https://drive.google.com/file/d/1QV9rnBRh9z-g-1E3WxssBfy7VaqSbewQ/view?usp=sharing>

5. <https://drive.google.com/file/d/1CYIN9XaxrLkZNZQXBQSe3F4mteDzuodG/view?usp=sharing>

Type de la politique	Variable	Domaine	Poids
Politique humaine	car_hit	{0,1}	-15
	go_straight	{0,1}	0.5
	elder_saved	{0,1}	20
	cat_hit	{0,1}	-15
	ambulance_hit	{0,1}	-50
Politique de l'agent	car_hit	{0,1}	-20
	go_straight	{0,1}	0.5

TABLE 2 – Fonction de récompense de la politique humaine et de notre agent. Les variables représentent les différentes infractions et actions qui ont un impact sur la récompense reçue. Si elle est égale à 1 c'est qu'elle est réalisée, sinon elle vaut 0. Les différentes variables sont multipliées par leur poids puis additionnées ensemble pour former la valeur de la récompense.

Nous avons modélisé la préférence de la politique humaine à éviter les ambulances en attribuant à cet objectif un poids négatif nettement supérieur à la variable *ambulance\_hit* quand elle est réalisée. Ce nouveau modèle nous permet d'obtenir les résultats présentés en figure 6.

Les graphes présentés en figure 6 montrent que cette méthode obtient des performances très satisfaisantes sur tous les critères en produisant une politique qui combine presque identiquement les meilleures performances des politiques classiques et humaines qui la forment. On remarque que son pourcentage sur l'évitement des autres véhicules est très proche de la politique optimale (6c) amorphe sur ce critère, ce qui est directement relié à la qualité de la politique humaine fournie. Cependant nous n'atteignons pas un taux nul pour l'évitement des ambulances (6a), que ce soit pour la politique humaine ou éthique. C'est une performance que nous souhaitons atteindre grâce à la seconde méthode présentée en section 3.4.2.

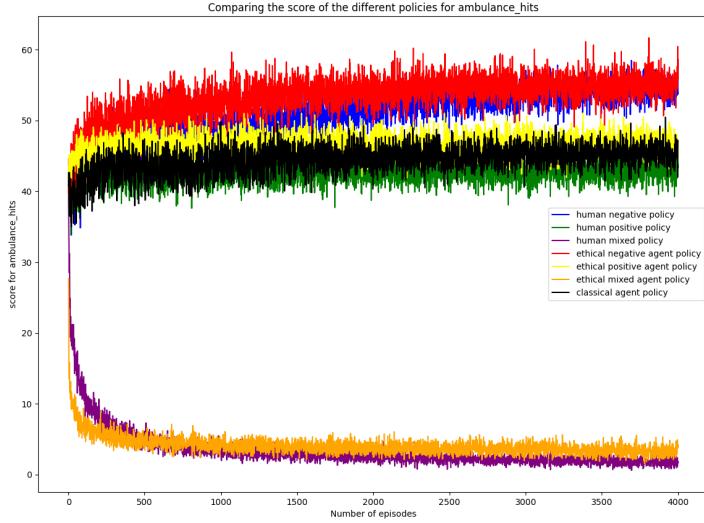
Cette méthode a pour avantage de ne pas demander à l'utilisateur de fixer un seuil de danger quant à la valeur de l'intensité de la sirène de l'ambulance. C'est grâce au processus d'apprentissage et au raisonnement qui en découle que l'agent va construire ce seuil lui-même en extrapolant la vitesse de l'ambulance grâce à la dérivée de son signal.

### 3.4.2 Ajout d'une loi stricte

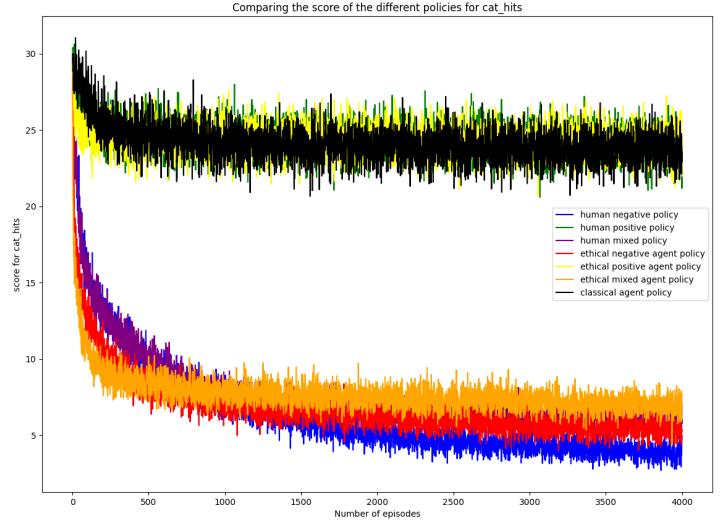
Dans cette seconde méthode, nous avons cherché à étudier la capacité de notre méthode **Data-driven** initiale à se coordonner avec une méthode dite **Rule-based**. Il serait en effet intéressant de pouvoir établir des règles fixes lorsque cela est possible, tout en laissant l'agent optimiser l'aspect éthique de son comportement via des données lorsqu'il est compliqué de fixer une loi stricte pour certains objectifs. C'est par exemple notre cas dans cette situation : on souhaite absolument éviter toutes les collisions avec une ambulance tout en évitant d'écraser un maximum de chats, sans pour autant fixer de seuil car cet objectif est jugé bien moins important.

Deux approches sont alors possibles : interdire les actions qui représentent un risque avant de choisir l'action la plus optimale parmi celles qui sont autorisées [14] ou vérifier le danger de l'action sélectionnée par notre politique, et l'interdire si il est trop grand en maintenant sa présence sur la même ligne au prochain pas de temps [12]. La première agit en amont de la décision prise par notre agent en filtrant les actions qui ne satisfont pas notre contrainte. La seconde agit plutôt comme un contrôleur logique en bloquant l'action après qu'elle ait été choisie. Pour avoir la possibilité de choisir une action qui soit la plus optimale même dans une situation défavorable, c'est la première approche que nous avons choisi d'adopter.

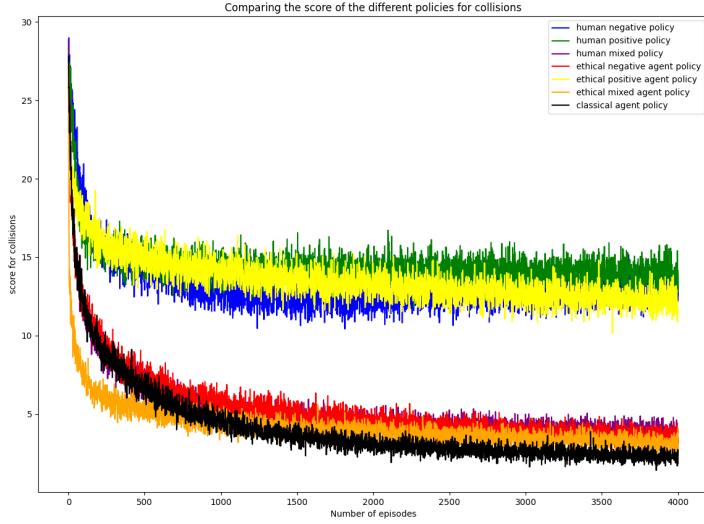
La notion d'états interdits ou de contraintes strictes avaient déjà été introduite par l'article [13]. Cependant c'est une solution qui a été développée pour des problèmes *Off-line* et qui demande une connaissance complète des



