Processus de décision markoviens éthiques

Mihail Stojanovski Nadjet Bourdache Grégory Bonnet Abdel-Illah Mouaddib

Normandie Univ., UNICAEN, ENSICAEN, CNRS, GREYC, Caen, France
{prenom.nom}@unicaen.fr

Résumé

Avec l'introduction d'agents autonomes dans la vie quotidienne, l'intégration de l'éthique dans les processus décisionnels de ces agents devient une question importante. Afin d'avoir un modèle générique fournissant un cadre expressif pour juger un comportement, nous proposons dans cet article les processus de décision markoviens éthiques, (E-MDP pour *Ethical Markov Decision Processes*), qui étendent les MDP classiques avec la représentation explicite des valeurs morales – positives ou négatives – que les décisions des agents peuvent promouvoir ou trahir. Nous proposons un algorithme pour résoudre les E-MDP et nous illustrons notre modèle sur trois cadres éthiques distincts : le cadre de la Théorie du Commandement Divin, le cadre des Devoirs *Prima Facie* et le cadre de l'Éthique de la Vertu, cadres qui raisonnent respectivement sur des états interdits, sur des paires état-action obligatoires et sur des trajectoires exemplaires représentées par des ensembles de transitions. Enfin, nous expérimentons le modèle sur une application de véhicule autonome et évaluons la perte de valeur entre les politiques morales et amorales.

Abstract

With the introduction of autonomous agents in everyday life, the integration of ethics in the decision-making processes of these agents becomes an important issue. In order to have a generic model which provides an expressive framework for judging behavior, we propose in this article the Ethical Markov Decision Processes (E-MDPs), that extend classical MDPs with the explicit representation of moral values – positive or negative – that agents' decisions may promote or demote. We propose an algorithm for solving E-MDPs and illustrate our model on three distinct ethical frameworks: the Divine Command Theory framework, the *Prima Facie* Duties framework, and the Virtue Ethics framework, which respectively focus on prohibited states, mandatory state-action pairs, and exemplary trajectories represented by sets of transitions. Finally, we experiment our model on an autonomous vehicle setting and evaluate the value loss between moral and amoral policies.

1 Introduction

Le développement d'agents autonomes interagissant avec des êtres humains peut poser des problèmes dans certaines situations en l'absence d'une composante éthique. Par exemple, dans le domaine des véhicules autonomes, le choix d'une vitesse de conduite peut être considéré comme plus ou moins éthique selon les circonstances. En effet, il peut être préférable de conduire à vitesse réduite devant une école à l'heure de la fin des cours, plutôt que de respecter simplement la limite de vitesse. Dans ce cas, l'éthique n'est pas une contrainte stricte mais plutôt une recommandation à laquelle l'agent devrait se conformer s'il en a la possibilité.

La littérature propose majoritairement des modèles qualitatifs fondés sur la logique, comme par exemple les cadres d'argumentation valuée [4], la logique modale [11], la logique non monotone [5], ou les architectures BDI [6]. Cependant, dans cet article, nous nous concentrons sur des processus décisionnels quantitatifs qui sont les *processus de décisions markoviens* (MDP). Traiter de l'éthique dans ce type de modèles est relativement nouveau et, à notre connaissance, les approches proposées manquent de généricité. En effet, elles ne permettent souvent de traiter qu'une unique règle parmi un ensemble de règles éthiques, appelées *principes éthiques*, qui indiquent de manière exogène au modèle quelles sont les valeurs morales importantes et quelles sont les décisions autorisées. Le problème est qu'il est nécessaire d'avoir un modèle différent pour chaque principe éthique auquel l'agent doit se conformer, ce qui peut être coûteux en temps et en effort.

Afin d'intégrer plus facilement l'éthique dans le raisonnement des agents autonomes, il est nécessaire de créer un modèle qui soit générique et qui permette d'exprimer avec richesse les comportements éthiques. Nous proposons alors le modèle des *processus de décision markoviens éthiques* (E-MDP) qui étend les MDP avec des valeurs morales explicites et utilise une fonction de récompense multicritère qui représente distinctement la satisfaction (promotion) et violation (trahison) de ces valeurs.

Remarquons qu'une fois une représentation de l'éthique choisie et spécifiée, le problème de décision éthique peut être vu comme un problème d'optimisation multicritère séquentiel. En effet, nous voulons d'abord assurer un comportement où l'agent fait le moins de mal possible, et parmi les comportements obtenus choisir ceux qui font le plus de bien, en accord avec le principe éthique choisi. Dans le cas où l'agent autonome a une tâche à accomplir, le comportement optimal par rapport à la tâche sera choisi parmi les comportements qui minimisent d'abord le mal et maximisent ensuite le bien.

Cet article est structuré comme suit. La section 2 présente un aperçu de l'état de l'art en termes d'intégration de l'éthique dans les MDP ainsi qu'un positionnement philosophique de notre proposition. Nous discutons d'une caractérisation de la prise de décision éthique dans la section 3 et détaillons le modèle E-MDP dans la section 4. Nous introduisons un algorithme pour résoudre les E-MDP en section 5 et illustrons notre modèle avec trois principes éthiques distincts en section 6. Enfin, la section 7 est consacrée à nos résultats expérimentaux.

2 État de l'art

Nous rappelons ici la définition des MDP, ainsi que quelques approches récentes visant à intégrer l'éthique dans ces modèles ; nous présentons ensuite une vision de haut niveau de l'éthique sur laquelle notre proposition est fondée.

2.1 Processus de décision markoviens

Un *processus de décision markovien* [14] est un modèle de prise de décision dans un environnement stochastique entièrement observable. Un MDP est un quadruplet $\langle S, A, \mathcal{T}, R\rangle$ où S est un ensemble fini d'états, A est un ensemble fini d'actions, $\mathcal{T} : S \times A \times S \to [0, 1]$ est une fonction telle que $\mathcal{T}(s, a, s')$ est la probabilité que choisir l'action $a \in A$ dans l'état $s \in S$ produise l'état résultant $s' \in S$, $R : S \times A \times S \to \mathbb{R}$ est une fonction telle que $R(s, a, s')$ est la récompense immédiate que l'agent recevra en arrivant à l'état $s' \in S$ après avoir choisi l'action $a \in A$ dans l'état $s \in S$.

Un solution à un MDP est une fonction, appelée *politique*, qui associe une action (ou un ensemble d'actions) à chaque état du MDP. Il existe plusieurs types de politiques comme des politiques déterministes ou stochastiques [14] ou encore non-déterministes [8]. Les politiques déterministes $\pi : S \to A$ décrivent quelle action doit faire l'agent dans chaque état s tandis que les politiques stochastiques $\pi : S \times A \to [0, 1]$ décrivent pour chaque état s la probabilité de choisir l'action a. Dans cet article, nous nous intéressons en particulier aux politiques non-déterministes. Une telle politique $\pi : S \to 2^{|A|}$ décrit l'ensemble des choix d'action que l'agent peut faire dans chaque état. Une solution optimale d'un MDP est une *politique optimale* π^*, qui maximise la récompense cumulée attendue dans chaque état $s \in S$ en résolvant l'équation de Bellman. Pour la maximisation, l'équation de Bellman prend la forme suivante $\forall s \in S$:

$$V^*(s) = \max_a \sum_{s' \in S} \mathcal{T}(s, a, s')[\mathcal{R}(s, a, s') + \gamma V^*(s')]$$

Les principaux algorithmes pour résoudre un MDP et obtenir des politiques optimales sont *Value Iteration* (VI) [3] et *Policy Iteration* (PI) [10]. Pour nos besoins, nous utilisons une variante de VI qui utilise les Q-valeurs [19].

2.2 Intégrer l'éthique dans les MDP

Plusieurs approches récentes ont été proposées pour intégrer l'éthique dans les MDP. Tout d'abord, certains travaux s'intéressent à l'apprentissage par renforcement, qui consiste à apprendre un comportement par essais et erreurs. Cependant, ces premières approches nécessitent une certaine part d'intervention humaine pour juger de la moralité des actions de l'agent. Par exemple, Abel *et al.* [1] utilisent l'expertise d'un humain pour juger a priori de la moralité de chaque action dans un MDP, tandis que Wu *et al.* [21] utilisent une base de données qui agrège de nombreuses réponses humaines pour un POMDP (MDP partiellement observable). Dans les deux cas, la récompense obtenue par l'agent dépend de la similarité entre la décision humaine et celle de l'agent.

Une autre manière d'intégrer l'éthique dans les MDP est d'utiliser une technique de façonnage de la fonction de récompense – *reward function shaping* – qui consiste à introduire des modifications locales de cette fonction. Pour cela, des récompenses et des pénalités fondées sur la moralité des actions choisies sont ajoutées à la fonction. Par exemple, dans le cas d'un véhicule autonome, De Moura *et al.* [7] utilisent une fonction de récompense fondée sur la proximité des autres usagers de la route, le respect de la loi et la distance au but. En cas de collisions inévitables, la fonction de récompense est modifiée selon des préférences éthiques pour décider de la collision la plus acceptable.

Une autre approche consiste à ajouter une composante éthique au MDP. Par exemple, Svegliato *et al.* [17] définissent des *systèmes autonomes éthiquement conformes* (ECAS pour ethically compliant autonomous systems) qui intègrent un principe moral particulier. Pour cela, une contrainte qui n'est pas directement exprimée dans le MDP est ajoutée au modèle. L'approche est alors fondée sur de l'optimisation sous contrainte, c'est-à-dire la recherche d'une politique qui satisfait cette contrainte. En conséquence, si les contraintes sont incohérentes, une politique éthique peut ne pas exister. Une extension multi-agent de l'ECAS est proposée dans Nashed *et al.* [13]. Ici, chaque agent a un espace d'état particulier et une fonction de valeur

sur celui-ci. Le principe moral est défini pour tenir compte des communautés morales auxquelles l'agent appartient. Dans ces approches, chaque nouveau principe moral nécessite une contrainte spécifique.

Une autre approche, proposée par Rodriguez-Soto *et al.* [16], consiste à utiliser une combinaison de fonctions de récompense. Ici, l'éthique est modélisée par une fonction normative et une fonction évaluative, qui, avec une fonction de récompense classique, guident les agents. La fonction normative pénalise l'agent qui viole une norme (agit d'une manière interdite ou néglige une obligation) tandis que la fonction évaluative le récompense pour des actions autorisées ou obligatoires. La fonction de valeur agrège toutes ces fonctions, et une politique éthique est une politique qui ne viole aucune norme tout en étant louable. Une telle politique peut ne pas toujours exister. Dans ce cas, en raison de l'agrégation, le modèle ne peut pas différencier les politiques en fonction de leurs valeurs. En effet, les politiques ayant de nombreuses normes violées et de nombreuses actions louables peuvent être classées au même rang que celles ayant peu de normes violées et peu d'actions louables.

2.3 Une approche à haut niveau pour l'éthique

Plusieurs travaux en éthique computationnelle [5, 6] distinguent clairement la morale de l'éthique, en se fondant sur la littérature en philosophie. Ici, les options sont évaluées en termes de morale, c'est-à-dire en leur associant le fait de causer du bien ou du mal par rapport à leur conformité aux mœurs, valeurs et usages d'un groupe ou d'une personne [18]. La morale est en effet fondée sur des valeurs morales, qui sont des notions abstraites – positives (ex. courage, sens de la justice) ou négatives (ex. avidité ou cruauté) – caractérisant un large ensemble d'objets : états, actions ou même normes. Prendre une décision qui promeut une valeur morale positive cause du bien, tandis qu'en trahir une correspond généralement à causer du mal. Lorsque deux options sont soutenues par des valeurs morales différentes, chacune apportant un certain regret étant donné que l'exécution des deux est impossible, il s'agit d'un *dilemme moral* [12]. Les sciences sociales ont montré que ces valeurs varient en importance et forment des organisations hiérarchiques appelées systèmes de valeurs [20]. L'éthique est alors à la fois la manière d'éliciter ces valeurs morales (c'est-à-dire de choisir les valeurs qui sont importantes pour le décideur), et de les agréger afin de prendre une décision juste.

Les premières réflexions sur l'éthique distinguent les obligations et les interdictions, des recommandations (positives ou négatives) : ne pas remplir une obligation ou violer une interdiction devrait être sanctionné par une importante pénalité, tandis que faire une action non recommandée ou ne pas faire une action recommandée devrait être sanctionné plus faiblement [2]. D'autres idées – par exemple, la doctrine du double effet [9] – réfutent la dualité entre causer le bien et causer le mal : nous ne pouvons pas facilement justifier de causer le mal en causant le bien car l'un n'est pas la négation de l'autre. Plus encore, la doctrine du faire et du permettre (*Doing and Allowing*) [15] met en évidence la différence entre faire du mal et permettre que du mal se produise. En effet, causer du mal est plus blâmable que de permettre que du mal se produise à l'avenir.

Bien que nous soyons conscients que ces notions de morale et d'éthique peuvent être critiquées ou affinées, nous avons choisi de représenter l'éthique grâce aux notions précédentes. Par conséquent, le modèle que nous devons définir doit : (1) avoir une représentation explicite des valeurs morales qui peuvent être élicitées comme positives ou négatives par le décideur ; (2) considérer explicitement le mal et le bien comme distincts, c'est-à-dire que causer un mal (resp. empêcher un mal) ne signifie pas nécessairement que le bien a été empêché (resp. le bien a été causé) ; (3) faire la distinction entre causer un mal (resp. un bien) et permettre un mal (resp. un bien) ; (4) être capable d'exprimer des obligations, des interdictions et des actions recommandées.

3 Prise de décision éthique

Décider d'un point de vue éthique nécessite un *contexte éthique* qui est propre au décideur. Ce contexte représente les valeurs morales de l'agent qui peuvent être divisées en valeurs positives, négatives et neutres. En effet, la polarité des valeurs peut différer d'un agent à l'autre. Par exemple, l'*obéissance* est une valeur qui peut être positive ou négative selon la position philosophique de l'agent.

Définition 1 *Soit* $\mathcal{V} = \{v_1, \ldots, v_k\}$ *un ensemble de valeurs morales. Un* contexte éthique C *est un tuple* $\{C_1, \ldots, C_k\}$ *où* $C_i \in \{1, -1, 0\}$. *Une évaluation de* 1 *(resp.* -1 *et* 0*) signifie que la valeur est considérée comme positive (resp. négative et neutre) pour l'agent. Soit* $\mathcal{G}_C$ *(resp.* $\mathcal{B}_C$ *et* $\mathcal{N}_C$*) l'ensemble des valeurs positives (resp. négatives et neutres) dans le contexte* C *:* $\mathcal{G}_C = \{v_i \in \mathcal{V} : C_i = 1\}$ *(resp.* $\mathcal{B}_C = \{v_i \in \mathcal{V} : C_i = -1\}$ *et* $\mathcal{N}_C = \{v_i \in \mathcal{V} : C_i = 0\}$*).*

Remarquons que les valeurs ne diffèrent pas en importance entre elles : une valeur positive ne peut pas être considérée comme meilleure qu'une autre (resp. négative, pire). La prise en compte d'une hiérarchie entre les valeurs est laissée à des travaux futurs. À ce stade, une question se pose : comment un agent doit-il prendre une décision en fonction d'un contexte éthique ? D'un point de vue général, une valeur peut être promue ou trahie par une décision. La promotion d'une valeur signifie que le comportement de l'agent est conforme à celle-ci, tandis que la trahison d'une valeur signifie qu'elle est violée par le comportement de l'agent (que la valeur soit positive ou négative). Il semble donc naturel qu'une prise de décision éthique intéressante maximise les valeurs morales positives promues

et minimise les valeurs négatives promues. De plus, elle doit maximiser les valeurs négatives trahies et minimiser les valeurs positives trahies. Comme les valeurs positives ou négatives promues et trahies ne sont pas duales, un tel processus de décision est un processus de décision multi-critère. On souhaite qu'il satisfasse la propriété suivante.