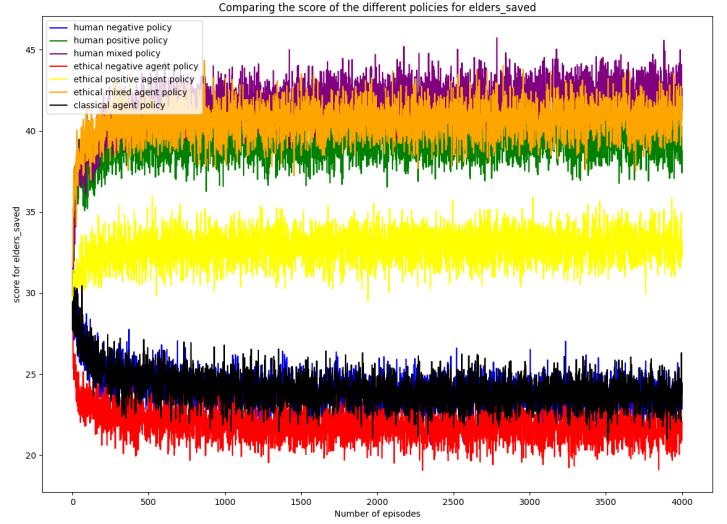
(a) Evolution des **ambulances percutées** (à minimiser)



(b) Evolution des **chats écrasés** (à minimiser)



(c) Evolution des **collisions** (à minimiser)



(d) Evolution des **personnes agées sauvées** (à maximiser)

FIGURE 6 – Evolution en pourcentage du nombre de collisions avec des ambulances(a), du nombre de chats blessés percutés(b), du nombre de collisions avec d'autres véhicules(c), et du nombre de personnes agées démentes sauvées(d) au cours de l'apprentissage pour des expériences de 300 pas. Moyenne réalisée sur 100 runs d'apprentissage, tel que  $c_n = 1.00$ ,  $\mathcal{T}_n = 0.20$ ,  $c_p = 2.00$  et  $\mathcal{T}_p = 0.50$ .

états possibles et de la fonction de transition. Mais l'approche proposée nous a tout de même fortement inspiré dans la construction de notre loi visant à éviter les ambulances. La seule information demandée est le seuil auquel l'intensité devient trop grande pour ne pas être considérée au niveau de l'horizon de l'agent, et donc dangereuse. Il est fixé à 8 dans nos cas. Il peut être fixé plus bas que nécessaire pour augmenter la taille de la zone de risque, mais cela peut jouer drastiquement sur les performances de notre agent.

Grâce à la valeur de l'intensité de la sirène et sa dérivée contenue dans l'état, on extrapole la valeur de celle-ci en lui ajoutant deux fois sa dérivée afin de projeter la valeur qu'elle aura atteinte sur les deux prochains pas de temps. Nous avons mené des expériences en amont avec un seul pas de temps, mais les résultats se sont montrés insuffisants car l'agent était déjà trop proche du danger pour pouvoir l'éviter. On cherche à la fois à éviter une

collision au prochain tour mais aussi à ne pas nous retrouver dans un état dit "bloquant" qui qu'importe l'action choisie nous mènerait à un état interdit (collision avec une ambulance). Ainsi si dans le prochain ou second pas de temps, une action nous mène à dépasser le seuil donné tout en se trouvant sur une des deux voies de gauche simultanément, alors nous l'interdisons. C'est à partir de ce sous-ensemble d'actions que notre agent est autorisé à choisir l'action qui lui semble la plus optimale. Si aucune action n'est disponible, alors il choisit parmi toutes les actions initialement possibles, et une infraction est comptabilisée.

Nous avons réalisé cette expérience en employant deux types de politiques humaines : l'une prenant déjà en compte l'évitement des ambulances présentée dans la section précédente, et l'autre n'en tenant pas compte. On remarque alors que les résultats sont aussi bons dans un cas comme dans l'autre, et qu'il n'est donc pas nécessaire que la politique humaine prenne en compte cette nouvelle contrainte qui sera contrôlée par la loi. Nous rappelons que cette observation tient aussi de part l'alignement des objectifs entre eux, car la règle aurait pu totalement écraser les performances de notre agent sur d'autres critères qui entrent en opposition avec elle. Le tableau 3 regroupe les performances réalisées en fin d'apprentissage avec chacune des deux politiques humaines. Les graphes correspondants aux performances des différents critères sont mis à disposition en annexe (figures 9 & 10).

Score	Politique humaine avec prise en compte des ambulances	Politique humaine sans prise en compte des ambulances
Ambulances percutées	0 %	0 %
Chats écrasés	7.5 %	7 %
Personnes agées sauvées	41 %	42.5 %
Collisions	3.5 %	3.5 %
Récompense cumulée	76	78
Infractions	0	0
Prix de la moralité $\psi$	8	9

TABLE 3 – Comparaison de la seconde méthode **Rule-based** en employant deux types de politique humaine pour l'*Ethics Shaping*.

Dans le tableau 3, en plus de souligner l'égalité des performances des deux expériences, nous introduisons la notion de "*Prix de la moralité*". Ce nouveau critère est extrait de l'article [13]. Il représente la différence maximale de récompense totale espérée entre la politique optimale amorphe, et la nouvelle politique morale. Elle permet de mesurer à quel point la politique que nous construisons par *Ethics Shaping* s'éloigne de la politique qui remplit notre objectif principal sans prendre en compte aucune loi éthique, notre politique **amorphe**.

Contrairement au papier présentant cette mesure, nous n'utilisons pas une fonction d'état  $V$  dans notre apprentissage mais une fonction d'état-action  $Q$ . Nous avons donc redéfini le prix de la moralité  $\psi$  pour qu'il s'applique à notre usage tel que :

$$\psi = \max_{s \in S} |Q^{\pi_\rho^*} - Q^{\pi^*}|_{a=a^*}$$

avec

- $Q^{\pi_\rho^*}$  : la fonction d'état-action de la politique générée par *Ethics Shaping* qui est supposée **optimale et morale**
- $Q^{\pi^*}$  : la fonction d'état-action de la politique classique, dont l'objectif est d'uniquement éviter les collisions avec les autres véhicules dans notre cas, qui est supposée **optimale et amorphe**
- $a^*$  : l'action optimale sélectionnée par la politique classique, pour un état  $s$  on a :

$$a^* = \operatorname{argmax}_{a \in A} Q^{\pi^*}(s, a)$$

Le prix de la moralité correspond donc dans notre cas à la plus grande différence entre nos deux fonctions d'état-action pour un même état pour l'action optimale choisie par la politique amorphe. Comme on peut voir grâce à la figure 7, malgré des politiques humaines avec un prix de la moralité élevé, la politique éthique générée est toujours la plus proche de la politique classique amorphe. Cela confirme l'idée que cette approche génère des politiques qui vont conserver l'objectif initial en apportant des modifications autour de lui.