Propriété 1 *Un processus décisionnel éthique maximise la promotion des valeurs positives et la trahison des valeurs négatives, tout en minimisant la trahison des valeurs positives et la promotion des valeurs négatives.*

Cependant, il n'est pas toujours possible de satisfaire la propriété 1. Par conséquent, nous proposons le compromis exprimé par la propriété 2.

Propriété 2 *Un processus décisionnel éthique minimise d'abord le mal en minimisant les valeurs négatives promues et en maximisant celles trahies (le 1^e critère étant plus important que le 2^e), puis maximise le bien en maximisant les valeurs positives promues et en minimisant celles trahies (le 1^e critère étant plus important que le 2^e).*

4 Processus de décision markoviens éthiques

Nous proposons des *processus de décision markoviens éthiques* (E-MDP), qui sont des MDP étendus avec une morale représentée par des valeurs promues ou trahies associées aux transitions, et une éthique, qui dicte comment l'agent optimise sa décision par rapport à la morale.

Définition 2 (MDP éthique) *Un E-MDP est un sextuplet $\langle S, A, \mathcal{T}, C, \mathcal{E}, \mathcal{R}\rangle$ où S, A et $\mathcal{T}$ sont les ensembles classiques d'états, d'actions et la fonction de transition (voir section 2.1), C est un contexte éthique (voir définition 1), $\mathcal{E}$ est une fonction d'évaluation morale (voir définition 3) et $\mathcal{R}$ est une fonction de récompense éthique (voir définition 4).*

Pour exprimer les principes éthiques, nous devons décrire l'alignement du comportement de l'agent avec les valeurs morales. Nous introduisons donc l'évaluation morale des transitions, qui mesure si une transition promeut, trahie ou ne considère pas une valeur donnée du contexte.

Définition 3 (Évaluation morale des transitions) *Chaque transition $(s, a, s') \in S \times A \times S$ où $\mathcal{T}(s, a, s') > 0$, est associée à un tuple $\mathcal{E}(s, a, s') = \langle \epsilon_1, \ldots, \epsilon_k\rangle$ représentant son évaluation morale. Le i-ème élément de $\mathcal{E}(s, a, s')$, noté $\mathcal{E}(s, a, s')_i$, prend une valeur dans $\{1, -1, 0\}$, qui signifie que la valeur morale $v_i \in \mathcal{V}$ est respectivement promue, trahie ou n'est pas considérée.*

4.1 Modélisation de l'éthique

La fonction de récompense des E-MDPs décrit comment les valeurs morales du contexte éthique de l'agent sont alignées avec son comportement. L'agent peut promouvoir ou trahir une valeur, ce qui donne quatre comportements distincts : la promotion d'une valeur positive ou négative – c'est-à-dire *causer du bien ou du mal* – et les comportements trahissant une valeur positive ou négative – c'est-à-dire *empêcher le bien ou le mal*.

Définition 4 (Fonctions de récompense éthique) *La fonction de récompense $\mathcal{R}$ produit un quadruplet où Δ compte le bien causé, ∇ le mal causé, $\overline{\Delta}$ le bien empêché, $\overline{\nabla}$ le mal empêché. Ainsi, $\mathcal{R}(s, a, s') = \langle \Delta, \nabla, \overline{\Delta}, \overline{\nabla}\rangle$ avec $\Delta, \nabla, \overline{\Delta}, \overline{\nabla}$ tels que :*

$$\Delta = \sum_{v_i \in \mathcal{G}_C} x_i \text{ et } \nabla = \sum_{v_i \in \mathcal{B}_C} x_i \text{ où } x_i = \begin{cases} 1 & \text{si } \mathcal{E}(s, a, s')_i = 1, \\ 0 & \text{sinon.} \end{cases}$$

$$\overline{\Delta} = \sum_{v_i \in \mathcal{G}_C} x_i \text{ et } \overline{\nabla} = \sum_{v_i \in \mathcal{B}_C} x_i \text{ où } x_i = \begin{cases} 1 & \text{si } \mathcal{E}(s, a, s')_i = -1, \\ 0 & \text{sinon.} \end{cases}$$

Nous notons $\mathcal{R}_\star(s, a, s')$ avec $\star \in \{\Delta, \nabla, \overline{\Delta}, \overline{\nabla}\}$ la composante $\star$ de la fonction de récompense pour une transition donnée. Suivant la définition 4, une politique d'un E-MDP peut être évaluée selon plusieurs fonctions de valeur éthiques qui donnent la valeur d'une politique dans un état donné selon une composante $\star$ de $\mathcal{R}$.

Définition 5 (Fonction de valeur éthique) *Une fonction de valeur éthique $V^\pi_\star(s)$ pour $\star \in \{\Delta, \nabla, \overline{\Delta}, \overline{\nabla}\}$ est définie comme :*

$$V^\pi_\star(s) = \sum_{a \in A} \pi(a \mid s) \sum_{s' \in S} p(s' \mid s, a)\big(\mathcal{R}_\star(s, a, s') + \big(\gamma V^\pi_\star(s')\big)\big)$$

Étant donné que causer et empêcher le mal et le bien sont distincts et potentiellement conflictuels, la notion d'optimalité devient subjective. De ce fait, nous voulons être en mesure de les exprimer de manière explicite et nous n'agrégeons pas les quatre aspects car nous risquons de perdre des informations nécessaires pour la prise de décision. Au regard des propriétés 1 et 2, une question se pose : « La fin justifie-t-elle les moyens ? En d'autres termes, cherchons-nous à produire le plus grand bien, qui peut résulter d'un mal, ou nous concentrons-nous sur le fait de faire le moins de mal possible, ce qui peut dans certaines circonstances empêcher de faire beaucoup de bien ? » Comme il n'y a pas de hiérarchie entre les valeurs, nous estimons qu'il est plus important pour l'agent d'éviter de causer du mal autant que possible, puis de se focaliser sur le fait de faire autant de bien que possible dans l'espace de décision restant. À cette fin, nous utilisons un ordre lexicographique qui consiste à privilégier les politiques qui causent le moins de mal, puis, à partir de l'ensemble de ces politiques avec un minimum de mal, nous choisissons celles qui font le plus de bien, comme le montrent les définitions suivantes.

Définition 6 *Les politiques optimales π^*_B par rapport aux valeurs négatives sont données par :*

$$\pi^*_B \in \operatorname*{argmin}_\pi V^\pi_\nabla(s) - V^\pi_{\overline{\nabla}}(s) + \epsilon V^\pi_{\overline{\nabla}}(s) \text{ où } \epsilon > 0.$$

Le troisième terme, c'est-à-dire la valeur pondérée du mal empêché, permet de mesurer l'importance de la prévention du mal, qui empêche l'agent d'être absous de causer du mal en le réparant (totalement ou partiellement). Cela incite l'agent à éviter complètement le mal, au lieu de le causer intentionnellement pour le réparer plus tard. En outre, si seule une soustraction entre le mal causé et le mal empêché était prise en compte, les politiques dans lesquelles un mal a été causé et entièrement réparé ne se distingueraient pas de celles dans lesquelles aucun mal n'a été causé. Cela s'applique également aux politiques optimales en ce qui concerne les valeurs positives du contexte.

Définition 7 *Les politiques optimales π^*_G par rapport aux valeurs positives, prises parmi les politiques optimales par rapport aux valeurs négatives, sont données par :*

$$\pi^*_G \in \underset{\pi \in \pi^*_B}{\operatorname{argmax}}\ V^{\pi}_{\Delta}(s) - V^{\pi}_{\overline{\Delta}}(s) - \epsilon' V^{\pi}_{\overline{\Delta}}(s) \text{ où } \epsilon' > 0.$$

Remarquons que la politique optimale par rapport au bien est choisie parmi les politiques qui sont déjà optimales par rapport au mal. Ce choix contraint la notion de politique optimale par rapport au bien, mais permet de nous assurer que l'agent ne considérera pas comme optimal de faire du mal pour obtenir un plus grand bien.

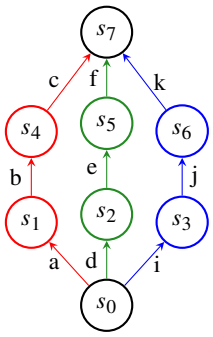
Propriété 3 *La résolution d'un E-MDP en utilisant les définitions 6 et 7 nous permet de calculer une politique qui satisfait la propriété 1 lorsqu'une telle politique existe, ou qui satisfait la propriété 2 sinon.*

Démonstration (Esquisse). Si une politique satisfaisant la propriété 1 existe, alors calculer une politique grâce à la définition 6 puis à la définition 7 retourne cette politique. En effet, si une telle politique existe, notons la π^*, alors elle est telle que pour toute autre politique π, et pour tout $s \in \mathcal{S}, V^{\pi^*}_{\bigstar}(s) \geq V^{\pi}_{\bigstar}(s)$ (resp. $V^{\pi^*}_{\bigstar}(s) \leq V^{\pi}_{\bigstar}(s)$) pour $\bigstar \in \{\Delta, \overline{\nabla}\}$ (resp. $\bigstar \in \{\overline{\Delta}, \nabla\}$). Par conséquent, π^* est optimale pour les critères données par les définitions 6 et 7 et sera donc retournée par tout algorithme optimisant ces derniers. S'il n'existe aucune politique satisfaisant la propriété 1, alors tout algorithme optimisant les critères donnés par les définitions 6 et 7 retournera une politique satisfaisant la propriété 2 par définition. □

Une fois que ces deux critères ont été satisfaits, nous obtenons un ensemble de politiques où l'agent fait le plus de bien possible tout en causant le moins de mal possible. Pour illustrer ces politiques optimales, nous présentons en figure 1 ci-dessous un exemple comparatif de politiques.

4.2 Traitement des tâches amorales

Comme la fonction de récompense éthique ne prend en compte que la moralité, le traitement d'une tâche amorale nécessite un traitement particulier. Une première approche peut consister à considérer que la réalisation de cette tâche promeut une valeur positive spécifique. Ainsi, cette valeur particulière servira d'incitation (dans un sens éthique) pour que l'agent s'efforce également d'atteindre son objectif tout en se comportant de manière éthique. Cependant, comme il n'y a pas de hiérarchie entre les valeurs, si le modèle contient d'autres valeurs morales positives, la tâche amorale pourrait être négligée au profit d'autres valeurs positives. Pour cette raison, nous ajoutons explicitement un autre critère d'optimisation. Nous supposons alors que nous avons une fonction de récompense classique $\mathcal{R}_T$ et, étant donnée une politique optimale par rapport aux valeurs positives, nous nous servons de cette fonction de récompense pour sélectionner la politique la plus optimale au regard de la tâche.



T	$\mathcal{E}(T)_n$	$\mathcal{E}(T)_p$
(s_0, a, s_1)	1	1
(s_1, b, s_4)	1	1
(s_4, c, s_7)	0	1
(s_0, d, s_2)	1	1
(s_2, e, s_5)	1	1
(s_5, f, s_7)	−1	0
(s_0, i, s_3)	1	1
(s_3, j, s_6)	0	1
(s_6, k, s_7)	0	0

FIGURE 1 – La figure présente trois politiques : rouge, verte et bleue. La table indique les évaluations morales positives ($\mathcal{E}(T)_p$) et négatives ($\mathcal{E}(T)_n$) pour chaque transition de ces politiques. Comme nous l'avons mentionné, une politique qui satisfait la propriété 1 n'existe pas toujours. Dans cet exemple, nous pouvons facilement voir qu'aucune politique n'optimise tous les critères en même temps. En effet, la politique rouge maximise le bien causé ($\Delta = 3$ contre $\Delta = 2$ pour les politiques bleue et verte), la politique verte maximise la quantité de mal réparé ($\overline{\nabla} = 1$ contre $\overline{\nabla} = 0$ pour les politiques rouge et bleu), tandis que la politique bleu minimise la quantité de mal causé ($\nabla = 1$ contre $\nabla = 2$ pour les politiques rouge et verte). En utilisant les définitions 6 et 7 nous pouvons déduire l'ordre de préférence suivant : la politique bleue est la meilleure politique (c'est celle qui cause le moins de mal), la politique verte est la deuxième meilleure (elle cause autant de mal que la rouge, mais en répare une partie), et la politique rouge est la moins bonne (même si c'est celle qui fait le plus de bien).

5 Résolution des E-MDP

Pour résoudre les E-MDP, nous adaptons l'algorithme VI pour prendre en compte les spécificités de la fonction de récompense éthique (voir section 4). Notre algorithme fournit une politique non déterministe, c'est-à-dire que plu-

sieurs actions optimales sont associées à chaque état sans choix probabiliste. Remarquons qu'il n'est pas possible de calculer de manière simultanée les Q-valeurs selon chaque critère. En effet, le calcul des Q-valeurs s'appuie sur les actions optimales calculées précédemment. Or, il est possible que des actions non optimales par rapport au mal le soient par rapport au bien. Ainsi, les Q-valeurs propagées d'état en état ne se fonderont pas sur les mêmes actions selon le critère. Pour cette raison, les Q-valeurs et la politique doivent être calculées de manière successive et affinées : une politique par rapport à un critère doit être calculée en ne considérant que les actions optimales pour le critère précédent. Notre approche suit donc les étapes ci-après :

1. Calculer les Q-valeurs en fonction du mal et en extraire une politique optimale, en partant d'une politique avec toutes les actions possibles pour chaque état.

2. Calculer les Q-valeurs en fonction du bien et en extraire une politique optimale, en ne considérant que les actions de la politique optimale par rapport au mal dans chaque état.

3. Si une tâche doit être accomplie (voir section 4.2), calculer les Q-valeurs en fonction de $\mathcal{R}_T$ et en extraire une politique optimale, en ne considérant, dans chaque état, que les actions de la politique optimale par rapport au bien.

Cette procédure assure que la politique optimale finale est celle qui adhère d'abord aux critères éthiques puis, s'il y a une tâche amorale, qu'elle accomplisse la tâche si possible. Pour chaque étape, nous utilisons une variante de l'algorithme VI dont la forme générique est présentée dans l'algorithme 1.