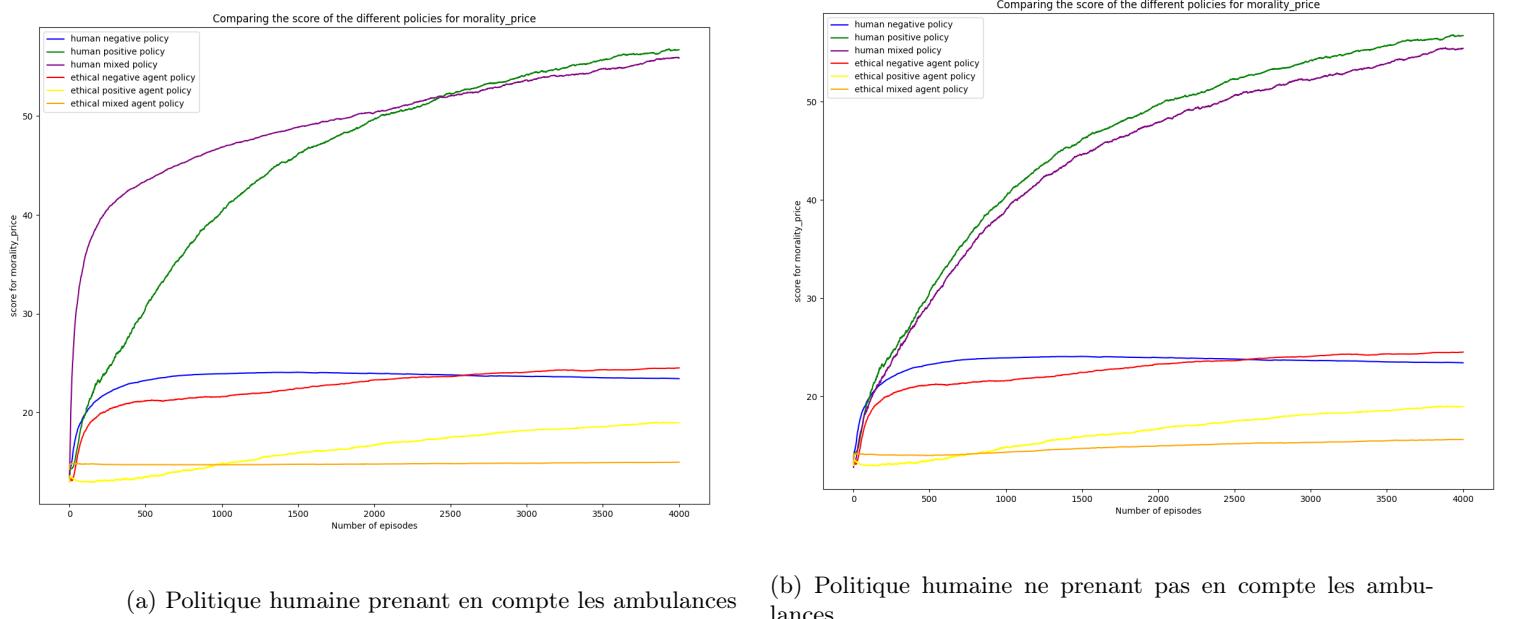


FIGURE 7 – Evolution du prix de la moralité au cours de l'apprentissage pour des expériences de 300 pas. Moyenne réalisée sur 100 runs d'apprentissage, tel que  $c_n = 1.00$ ,  $\mathcal{T}_n = 0.20$ ,  $c_p = 2.00$  et  $\mathcal{T}_p = 0.50$ .

De plus, on peut remarquer que la politique humaine, qu'elle prenne ou non en compte l'évitement des ambulances dans son apprentissage, ressemble d'avantage à la politique humaine positive qui prend en compte uniquement le sauvetage des personnes agées. On peut en déduire que c'est la contrainte liée aux personnes agées qui est dominante face aux autres contraintes prises en compte.

Pour finir, cette seconde méthode présente aussi un réel avantage en complétant grâce à des lois des données qui peuvent être incomplètes sur des objectifs recherchés. Cela nous permet de pouvoir réutiliser la politique humaine du scénario *Driving and Avoiding and Rescuing* tout en obtenant des résultats similaires sur la politique finale de l'agent. Elle peut à la fois venir remplacer ou compléter des contraintes de notre modèle.

### 3.5 Réinterprétation du problème dans un contexte d'apprentissage par renforcement multi-objectif

Nous avons présenté dans les sections précédentes les performances obtenues grâce à l'application de la méthode par *Ethics Shaping* proposée par l'article [22] et ses dérivées. Malgré des résultats très satisfaisants, elle laisse de côté un point important que nous avions abordé en introduction : la transparence des règles intégrées par la politique de notre agent. En effet, elle libère du développeur la lourde tâche de définir explicitement les comportements éthiques à adopter mais de cette manière nous n'avons pas de définition précise de ceux-ci.

C'est dans l'optique de vérifier que cette méthode permet d'atteindre des performances proche du cas idéal où les conditions éthiques peuvent être clairement définies et exprimées que nous avons choisi de reprendre le même problème mais cette fois-ci avec une approche multi-objective. Nous avons implémenté les méthodes proposées par l'article [15] pour répondre à cette nouvelle problématique.

Nous avons dû adapter notre modèle initial en transformant notre signal de récompense  $R$  initialement représenté par un scalaire en un vecteur où chaque composante est propre à un objectif. Par extension, les valeurs prises par la fonction d'état-action  $Q$  prennent aussi une forme vectorielle au lieu de celle d'un simple scalaire. Ainsi chaque objectif est amené à évoluer individuellement et n'est pas invisibilisé par une somme. Par cette transformation, nous exprimons notre modèle sous la forme d'un MOMDP et non plus celle d'un MDP. L'ensemble des états et des actions du modèle de départ sont quant à eux conservés. Les objectifs sont alors fournis directement à l'agent, qui n'est plus influencé par une politique humaine.

Après cette modification, deux approches sont alors proposées : une première qui construit une politique unique qui est une extension directe de notre méthode de résolution précédente étendue à notre modèle multi-objectif, et une seconde qui va contruire un ensemble de politiques qui sont Pareto dominantes que l'agent pourra sélectionner. Ces deux approches et leurs résultats sont présentés dans les sous-sections suivantes.

Afin de favoriser l'exploration au début de l'apprentissage, pour chacune des deux approches, notre agent va sélectionner une action de manière  $\epsilon$ -gloutonne, c'est-à-dire qu'une valeur est tirée entre 0 et 1, et si elle est inférieure à  $\epsilon$  alors notre agent sélectionne l'action qu'il définit comme la plus optimale, sinon il la sélectionne aléatoirement. Nous avons fixé  $\epsilon$  à 0.9, et sa valeur décroît proportionnellement au nombre d'épisodes écoulés.

### 3.5.1 Politique unique

Pour mettre à jour les valeurs de notre fonction  $\widehat{Q}$  d'état-action désormais définie par un vecteur pour chaque couple, nous réutilisons la formule utilisée pour l'apprentissage d'un seul objectif en l'appliquant pour les différents objectifs. Cela nous donne la pour formule suivante :

$$\forall o \in \mathcal{O}, \quad \widehat{Q}_o(s, a) = \widehat{Q}_o(s, a) + \alpha(r_o + \gamma \widehat{Q}_o(s', a') - \widehat{Q}_o(s, a))$$

avec  $\mathcal{O}$  l'ensemble des objectifs pris en compte par notre agent,  $\alpha$  le taux d'apprentissage,  $r_o$  la récompensé liée à cette objecitif dans le vecteur de récompense reçu,  $s'$  l'état dans lequel se trouve l'agent en exécutant l'action  $a$  à partir de  $s$ , et  $a'$  l'action sélectionnée dans l'état  $s'$ .

La plus grande différence apportée par cette approche, en plus d'expliciter les objectifs de l'agent, est la méthode de sélection d'une action. Nous disposons désormais d'une estimation du gain espéré sur chacun des objectifs qui évoluent avec notre agent indépendemment les uns des autres. L'action optimale  $a*$  dans un état  $s$  est définie comme l'action  $a$  avec la plus petite distance pondérée  $d$  entre sa solution  $\widehat{Q}(s, a)$  et le **point idéal**  $\widehat{Q}^{ideal}$ . Le point idéal correspond aux meilleures valeurs de  $\widehat{Q}_o(s, a)$  rencontrées lors de l'apprentissage sur chacun des objectifs, en majorant cette valeur d'une constante  $\tau$ . Il est en opposition avec le point Nadir. Le but ce point de référence est de faire converger notre politique vers ce point qui évolue au cours de l'apprentissage. Cette distance est évaluée par la formule suivante :

$$d(\widehat{Q}(s, a), \widehat{Q}^{ideal}) = \sqrt{\sum_{o \in \mathcal{O}} w_o * (\widehat{Q}_o(s, a) - \widehat{Q}_o^{ideal})^2}$$

avec  $w_o$  le poids associé à un objectif  $o$ . L'ensemble des poids utilisés pour cette expérience est représenté dans le tableau 4. Leur somme sur tous les objectifs pour une politique donnée est toujours égale à 1.

Nous avons aussi mené des expériences avec la mesure de *Chebyshev* qui était recommandée par le papier et qui ne retient que la valeur la plus grande de  $w_o * |\widehat{Q}_o(s, a) - \widehat{Q}_o^{ideal}|$  parmis tous les objectifs. Malheureusement son utilisation mène à une négligence des collisions car les récompenses ne sont pas toujours du même ordre de grandeur selon les objectifs, particulièrement pour les ambulances qui ont un poids négatif deux fois plus grand. Cette norme pondérée par rapport au point idéal offre des résultats plus uniformes sur l'ensemble des objectifs.