— Le paramètre $\bigstar$ indique le critère pour lequel les politiques sont calculées : $\nabla\overline{\nabla}$ est le critère par rapport au mal et $\Delta\overline{\Delta}$ le critère par rapport au bien.
— Pour le critère $\nabla\overline{\nabla}$, la politique d'entrée est une politique générique qui considère toutes les actions possibles pour chaque état. Pour le critère $\Delta\overline{\Delta}$, la politique d'entrée est la politique optimale par rapport au mal, qui ne considère que les actions optimales pour le mal. Dans le cas où une tâche est présente, la politique d'entrée ne considère que les actions optimales extraites du critère $\Delta\overline{\Delta}$.
— La fonction f calcule les Q-valeurs soit par rapport au mal, soit par rapport au bien, soit classiquement par rapport à $\mathcal{R}_T$. Dans le cas du mal ou du bien, les formules suivantes sont celles présentées dans les définitions 6 et 7. Formellement :

$$f(Q_{\nabla\overline{\nabla}}(s,a)) = \sum_{s'} T(t)\Big[\big[R_{\nabla}(t) - R_{\overline{\nabla}}(t) + \epsilon R_{\overline{\nabla}}(t)\big] + \gamma \min_{a'} Q_{\nabla\overline{\nabla}}(s',a')\Big],$$

$$f(Q_{\Delta\overline{\Delta}}(s,a)) = \sum_{s'} T(t)\Big[\big[R_{\Delta}(t) - R_{\overline{\Delta}}(t) - \epsilon' R_{\overline{\Delta}}(t)\big] + \gamma \max_{a'} Q_{\Delta\overline{\Delta}}(s',a')\Big].$$

où $t = (s,a,s'), \epsilon > 0, \epsilon' > 0$.
— De même, la fonction *Opt* se concentre sur la valeur minimale ou maximale donnée par Q pour une action $a \in A$ en fonction de $\bigstar$. Pour le mal, nous voulons les politiques dans lesquelles le moins de mal a été fait. Pour le bien, nous voulons les politiques dans lesquelles le plus de bien a été fait. Formellement :

$$Opt_{\nabla\overline{\nabla}}(Q_{\nabla\overline{\nabla}},\pi,s) = \big\{a \in \pi(s) : Q_{\nabla\overline{\nabla}}(s,a) = \operatorname*{argmin}_{\pi(s)} Q_{\nabla\overline{\nabla}}(s,\pi(s))\big\},$$

$$Opt_{\Delta\overline{\Delta}}(Q_{\Delta\overline{\Delta}},\pi,s) = \big\{a \in \pi(s) : Q_{\Delta\overline{\Delta}}(s,a) = \operatorname*{argmax}_{\pi(s)} Q_{\Delta\overline{\Delta}}(s,\pi(s))\big\}.$$

Algorithme 1 Itération de la valeur lexicographique

Entrée : la politique π et le critère $\bigstar$
Sortie : la politique optimale $\pi^*_{\bigstar}$

for all $s \in S$ **do**
  for all $a \in \pi(s)$ **do**
    $Q_{\bigstar}(s,a) \leftarrow 0$
  end for
end for
$\Delta \leftarrow 0$
while $\Delta \geq \Delta_o$ **do**
  $\Delta \leftarrow 0$
  for all $s \in S$ **do**
    for all $a \in \pi(s)$ **do**
      $\texttt{temp} \leftarrow Q_{\bigstar}(s,a)$
      $Q_{\bigstar}(s,a) \leftarrow f(Q_{\bigstar}(s,a))$
      $\Delta \leftarrow \max(\Delta, |\texttt{temp} - Q_{\bigstar}(s,a)|)$
    end for
  end for
end while
$\pi^*_{\bigstar}(s) \leftarrow Opt_{\bigstar}(Q^*_{\bigstar},\pi,s) \quad \forall s \in S$
retourner π^*

6 Cadres éthiques dans les E-MDP

Afin de mettre en application les concepts présentés en sections 4 et 5, nous modélisons les cadres éthiques considérés par Svegliato *et al.* [17] : la *théorie du commandement divin* (DCT), les *devoirs prima facie* (PFD) et l'*éthique de*

la vertu (VE). Ces cadres éthiques sont intéressants car ils se focalisent respectivement sur les états, les couples état-action et les transitions. De plus, Svegliato *et al.* ont produit des résultats expérimentaux qui peuvent être comparés à notre modèle. Pour chaque cadre éthique, nous définissons le contexte éthique et les contraintes qu'il implique sur la fonction de l'évaluation morale.

Dans la suite de l'article, nous désignons par T l'ensemble des transitions avec une probabilité non nulle : $T = \{(s, a, s') \in S \times A \times S \mid \mathcal{T}(s, a, s') > 0\}$.

6.1 Théorie du commandement divin

La DCT suppose que certains états sont interdits, et que l'agent cause du mal en y entrant. Nous définissons ces états ci-dessous et les contraintes qu'ils imposent à la fonction de valeur morale.

Définition 8 (États interdits) *Soit $\mathcal{F} = \{\mathcal{F}_1, \ldots, \mathcal{F}_k\}$ un ensemble d'ensembles d'états interdits $\mathcal{F}_i \subseteq S$. Chaque $\mathcal{F}_i$ est associé à une valeur négative $v_i \in \mathcal{B}_C$ du contexte.*

Le fait de disposer de plusieurs sous-ensembles d'états interdits nous permet de les associer à des valeurs morales différentes. Par exemple, un état pourrait être interdit pour éviter de mettre en danger des personnes, tandis qu'une autre pourrait être interdit pour protéger la vie privée des personnes. Comme il s'agit de valeurs différentes, chaque état fera partie d'un sous-ensemble différent d'états interdits.

Transiter dans un état appartenant à un sous-ensemble interdit promeut la valeur négative associée, car l'agent cause du mal en faisant cela.

Définition 9 (Entrée dans un état interdit) *Les transitions qui se terminent dans un état interdit promeuvent la valeur associée.*

$$\forall (s, a, s') \in T, \forall \mathcal{F}_i \in \mathcal{F} \text{ t.q. } s' \in \mathcal{F}_i : \mathcal{E}(s, a, s')_i = 1$$

Lorsque l'agent se trouve dans un état interdit, il doit en sortir le plus rapidement possible. Pour inciter l'agent à quitter ces états, transiter d'un état interdit vers un état qui ne l'est pas permet d'éviter le mal.

Définition 10 (Sortie d'un état interdit) *Les transitions qui commencent dans un état interdit et se terminent dans un état qui n'est pas interdit trahissent la valeur associée.*

$$\forall (s, a, s') \in T, \forall \mathcal{F}_i \in \mathcal{F} \text{ t.q. } (s \in \mathcal{F}_i \wedge s' \notin \mathcal{F}_i) : \mathcal{E}(s, a, s')_i = -1$$

Les transitions dans lesquelles ni l'état de départ ni l'état résultant ne sont interdits ont une évaluation morale qui ne considère aucune valeur.

Définition 11 (Autres transitions)

$$\forall (s, a, s') \in T, \forall \mathcal{F}_i \in \mathcal{F} \text{ t.q. } (s \notin \mathcal{F}_i \wedge s' \notin \mathcal{F}_i) : \mathcal{E}(s, a, s')_i = 0$$

La figure 2 illustre les définitions précédentes.

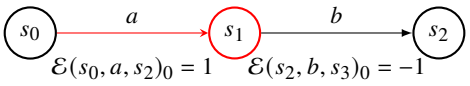


FIGURE 2 – Un exemple de cadre DCT. Soit le contexte éthique $C = \langle -1 \rangle$ et les états interdits $\mathcal{F}_0 = \{s_1\}$. L'état s_1 est interdit et la transition qui y entre promeut le mal. La transition qui en sort et qui arrive en s_2 (qui n'est pas interdit) trahit la valeur négative, et empêche ainsi le mal.

6.2 Devoirs *Prima Facie*

Les PFD supposent l'existence de devoirs fondamentaux que l'agent doit accomplir. Nous définissons ci-dessous ces devoirs et les contraintes qu'ils imposent à la fonction d'évaluation morale.

Définition 12 (Devoir) *Soit $\mathcal{D} = \{\delta_1, \ldots, \delta_k\}$ un ensemble de devoirs. Un devoir δ_i est un ensemble de couples état-action : $\delta_i = \{(s_0, a_0), \ldots, (s_n, a_n)\}$. Chaque devoir est associé à une valeur négative v_i dans C.*

L'agent *accomplit le devoir* en choisissant l'action du devoir dans l'état associé, et *néglige le devoir* si une action différente est choisie.