Nous avons choisis d'évaluer deux cas. Un premier où l'évitement des ambulances est traité comme un objectif de notre modèle, c'est-à-dire qu'il est contenu dans le vecteur de récompense  $\mathbf{r}$  et ainsi, aussi dans celui de notre fonction  $\widehat{Q}$ . De cette manière, la valeur prise par cette variable n'est pas bornée et est optimisée de la même manière que les autres objectifs. Dans le second cas, on combine notre approche **multi-objective** avec la méthode **Rule-based** appliquée pour l'évitement des ambulances comme nous l'avions fait précédemment. Dans ce cas, l'évitement des ambulances n'est pas considéré parmi les objectifs optimisé et est traité comme une contrainte. Ce sont les résultats de ce second cas qui sont présentés en figure 8 car ils sont similaires au premier cas tout en ayant un taux

Type de la politique	Objectif	Poids sans règle fixe sur les ambulances	Poids avec règle fixe sur les ambulances
Politique de l'agent	Evitement des collisions	0.9	0.9
	Conduire droit	0.1	0.1
Politique humaine <i>Driving and Avoiding</i>	Evitement des chats	0.4	0.4
	Evitement des collisions	0.5	0.5
	Conduire droit	0.1	0.1
Politique humaine <i>Driving and Rescuing</i>	Sauvetage des personnes agées	0.4	0.4
	Evitement des collisions	0.5	0.5
	Conduire droit	0.1	0.1
Politique humaine <i>Driving and Avoiding and Rescuing</i>	Evitement des ambulances	0.3	0
	Evitement des collisions	0.3	0.4
	Sauvetage des personnes agées	0.2	0.3
	Evitement des chats	0.15	0.25
	Conduire droit	0.05	0.05

TABLE 4 – Tableau regroupant les poids utilisés pour le calcul de la distance pondérée au point idéal  $\hat{Q}^{ideal}$  pour la résolution du problème sous forme multi-objective avec une politique unique.

de collision nul avec des ambulances contre 4% en fin d'apprentissage pour le premier cas.

Les résultats montrent que dans les mêmes conditions, c'est-à-dire pour des expériences de même longueur (300 pas) et le même nombre d'épisode (4000), les performances obtenues sont inférieures à celles de notre premier modèle. Nous n'atteignons pour notre politique éthique qu'un score de 12,5% pour l'évitement des chats (figure 8a), et de 7% pour la collision avec les autres véhicules (figure 8b). Pour le sauvetage des personnes agées (figure 8c), le score est quant à lui presque identique avec 42% de personnes sauvées sur toutes celles intégrées à l'environnement. Un apprentissage plus long ou une modification des poids des objectifs pourraient être responsables de ces résultats puisque contrairement à notre approche par *Reward Shaping*, la politique éthique n'est pas guidée par la politique humaine qui elle est déjà entraînée dans le même environnement et lui apporte une aide non négligeable.

### 3.5.2 Politique multiple

Nous reprenons la méthode "*Pareto Q-learning*" présentée dans l'article [15] qui est à l'initiative de la transformation de notre MDP initial en un nouvel MOMDP. Contrairement à la méthode visant à développer une politique unique, elle va construire lors d'un apprentissage en ligne un ensemble de politiques Pareto dominantes. Une fois cet ensemble calculé, il est possible de traquer aisément la politique que l'on souhaite exécuter. La différence avec notre environnement est que cette méthode est initialement appliquée sur un environnement déterministe, c'est-à-dire qu'on ne peut arriver que dans un unique état  $s'$  lorsque l'on exécute l'action  $a$  dans l'état  $s$  ( $\mathcal{T}(s, a, s') = 1$ ). Nous avons essayé dans notre implémentation de revaloriser cette méthode pour pouvoir l'appliquer à notre environnement stochastique pourvu d'une fonction de transition  $\mathcal{T}$  inconnue.

La structure du processus d'apprentissage diffère drastiquement de notre approche initiale puisqu'elle sépare l'apprentissage de la récompense immédiate moyenne  $\bar{\mathcal{R}}(s, a, s')$  et celui de l'ensemble des solutions non-dominées  $\mathcal{ND}(s, a, s')$  qui sont accessibles depuis notre état  $s'$  obtenu en exécutant l'action  $a$  dans l'état  $s$ . Dans leur approche, l'ensemble  $\mathcal{ND}(s, a, s')$  était dépendant du temps écoulé depuis le début de l'expérience, ou du nombre de frames écoulées dans un espace de temps discret. Nous avons choisi de rendre cet ensemble indépendant du temps car dans notre cas nous n'avons pas d'état de fin contrairement à leur expérience, et nos expériences sont uniquement limitées par la taille que nous leur fixons.

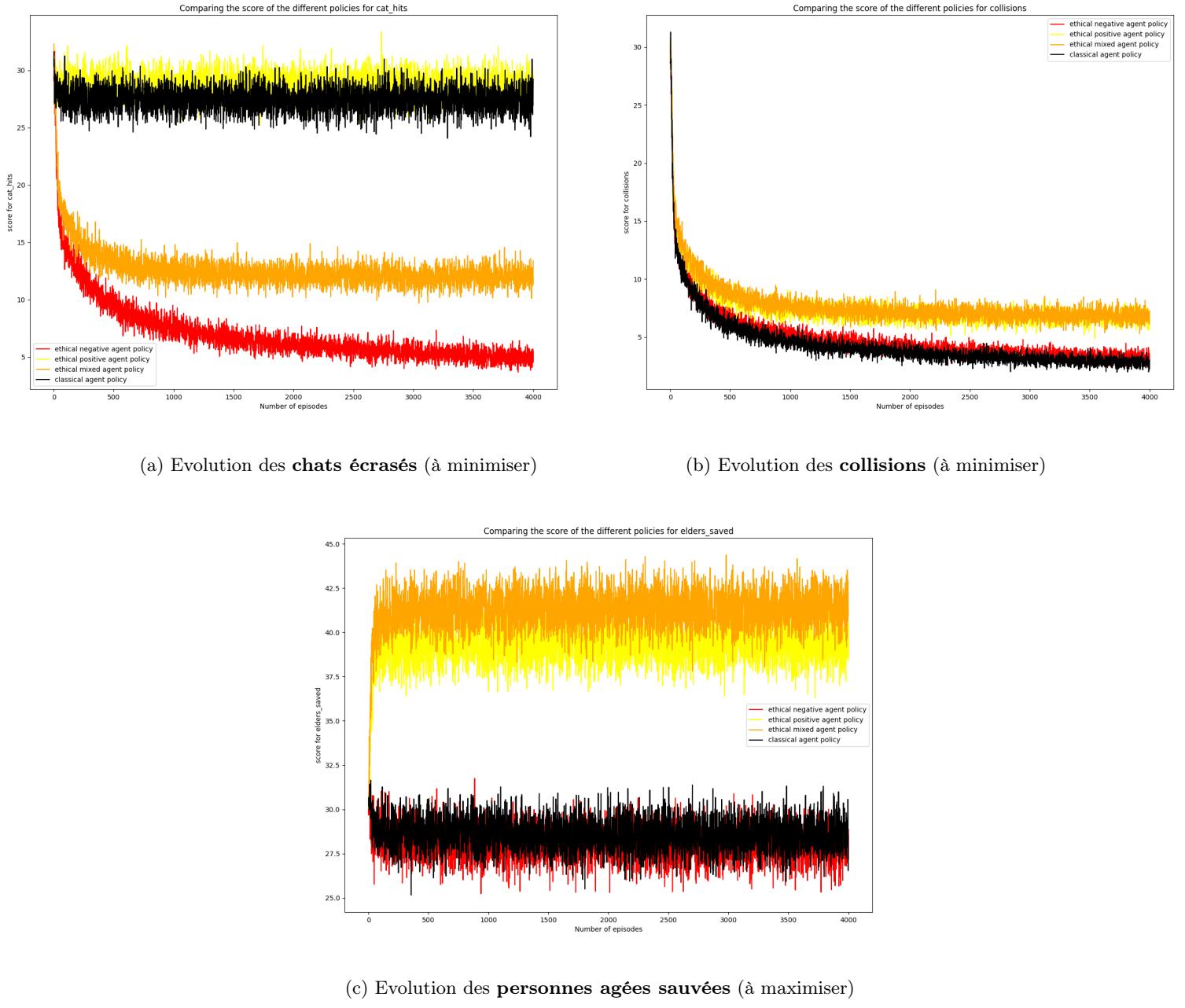


FIGURE 8 – Evolution en pourcentage du nombre de chats blessés percutés(a), du nombre de collisions avec d’autres véhicules(b), et du nombre de personnes agées démentes sauvées(c) au cours de l’apprentissage pour des expériences de 300 pas. Moyenne réalisée sur 100 runs d’apprentissage.

Cette caractéristique pourrait être utile si nous souhaitions inférer un comportement qui évolue au cours du temps, comme un changement de priorité dans les objectifs par exemple.

Il est aussi important pour cette approche d’introduire la notion de **dominance de Pareto** qui sera notre outil de comparaison entre les solutions collectées. On a qu’une solution  $x$  domine au sens de Pareto une solution  $y$  si et seulement si :

$$\forall o \in \mathcal{O}, x_o \geq y_o$$

avec  $\mathcal{O}$  l'espace des objectifs.

Les solutions contenues dans notre fonction d'état-action  $\widehat{Q}_{set}$  sont alors mises à jour à chaque pas de temps  $t$  à l'aide des formules suivantes :

$$\begin{aligned}\bar{\mathcal{R}}(s, a, s') &= \bar{\mathcal{R}}(s, a, s') + \frac{\mathbf{r} - \bar{\mathcal{R}}(s, a, s')}{n(s, a, s')} \\ \mathcal{ND}(s, a, s') &= \mathcal{ND}_{\mathcal{P}}\left(\bigcup_{a' \in \mathcal{A}} \bigcup_{s'' \in \mathcal{S}} \widehat{Q}_{set}(s', a', s'')\right) \\ \widehat{Q}_{set}(s, a, s') &= \bar{\mathcal{R}}(s, a, s') \oplus \gamma \mathcal{ND}(s, a, s')\end{aligned}$$

avec  $s'$  l'état dans lequel se trouve notre agent après avoir exécuté l'action  $a$  dans l'état  $s$ ,  $\mathbf{r}$  la récompense qu'il en reçoit,  $n(s, a, s')$  le nombre de fois où  $a$  a été réalisée à partir de l'état  $s$ . L'opérateur  $\mathcal{ND}_{\mathcal{P}}$  implique une réduction de l'ensemble considéré par application de la dominance de Pareto.

On a aussi que  $s''$  valide la condition suivante :  $\mathcal{T}(s', a, s'') \neq 0$ , c'est-à-dire que nous avons rencontré cette transition au moins une fois avant. Si ce n'est pas le cas, la valeur de l'ensemble  $\widehat{Q}_{set}(s', a', s'')$  est forcément vide puisque nous n'avons pas récolter d'informations dessus.

Les actions sont sélectionnées par notre agent en suivant la procédure qui a montré les meilleures performances dans l'article : par **évaluation de l'ensemble de Pareto**, *PO-PQL*. Quand notre agent se trouve dans l'état courant  $s$ , nous construisons pour chaque action réalisable l'ensemble  $NDQ_{set}(s, a)$  qui regroupe les solutions qui ne sont pas Pareto dominées parmi tous les états accessibles  $s'$  après l'exécution de l'action. On a alors que :

$$NDQ_{set}(s, a) = \mathcal{ND}_{\mathcal{P}}\left(\bigcup_{s'} \widehat{Q}_{set}(s, a, s')\right) \quad \text{s.t. } s' \in \mathcal{S} \text{ and } \mathcal{T}(s, a, s') \neq 0$$

Une fois que nous avons déterminé les ensembles  $NDQ_{set}(s, a)$  pour chacune des actions de notre ensemble  $\mathcal{A}$ , nous appliquons la dominance de Pareto entre chacun d'eux. L'agent choisit alors aléatoirement parmi les ensembles  $Q_{set}(s, a)$  qui ne sont pas vides après cette dernière opération. Nous ne sélectionnons pas l'action qui possède l'ensemble de solutions non Pareto dominées le plus grand car cette méthode qui évalue cardinalement les ensembles de solutions car elle ne traite pas équitablement chaque solution non-dominée.

Nous nous sommes rapidement rendu compte que le nombre de solutions s'accroissait très rapidement, ce qui rendait impossible l'exécution de l'apprentissage au bout de seulement une centaine d'épisodes. Nous avons donc tenté d'approcher les résultats tout en limitant le nombre de solutions retenues en appliquant une méthode de comparaison plus souple : la  $\epsilon$ -dominance de Pareto. On a désormais qu'une solution  $x$   $\epsilon$ -domine au sens de Pareto une solution  $y$  si et seulement si :

$$\forall o \in \mathcal{O}, (1 + \epsilon) x_o \geq y_o$$

Nous avons aussi retenu uniquement les  $k$ -meilleurs cas pour un ensemble  $\widehat{Q}_{set}(s, a, s')$  en retenant les  $k$  solutions qui avaient la plus petite distance au point idéal. Ces deux méthodes nous ont permis de considérablement réduire le temps de calcul de notre apprentissage, mais il n'en reste pas moins très important en comparaison des méthodes précédentes. Nous n'avons donc ainsi pas pu produire de résultats concluants à partir de cette méthode qui demande des ressources bien plus importantes.

## 4 Critique

D'après les résultats que nous avons obtenus pendant ce stage, l'approche par *Ethics Shaping* est celle qui nous permet d'atteindre les meilleures performances. Cependant comme nous l'avions souligné plus tôt, ses résultats sont extrêmement dépendants de la qualité de la politique humaine fournie. La politique éthique va s'aligner sur les performances de la politique humaine. Ainsi si les données fournies sont sous-optimales alors notre politique humaine le sera, et la politique éthique aussi. Il est donc primordiale de contrôler la qualité des données d'entrée avant de les utiliser, en les transformant en une politique humaine et en évaluant ses performances. Combiner cette approche avec des règles fixes est aussi un moyen de palier au manque d'optimalité de certains critères, et même de compléter la politique humaine.

L'approche proposée par l'*Ethics Shaping* propose d'utiliser des données qui adoptent un comportement généralement éthique mais qui peut être éloigné du but initial. Hors nous avons remarqué que l'alignement des objectifs entre eux joue un rôle important, et que donc des données trop déconnectées de l'objectif initial ne permettent pas d'assurer cet alignement. En effet, les préférences modélisées par les données ne sont pas explicites, et peuvent entrer en contradiction avec nos objectifs. Des données trop générales peuvent aussi masquer des préférences moins représentées mais existantes qui sont indésirables. Il est donc important de pouvoir évaluer la qualité des données, mais aussi le sens des préférences qu'elles expriment. La transparence des données utilisées est un enjeu important de l'intelligence artificielle.

Pour finir sur la critique de cette méthode, il est demandé au développeur de définir lui-même l'état éthique qui est employé par la politique humaine exprimée par les données. Or comme nous avons pu le voir, il peut être assez compliqué de formuler l'état éthique qui correspond à nos attentes et qui nous permettra d'obtenir les meilleurs résultats. En utilisant un état éthique trop général, nous risquons d'appliquer le *Rewards Shaping* plus fréquemment, et la politique de l'agent se rapprochera d'avantage de la politique humaine que de la politique qui optimise l'objectif initial que nous lui avons fixé. Employer l'état général comme état éthique permet de mettre l'humain et l'agent dans le même contexte pour la prise de décision, et nous a permis d'atteindre les performances que nous souhaitions.