Accomplir ou négliger un devoir peut être interprété de deux manières : **positivement** où l'agent cause du bien chaque fois qu'un devoir est accompli, et **négativement** où l'agent cause du mal chaque fois qu'un devoir est négligé. Dans ce cas précis, les notions de *causer le bien* et *causer le mal* deviennent duales, et nous pouvons donc soit nous concentrer sur les devoirs accomplis (sur la maximisation du bien), soit nous concentrer sur les devoirs négligés (sur la minimisation du mal). Nous choisissons ici de nous concentrer sur l'interprétation négative. Comme les devoirs sont considérés comme fondamentaux, c'est-à-dire que l'agent est tenu de les accomplir, le fait de revenir à un état où un devoir a été précédemment négligé et de l'accomplir ne sera pas considéré comme empêchant un mal. Par conséquent, seul un nouveau mal peut être évité et, en raison de la nature des devoirs, le mal causé ne peut pas du tout être empêché dans ce cadre.

Définition 13 (Négliger un devoir) *Négliger un devoir δ_i consiste à passer par des transitions (s, a, s') où il existe $(s, a') \in \delta_i$ et où $(s, a) \notin \delta_i$. Dans ce cas, (s, a, s') promeut*

la valeur négative v_i.

$$\forall (s,a,s') \in T \text{ t.q. } \exists\, \delta_i \in \mathcal{D}, (s,a') \in \delta_i \ \wedge\ (s,a) \notin \delta_i : \\ \mathcal{E}(s,a,s')_i = 1.$$

Les autres transitions ne prennent en considération aucune valeur dans leur évaluation.

Définition 14 (Autres transitions) *Toutes les transitions (s,a,s') où ni l'état de départ, ni l'action ne font partie d'un devoir δ_i ont une évaluation morale non applicable pour la valeur négative associée $v_i \in \mathcal{B}_C$:*

$$\forall (s,a,s') \in T, \forall\, \delta_i \in \mathcal{D} \text{ t.q. } (s,a) \notin \delta_i : \mathcal{E}(s,a,s')_i = 0$$

La figure 3 illustre les définitions précédentes.

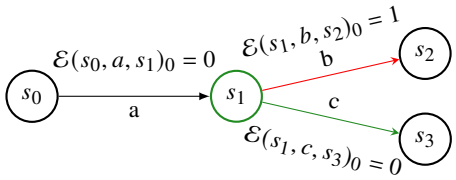


FIGURE 3 – Un exemple de cadre PFD. Soit un contexte $C = \langle -1 \rangle$ et le devoir $\delta = (s_1, c)$. Choisir l'action c dans s_1 accomplit le devoir et ne cause pas de mal. Le choix de b promeut la valeur négative et cause donc un mal.

6.3 L'éthique de la vertu

Le cadre VE suppose qu'il existe un exemplaire moral qui se comporte comme nous voulons que notre agent se comporte. L'agent doit alors se conformer à ce comportement, représenté par des *trajectoires morales* (MT).

Définition 15 (Trajectoire morale) *Une trajectoire morale d'un exemplaire moral est un ensemble $MT = \{(s_0,a_0,s'_0),\ldots,(s_n,a_n,s'_n)\}$ de transitions. Une trajectoire morale est associée à deux valeurs morales v_n et v_p telles que $C_n = -1$ et $C_p = 1$: v_n est une valeur négative et v_p une valeur positive.*

Comme les résultats des actions ne sont pas déterministes, il peut y avoir des sorties de trajectoire non intentionnelles : l'agent essaie de suivre la trajectoire mais n'y parvient pas en raison d'effets stochastiques. Dans les autres cadres éthiques, ceci n'a pas d'importance car ils reposent sur des interdictions, i.e. ne pas atteindre un état ou ne pas négliger un devoir. Dans l'éthique de la vertu, au contraire, ces sorties accidentelles ont leur importance car ce n'est pas tant le résultat de l'action qui compte que l'intention qui était exprimée derrière. Deux valeurs sont alors nécessaires pour prendre cela en compte, l'une pour récompenser le suivi intentionnel ou pénaliser les sorties non intentionnelles, et l'autre pour pénaliser les sorties intentionnelles. Dans la suite, nous distiguons les *trajectoires moralement bonnes* (MGT) et les *trajectoires moralement mauvaises* (MBT), respectivement les trajectoires morales que nous voulons que l'agent suive et les trajectoires que nous voulons que l'agent évite. Les contraintes sur la fonction d'évaluation morale dépendent alors du type de MT.

6.3.1 Trajectoires moralement bonnes

Suivre une trajectoire définie par un exemplaire moral est considéré comme moralement bon. Ainsi, les transitions qui font partie d'une MGT promeuvent la valeur positive associée et causent du bien.

Définition 16 (Maintenir une MGT) *Les transitions qui appartiennent à une MGT promeuvent la valeur positive associée.*

$$\forall (s,a,s') \in MGT : \mathcal{E}(s,a,s')_p = 1.$$

Lorsque l'agent choisit une action qui suit une MGT mais que l'état successeur n'est pas celui attendu, il fait une sortie accidentelle de MGT. Comme l'agent n'avait pas l'intention de quitter la trajectoire, il ne fait qu'empêcher le bien, sans pour autant causer du mal.

Définition 17 (Sortie accidentelle d'une MGT) *Les transitions qui ont le même état de départ et la même action qu'une transition dans une MGT, mais dont l'état successeur est différent, trahissent sa valeur positive.*

$$\forall (s,a,s') \in T \text{ t.q. } \exists (s_g,a_g,s'_g) \in MGT \\ \text{où } s = s_g \wedge a = a_g \wedge s' \neq s'_g \ \wedge\ (s,a,s') \notin MGT : \\ \mathcal{E}(s,a,s')_p = -1.$$

Tandis que dans une MGT, si l'agent choisit une action qui n'en fait pas partie tout en ayant la possibilité de choisir une action correcte, il choisit de sortir intentionnellement de la trajectoire. Par conséquent, non seulement il empêche le bien, mais il cause aussi du mal.

Définition 18 (Sortie intentionnelle d'une MGT) *Les transitions ayant un état de départ qui est un état de départ dans un MGT, mais dont l'action est différente de celle du MGT, promeuvent sa valeur négative et trahissent sa valeur positive.*

$$\forall (s,a,s') \in T \text{ t.q. } \exists (s_g,a_g,s'_g) \in MGT \\ \text{où} \quad s = s_g \wedge a \neq a_g : \\ \mathcal{E}(s,a,s')_n = 1 \quad \text{et} \quad \mathcal{E}(s,a,s')_p = -1.$$

Pour toutes les autres transitions, l'évaluation morale ne considère aucune valeur positive ou négative.

Définition 19 (Autres transitions) *Les transitions qui ne satisfont pas les définitions 16, 17, 18 sont évaluées comme suit :*

$$\mathcal{E}(s,a,s')_{n,p} = 0.$$

La figure 4 illustre les définitions précédentes.

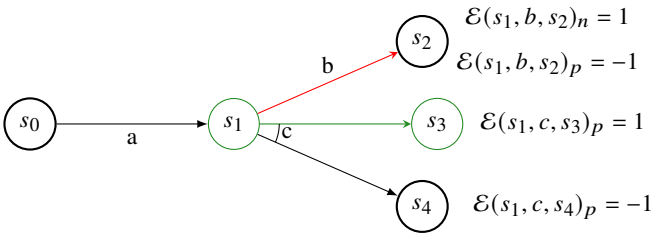


FIGURE 4 – Un exemple de MGT pour un cadre VE. Soit le contexte $C = \langle -1, 1\rangle$ et une MGT contenant (s_1, c, s_3). Si l'agent effectue l'action c, il peut se retrouver en s_3 et cause le bien, ou en s_4 et empêche le bien. Si l'agent choisit b, il sort intentionnellement de la trajectoire, causant du mal.