Ainsi, si il est possible d'exprimer les contraintes explicitement, il est favorable d'utiliser une approche multi-objective qui est plus transparente et donc plus facilement modifiable et interprétable. Or comme les résultats que nous avons obtenus l'ont montré, travailler avec la politique humaine permet de guider notre agent et facilite la convergence de notre politique vers un comportement souhaité lorsque les données sont de très bonne qualité.

Dernièrement, enrichir la structure de notre fonction  $Q$  pourrait être déterminant puisqu'en enrichissant notre scénario nous avons aussi considérablement agrandi l'espace de recherche. Une structure plus profonde et une approche de type *Deep Q-learning* pourrait offrir de meilleurs résultats.

## 5 Conclusion

Nous avons durant ce stage abordé différentes approches pour résoudre des problèmes éthiques grâce à l'apprentissage par renforcement. Les méthodes de résolution proposées se sont articulées autour d'agent modélisés par des MDPs et MOMDPs ce qui permet de traiter des problèmes assez variés. Les résultats obtenus par *Reward Shaping* ont particulièrement été remarqués, ainsi que la capacité de cette méthode à être étendue et combinée avec des méthodes dites *Rule-based*. Cependant comme le montre la section Critique, de nombreux aspects ont encore besoin d'être améliorés.

Dans un premier temps, il serait nécessaire de construire un outil capable de mesurer la qualité et le sens des données utilisées pour garantir la robustesse de la méthode par *Reward Shaping*. Il est important de pouvoir fournir une justification sur l'emploi de ces données, et de garantir la qualité de la politique générée avant qu'elle ne puisse être déployée.

Il serait aussi nécessaire de poursuivre les recherches sur la méthode de résolution multi-objective qui traitent l'existence de plusieurs politiques dominantes pour laquelle nous n'avons pas pu obtenir de résultats concluants. Il est aussi important de définir une méthode pour sélectionner la politique que l'on souhaite traquer en déterminant si elle doit être équilibrée sur tous les objectifs ou si des besoins stricts sur certains critères sont à prioriser.

Un dernier point serait aussi de travailler sur des apprentissages plus longs pour les méthodes multi-objectives qui prennent plus de temps à converger et demandent plus de ressources afin de conclure sur leur intérêt face à des méthodes plus flexibles et moins contraignantes comme le *Reward Shaping*.

## A Annexes

### A.1 Glossaire

**Décision négativement éthique** C'est lorsque l'action choisie par notre agent a une probabilité d'être sélectionnée supérieure à celle accordée par la politique humaine qui est très faible. C'est-à-dire quand l'agent exécute trop souvent une action dans un état que la politique humaine évite.

**Décision positivement éthique** C'est lorsque l'action choisie par notre agent a une probabilité d'être sélectionnée inférieure à celle accordée par la politique humaine qui est très haute. C'est-à-dire quand l'agent n'exécute pas assez souvent une action dans un état que la politique humaine favorise.

**Loi temporellement complexe** Principe à suivre qui implique une dépendance des états et/ou des actions pendant un certain nombre de pas. C'est-à-dire que par exemple l'état et l'action précédents notre état courant peuvent avoir un impact sur l'action qui va être sélectionnée. L'horizon peut être plus large.

**Point Ideal** Solution regroupant les meilleures valeurs obtenues sur les différents objectifs. Utilisé comme point de référence.

**Point Nadir** Solution regroupant les pire valeurs obtenues sur les différents objectifs. Utilisé comme point de référence.

**Politique amorphe** Politique qui ignore les aspects moraux du comportement qu'elle traduit. Elle ne va pas contre la morale, mais ne la prend simplement pas en compte.

**Politique classique** Politique qui intègre les objectifs liés à la performance uniquement. Elle est supposée optimale et amorphe.

**Politique éthique** Politique qui intègre les objectifs liés à la performance et ceux liés aux principes moraux. Elle est supposée optimale et morale.

**Politique humaine** Politique construite à partir de données retraçant des comportements humains. Elle supposée morale, mais pas forcément optimale. Elle est utilisée pour corriger la politique classique avec le *Reward Shaping*.

**Prix de la moralité** C'est la distance maximale, pondérée ou non, entre la fonction d'état-action  $Q$  de la politique classique et celle d'une autre politique parmi tous les états  $s$  appartenant à  $\mathcal{S}$  lorsque l'action optimale  $a^*$  d'après la politique classique est sélectionnée.

## A.2 Tableaux

Type de la politique humaine "éthique"	Etat éthique	Résultats du <i>Ethics Shaping</i>
Politique humaine <b>négative</b> (évitement des chats)	$(c_0, c_1, c_2)$	L'évitement des chats est bien transféré à notre agent tout en limitant son impact sur l'évitement des autres véhicules. Les performances obtenues sont proches des valeurs optimales. Les deux objectifs s'alignent.
Politique humaine <b>positive</b> (sauvetage des personnes agées)	$(e_0, e_1, e_2)$	Le sauvetage des personnes agées est moyennement transféré à notre agent avec un fort impact négatif sur l'évitement des autres véhicules. Les performances obtenues sont assez éloignées des valeurs optimales. Les deux objectifs ont du mal à s'aligner.
Politiques humaines <b>négative + positive</b> (utilisation des 2 politiques précédentes simultanément)	$(c_0, c_1, c_2)$ et $(e_0, e_1, e_2)$	Les performances observées sont inférieures à celles obtenues individuellement par les politiques précédentes. Combiner les objectifs des politiques humaines sans changer la représentation de l'état éthique ne leur permet pas de s'aligner puisqu'ils sont représentés dans des référentiels distincts. Les performances obtenues sont insuffisantes.
Politique humaine <b>mixte</b> (combinaison des 2 objectifs précédents dans une seule politique)	$(c_0, e_0, c_1, e_1, c_2, e_2)$ ,	Même si on observe une nette amélioration sur l'évitement des chats et le sauvetage des personnes agées avec des performances similaires à celles de la politique humaine, l'évitement des collisions avec les autres véhicules est totalement négligé par la politique résultante même si elle est bien prise en compte par la politique humaine. Le romodelage par <i>Ethics Shaping</i> a désaligné notre agent de son objectif initial d'éviter les collisions en renforçant celui d'éviter les chats.
Politique humaine <b>mixte</b> (combinaison des 2 objectifs précédents dans une seule politique)	$(v_0, c_0, e_0, v_1, c_1, e_1,$ $v_2, c_2, e_2)$	La politique résultante prend en compte les divers objectifs fixés par la politique humaine et la politique classique. Elle évite les chats et sauve les personnes agées avec des performances très proches de la politique humaine, tout en ayant des performances très proches de notre politique optimale pour l'évitement des autres véhicules. Ce modèle obtient les <b>meilleures performances</b> pour cette expérience.

TABLE 5 – Tableau résumant les différentes politiques humaines avec leur état éthique utilisées pour résoudre l'expérience *Driving and Avoiding and Rescuing*.

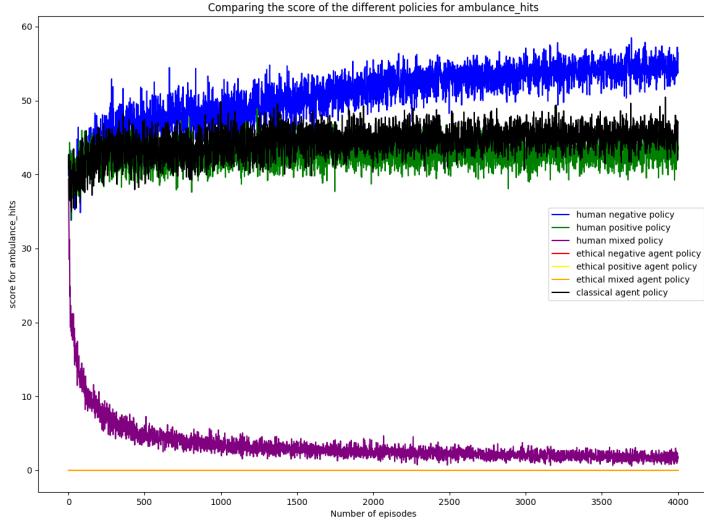
Type de la politique	Couleur	Evitements des collisions	Evitements des chats	Sauvetage des personnes agées	Apprentissage par par <i>Ethics Shaping</i>
Politique de l'agent	noir	X			
Politique humaine <i>Driving and Avoiding</i>	bleu		X		
Politique humaine <i>Driving and Rescuing</i>	vert			X	
Politique humaine <i>Driving and Avoiding and Rescuing</i>	violet		X	X	
Politique éthique <i>Driving and Avoiding</i>	rouge	X	X		X
Politique éthique <i>Driving and Rescuing</i>	vert	X		X	X
Politique éthique <i>Driving and Avoiding and Rescuing</i>	orange	X	X	X	X

TABLE 6 – Tableau regroupant les légendes et les objectifs des différentes politiques construites lors de nos expériences.

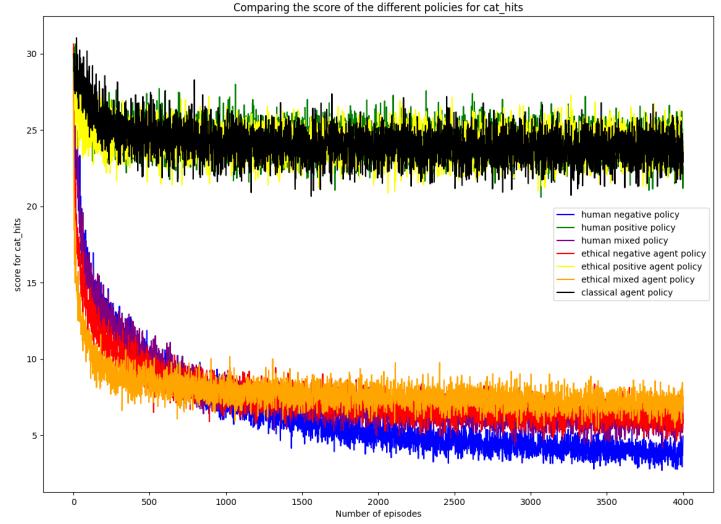
## A.3 Figures

### A.3.1 Performances pour la loi stricte sur l'ambulance

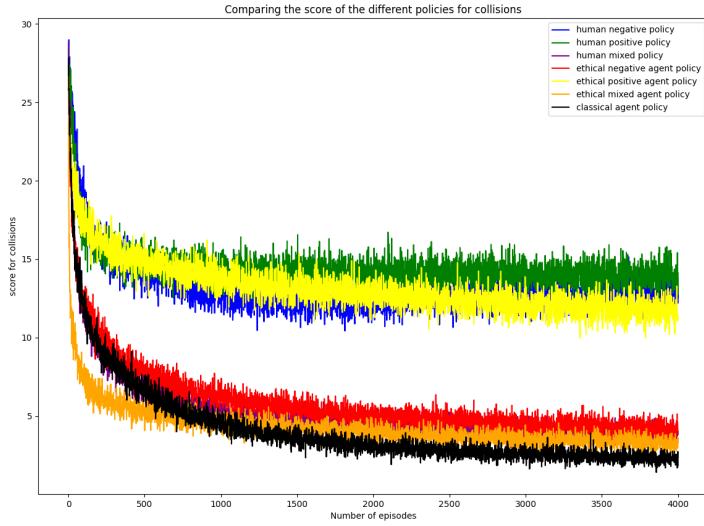
Politique humaine avec prise en compte des ambulances



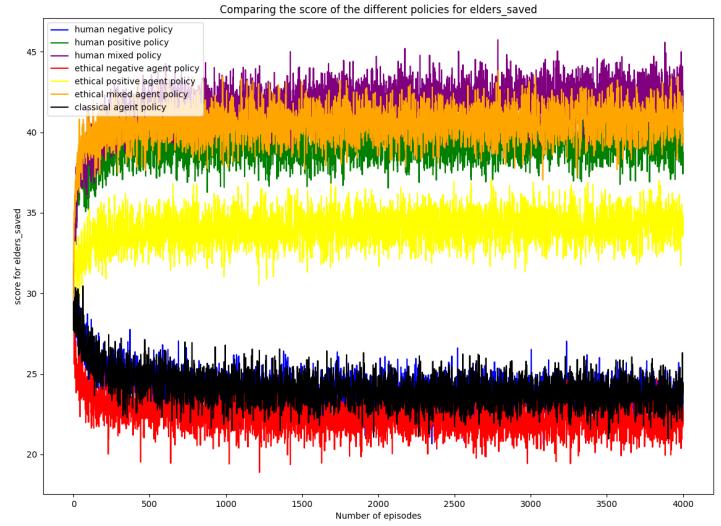
(a) Evolution des **ambulances percutées** (à minimiser)



(b) Evolution des **chats écrasés** (à minimiser)



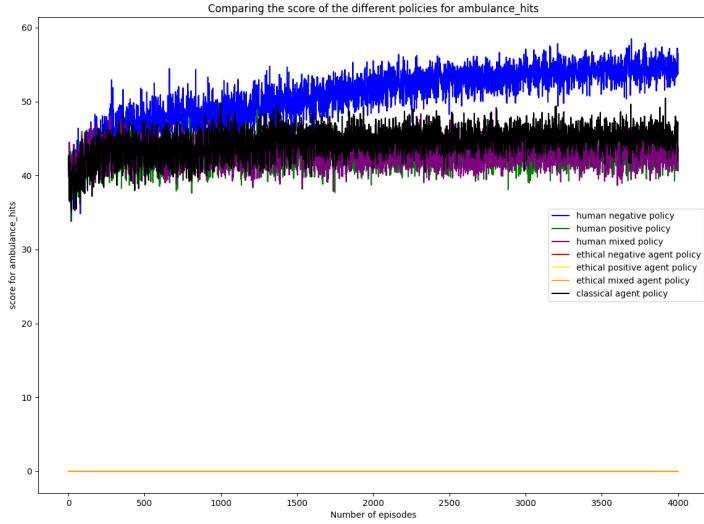
(c) Evolution des **collisions** (à minimiser)



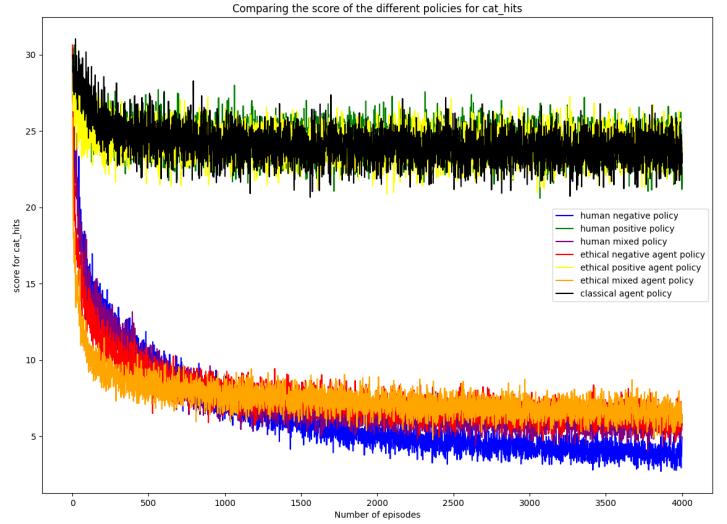
(d) Evolution des **personnes âgées sauvées** (à maximiser)

FIGURE 9 – Evolution en pourcentage du nombre de collisions avec des ambulances(a), du nombre de chats blessés percutés(b), du nombre de collisions avec d'autres véhicules(c), et du nombre de personnes âgées démentes sauvées(d) au cours de l'apprentissage pour des expériences de 300 pas. Moyenne réalisée sur 100 runs d'apprentissage, tel que  $c_n = 1.00$ ,  $\mathcal{T}_n = 0.20$ ,  $c_p = 2.00$  et  $\mathcal{T}_p = 0.50$ .