6.3.2 Trajectoires moralement mauvaises

Un agent se trouvant dans un état qui fait partie d'une MBT ne doit pas y rester. Par conséquent, les transitions qui partent de cet état et qui y retournent causent du mal, même si elle n'appartiennent pas à la MBT.

Définition 20 (Rester dans une MBT) *Les transitions, dont l'état de départ et l'état sucesseur sont les mêmes et qui sont aussi un état de départ d'une transition dans une MBT, promeuvent la valeur négative.*

$$\forall (s,a,s') \in T \text{ t.q. } \exists (s_b,a_b,s'_b) \in MBT$$
$$où\ s = s_b \wedge s = s' \ : \mathcal{E}(s,a,s')_n = 1.$$

Bien entendu, le fait de suivre une MBT est également considéré comme causant du mal.

Définition 21 (Maintenir une MBT) *Les transitions qui appartiennent à une MBT promeuvent sa valeur négative.*

$$\forall (s,a,s') \in MBT : \mathcal{E}(s,a,s')_n = 1.$$

À cause de la stochasticité, les sorties accidentelles des MBT sont également possibles. Cela signifie que l'agent a choisi de rester dans une MBT mais a échoué. Une telle transition cause certes du bien car l'agent est sorti de la MBT mais cause également du mal car l'intention était d'y rester.

Définition 22 (Sortie accidentelle d'une MBT) *Les transitions ayant le même état de départ et action qu'une transition dans une MBT, mais dont l'état successeur est différent, promeuvent à la fois sa valeur négative et positive.*

$$\forall (s,a,s') \in T \text{ t.q. } \exists (s_b,a_b,s'_b) \in MBT$$
$$où \quad s = s_b \wedge a = a_b \wedge s' \neq s'_b :$$
$$\mathcal{E}(s,a,s')_n = 1 \quad et \quad \mathcal{E}(s,a,s')_p = 1.$$

Dans un MBT, si l'agent choisit d'en sortir alors qu'il a la possibilité de le suivre, il empêche le mal et cause le bien.

Définition 23 (Sortie intentionnelle d'une MBT) *Les transitions ayant un état de départ qui est un état de départ d'une transition d'une MBT, mais dont l'action ne fait pas partie de celle-ci, trahissent la valeur négative et promeuvent la valeur positive.*

$$\forall (s,a,s') \in T \text{ t.q. } \exists (s_b,a_b,s'_b) \in MBT$$
$$où \quad s = s_b \wedge a \neq a_b \wedge \exists a' = a_b :$$
$$\mathcal{E}(s,a,s')_n = -1 \quad et \quad \mathcal{E}(s,a,s')_p = 1.$$

Pour toutes les autres transitions, l'évaluation morale ne considère aucune valeur positive ou négative.

Définition 24 (Autres transitions) *Les transitions qui ne satisfont pas les définitions 20, 21, 22, 23 sont évaluées comme suit :*

$$\mathcal{E}(s,a,s')_{n,p} = 0.$$

La figure 5 illustre les définitions précédentes.

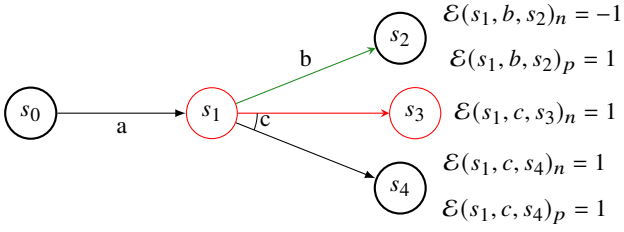


FIGURE 5 – Un exemple de MBT pour un cadre VE. Soit le contexte $C = \langle -1, 1\rangle$ et le MBT contenant (s_1, c, s_3). Si l'agent effectue l'action c, il peut se retrouver en s_3 et ne cause que du mal, ou en s_4 et cause à la fois du mal et du bien, car l'agent a choisi de suivre la mauvaise trajectoire, mais a eu une sortie accidentelle. Si l'agent choisit b, il choisit activement de sortir de la trajectoire, causant ainsi du bien et empêchant le mal.

7 Expériences

Pour évaluer notre modèle, nous avons implémenté le scénario de véhicule autonome proposé par Svegliato *et*

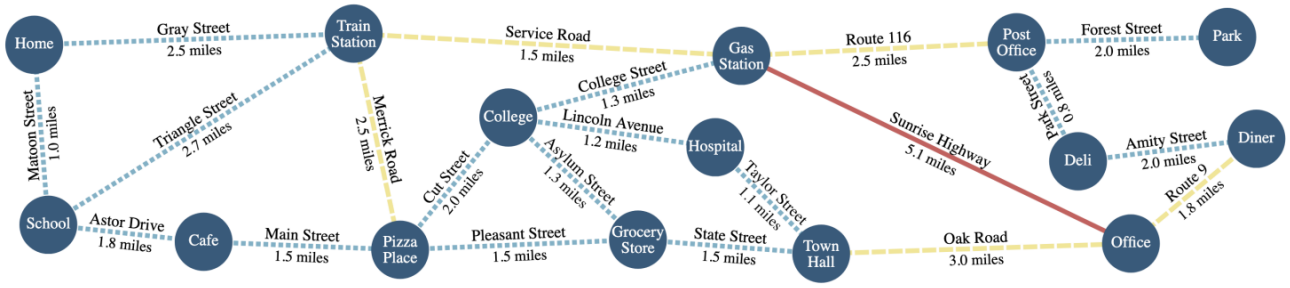


FIGURE 6 – Graphe des états (lieux et routes) utilisés dans nos expériences, tiré de Svegliato *et al.* [17].

al. [17]. Dans celui-ci, un véhicule doit naviguer dans une zone urbaine où différents lieux sont représentés par des états reliés entre eux par d'autres états représentant des routes. La figure 6 représente ces états et la manière dont ils sont reliés entre eux. Les états de route sont caractérisés par des informations sur leur type (encodant les limitations légales de vitesse), la circulation piétonne présente (élevée ou faible) et la vitesse à laquelle le véhicule roule (élevée, normale ou faible). Lorsque l'agent se trouve dans un état de lieu, il peut choisir sur quelle route il va tourner et s'il se trouve sur une route, il peut choisir la vitesse à laquelle il conduira jusqu'au prochain lieu.

Les expériences consistent en trois tâches de navigation entre un état initial et un état but donnés tout en respectant des contraintes éthiques qui peuvent être combinées. Les différentes contraintes éthiques pour chaque cadre sont données ci-après. (**DCT - H**) Tous les états où l'agent roule à vitesse élevée sont interdits. (**DCT - I**) Les états où l'agent roule à vitesse normale ou élevée alors qu'il y a une circulation piétonne élevée sont interdits. (**PFD -** δ_1) Dans tous les états où la circulation piétonne est faible, l'agent a le devoir de conduire à une vitesse normale ou élevée. (**PFD -** δ_2) Dans tous les états où la circulation piétonne est élevée, l'agent a l'obligation de conduire à vitesse faible. (**VE - C**) L'agent ne considère qu'une MGT composées des transitions où la conduite est à vitesse normale lorsqu'il y a une circulation piétonne faible et celles où la conduite est à faible vitesse lorsque la circulation piétonne est élevée. (**VE - P**) L'agent ne considère qu'une MBT composée de toutes les transitions qui atteignent un états de type « Autoroute » ou « École ».

Le table 1 indique le *prix de la moralité*, qui est le pourcentage de perte de valeur entre les politiques éthiques et les politiques optimales en absence de tout cadre éthique. Naturellement, le prix de la moralité est nul lorsqu'il n'y a pas de cadre éthique. De plus, il est important de noter que les valeurs prises par le prix de la moralité n'ont de sens que relativement au modèle de l'application, càd la manière dont la récompense amorale est construite. Par exemple, DCT-I et PFD-δ_1 ont un prix de la moralité nul car ces contraintes éthiques n'interdisent pas les trajectoires optimale amorale. Remarquons que combiner des contraintes,

Éthique	Contraintes	Tâche 1	Tâche 2	Tâche 3
Aucune	-	0%	0%	0%
DCT	H	25.82%	26.25%	34.22%
	I	0%	0%	0%
	H ∪ I	36.12%	36.98%	43.59%
PFD	δ_1	0%	0%	0%
	δ_2	15.47%	15.98%	22.46%
	$\delta_1 \cup \delta_2$	15.47%	15.98%	22.46%
VE	C	36.12%	36.98%	43.59%
	P	11.54%	52.22%	0%
	C ∪ P	64.38%	127.73%	50.28%

TABLE 1 – Prix de la moralité dans l'état de départ.

même certaines ayant individuellement un prix de la moralité nul, peut entraîner un prix de la moralité plus élevé. En effet, une contrainte individuelle peut impacter la politique morale optimale qui aurait été calculée indépendamment avec l'autre contrainte. Remarquons également que les tâches 1, 2 et 3 sont de difficulté croissante pour satisfaire tout à la fois le but amoral et les contraintes éthiques. Il est important de noter que le but des expériences n'est pas nécessairement d'obtenir des résultats quantifiés en termes de performance plus ou moins importante, mais d'illustrer le fait que notre modèle capture les différentes contraintes éthiques considérées par Svegliato *et al* [17]. L'objectif est donc de montrer que le prix de la moralité évolue de manière similaire pour les mêmes tâches et contraintes, mettant ainsi en lumière que le modèle est générique et peut représenter différents principes éthiques.

Nous voyons alors que de manière cohérente le prix de la moralité augmente puisque le modèle privilégie l'éthique sur le but amoral. Lorsque nous comparons ces résultats à ceux de Svegliato *et al.* [17] nous retrouvons les mêmes tendances dans la dégradation du prix de la moralité, à savoir une dégradation avec la difficulté de la tâche et avec un cumul des contraintes. La différence entre notre approche et celle de Svegliato *et al.* est que ces derniers font de l'optimisation sous contraintes, c'est-à-dire qu'ils calculent la politique optimale amorale et la dégrade jusqu'à satisfaire les contraintes éthiques, tandis que nous calculons des

politiques qui adhèrent d'abord aux cadres éthiques, puis sont optimisées sur la base des critères restants (comme les tâches amorales). Une autre différence est que notre approche est générique au sens où l'éthique est définie dans le modèle – dans la récompense elle-même et la fonction de valeur à partir des concepts de valeurs – et non pas comme contrainte exogène qui est une fonction spécifiquement définie avec des éléments qui lui sont propres.

8 Conclusion

Le modèle E-MDP intègre explicitement l'éthique dans les MDP comme des valeurs morales positives et négatives, promues ou trahies. Il utilise une fonction de récompense sous forme de quadruplet pour optimiser d'abord le fait de causer le moins de mal possible, puis de causer le plus de bien possible. Nous avons illustré ce modèle sur trois cadres éthiques, DCT, PFD et VE, qui peuvent être combinés. Enfin, nous avons comparé nos résultats avec ceux de Svegliato *et al.* [17], exhibant des similarités qui nous amènent à penser que notre approche est une manière générique adéquate d'intégrer l'éthique dans les MDP.

Comme perspectives sur la représentation de l'éthique, ce modèle doit être étendu pour des valeurs hiérarchiques mais aussi pour traiter des agents multiples. Intuitivement, chaque agent devrait avoir son propre contexte éthique, et il serait intéressant que des valeurs morales représentent la façon dont les agents prennent en compte les autres. Par exemple, une valeur égalitaire pourrait être promue lorsqu'une décision favorise ou trahit le même nombre de valeurs pour tous les agents.

Comme perspectives algorithmiques, nous pouvons envisager une extension de notre modèle aux observations partielles (PO-MDP). Il serait également intéressant d'étudier l'utilisation d'autres algorithmes de calcul de politique pour trouver les politiques optimales plus efficacement, tel que l'emploi de techniques de recherche heuristique.

Références

[1] Abel, David, James MacGlashan et Michael L. Littman: *Reinforcement Learning as a Framework for Ethical Decision Making*. Dans *AAAI Workshop on AI, Ethics, and Society*, pages 54–61, 2016.

[2] Averroes: *The Decisive Treatise*, 1178.

[3] Bellman, Richard: *A Markovian Decision Process*. Indiana Univ. Math. J., 6 :679–684, 1957.

[4] Bench-Capon, Trevor: *Persuasion in Practical Argumentation Using Value-Based Argumentation Frameworks*. J. Log. Comput., 13(3) :429–448, 2003.

[5] Berreby, Fiona, Gauvain Bourgne et Jean Gabriel Ganascia: *A Declarative Modular Framework for Representing and Applying Ethical Principles*. Dans *16e AAMAS*, page 96–104, 2017.

[6] Cointe, Nicolas, Grégory Bonnet et Olivier Boissier: *Ethical Judgment of Agents' Behaviors in Multi-Agent Systems*. Dans *15e AAMAS*, pages 1106–1114, 2016.

[7] De Moura, Nelson, Raja Chatila, Katherine Evans, Stéphane Chauvier et Ebru Dogan: *Ethical Decision Making for Autonomous Vehicles*. Dans *IEEE Intelligent Vehicles Symposium*, pages 2006–2013, 2020.

[8] Fard, Mahdi Milani et Joelle Pineau: *Non-Deterministic Policies in Markovian Decision Processes*. Journal of Artificial Intelligence Research, 40 :1–24, 2011.

[9] Foot, Philippa: *The Problem of Abortion and the Doctrine of the Double Effect*. Oxf. Rev., 5 :5–15, 1967.

[10] Howard, Ronald A.: *Dynamic Programming and Markov Processes*. MIT Press, Cambridge, MA, 1960.

[11] Lorini, Emiliano: *On the Logical Foundations of Moral Agency*. Dans *11e DEON*, tome 7393 de *LNCS*, pages 108–122. Springer-Verlag, 2012.

[12] McConnell, Terrance: *Moral Dilemmas*. Dans Zalta, Edward N. (rédacteur) : *The Stanford Encyclopedia of Philosophy*. 2018.

[13] Nashed, Samer, Justin Svegliato et Shlomo Zilberstein: *Ethically Compliant Planning within Moral Communities*. Dans *4e AIES*, pages 188–198, 2021.

[14] Puterman, M.: *Markov Decision Processes : Discrete Stochastic Dynamic Programming*. Wiley & Sons, 1994.

[15] Rickless, Samuel C.: *The Doctrine of Doing and Allowing*. Philos. Rev., 106(4) :555–575, 1997.

[16] Rodriguez-Soto, Manel, Maite Lopez-Sanchez et Juan A. Rodriguez-Aguilar: *A Structural Solution to Sequential Moral Dilemmas*. Dans *19e AAMAS*, pages 1152–1160, 2020.

[17] Svegliato, Justin, Samer Nashed et Shlomo Zilberstein: *Ethically Compliant Sequential Decision Making*. Dans *35e AAAI*, pages 11657–11665, 2021.

[18] Timmons, Mark: *Moral Theory : an Introduction*. Rowman & Littlefield Publishers, 2012.

[19] Watkins, Christopher: *Learning From Delayed Rewards*. Thèse de doctorat, Cambridge Univ., 1989.

[20] Wiener, Yoash: *Forms of Value Systems : A Focus on Organisational Effectiveness and Cultural Change and Maintenance*. Acad. Manage. Rev., 13(4) :534–545, 1988.

[21] Wu, Yueh Hua et Shou De Lin: *A Low-Cost Ethics Shaping Approach for Designing Reinforcement Learning Agents*. Dans *32e AAAI*, pages 1687–1694, 2018.