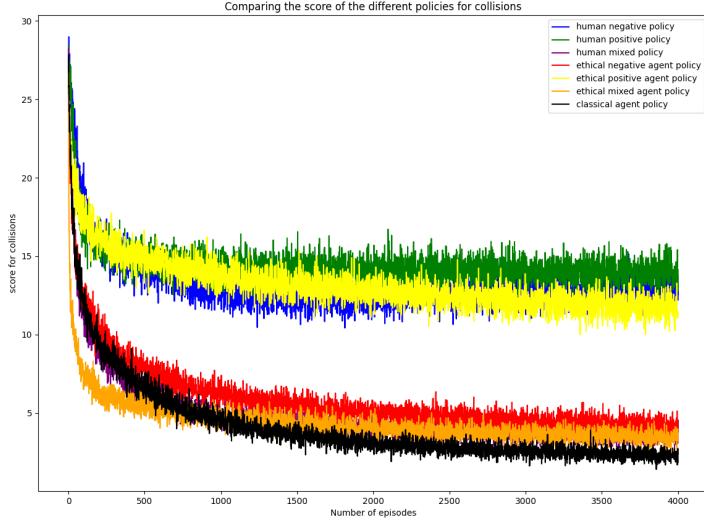
## Politique humaine sans prise en compte des ambulances



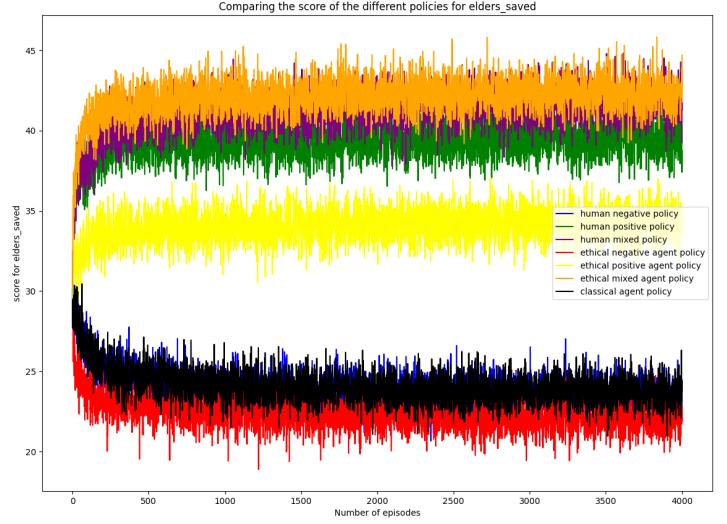
(a) Evolution des **ambulances percutées** (à minimiser)



(b) Evolution des **chats écrasés** (à minimiser)



(c) Evolution des **collisions** (à minimiser)



(d) Evolution des **personnes agées sauvées** (à maximiser)

FIGURE 10 – Evolution en pourcentage du nombre de collisions avec des ambulances(a), du nombre de chats blessés percutés(b), du nombre de collisions avec d'autres véhicules(c), et du nombre de personnes agées démentes sauvées(d) au cours de l'apprentissage pour des expériences de 300 pas. Moyenne réalisée sur 100 runs d'apprentissage, tel que  $c_n = 1.00$ ,  $\mathcal{T}_n = 0.20$ ,  $c_p = 2.00$  et  $\mathcal{T}_p = 0.50$ .

## Références

- [1] David Abel, J. MacGlashan, and M. Littman. Reinforcement learning as a framework for ethical decision making. In *AAAI Workshop : AI, Ethics, and Society*, 2016.
- [2] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety, 2016.
- [3] Mauricio Araya-López, Vincent Thomas, Olivier Buffet, and François Charpillet. A Closer Look at MOMDPs. In *22nd International Conference on Tools with Artificial Intelligence - ICTAI 2010*, Proceedings of the 22nd International Conference on Tools with Artificial Intelligence, Arras, France, October 2010. IEEE.
- [4] Thomas Arnold, Daniel Kasenberg, and Matthias Scheutz. Value alignment or misalignment – what will keep systems accountable? In *AAAI Workshop on AI, Ethics, and Society*, 2017.
- [5] Djallel Bounedjouf and Irina Rish. A survey on practical applications of multi-armed and contextual bandits. *IJCAI 2019, the 28th International Joint Conference on Artificial Intelligence*, 2019.
- [6] Djallel Bounedjouf, Irina Rish, Guillermo Cecchi, and Raphaël Féraud. Context attentive bandits : Contextual bandit with restricted context. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 1468–1475, 2017.
- [7] Ronen I. Brafman and Giuseppe De Giacomo. Regular decision processes : A model for non-markovian domains. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5516–5522. International Joint Conferences on Artificial Intelligence Organization, 7 2019.
- [8] Yang Gao, Huazhe Xu, Ji Lin, Fisher Yu, Sergey Levine, and Trevor Darrell. Reinforcement learning from imperfect demonstrations. *ICLR 2018 Conference Blind Submission*, 02 2018.
- [9] Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles L Isbell, and Andrea L Thomaz. Policy shaping : Integrating human feedback with reinforcement learning. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.
- [10] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.
- [11] Mingxuan Jing, Xiaojian Ma, Wenbing Huang, Fuchun Sun, Chao Yang, Bin Fang, and Huaping Liu. Reinforcement learning from imperfect demonstrations under soft expert guidance. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04) :5109–5116, Apr. 2020.
- [12] Wang Junjie, Qichao Zhang, Dongbin Zhao, and Chen Yaran. Lane change decision-making through deep reinforcement learning with rule-based constraints. pages 1–6. 2019 International Joint Conference on Neural Networks (IJCNN), 07 2019.
- [13] Shlomo Zilberstein Justin Svegliato, Samer B. Nashed. Ethically compliant sequential decision making. In *Proceedings of the Thirty-Fifth Association for the Advancement of Artificial Intelligence, AAAI-21*, 2021.
- [14] Amarildo Likmeta, Alberto Maria Metelli, Andrea Tirinzoni, Riccardo Giol, Marcello Restelli, and Danilo Romano. Combining reinforcement learning with rule-based controllers for transparent and general decision-making in autonomous driving. *Robotics and Autonomous Systems*, 131 :103568, 2020.
- [15] Kristof Van Moffaert and Ann Nowé. Multi-objective reinforcement learning using sets of pareto dominating policies. *Journal of Machine Learning Research*, 15(107) :3663–3692, 2014.
- [16] Ritesh Noothigattu, Djallel Bounedjouf, Nicholas Mattei, Rachita Chandra, Piyush Madan, Kush R. Varshney, Murray Campbell, Moninder Singh, and Francesca Rossi. Teaching ai agents ethical values using reinforcement learning and policy orchestration. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 6377–6381. International Joint Conferences on Artificial Intelligence Organization, 7 2019.
- [17] Syed Raza, Benjamin Johnston, and Mary-Anne Williams. Reward from demonstration in interactive reinforcement learning. Twenty-Ninth International Florida Artificial Intelligence Research Society Conference, 01 2016.
- [18] Francesca Rossi and Nicholas Mattei. Building ethically bounded ai. The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19), 12 2018.
- [19] Renata Saraiva Perkusich, João Nunes, Mirko Perkusich, Hyggo Almeida, and Cláudivton Siebra. A hybrid approach using case-based reasoning and rule-based reasoning to support cancer diagnosis : A pilot study. volume 216, 08 2015.

- [20] Matthias Scheutz. The case for explicit ethical agents. *AI Magazine*, 38(4) :57–64, Dec. 2017.
- [21] Yueh-Hua Wu, Nontawat Charoenphakdee, Han Bao, Voot Tangkaratt, and Masashi Sugiyama. Imitation learning from imperfect demonstration. Proceedings of the 36th International Conference on Machine Learning, 02 2019.
- [22] Yueh-Hua Wu and Shou-De Lin. A low-cost ethics shaping approach for designing reinforcement learning agents. The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), 12 2018.
- [23] Han Yu, Zhiqi Shen, Chunyan Miao, Cyril Leung, Victor R. Lesser, and Qiang Yang. Building ethics into artificial intelligence. *CoRR*, abs/1812.02953, 2018.