Comment rendre des comportements plus prédictibles

Salomé Lepers Vincent Thomas Olivier Buffet

Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
prénom.nom@loria.fr

Résumé

Dans cet article, nous nous intéressons à des problèmes de prédictibilité, c'est à dire pour lesquels un agent doit choisir sa stratégie dans le but d'optimiser les prédictions que pourrait faire un observateur extérieur. Nous abordons ces problèmes en tenant compte des incertitudes sur la dynamique de l'environnement et sur la politique de l'agent observé. Dans ce but, nous faisons l'hypothèse que l'observateur 1. cherche à prédire l'action ou l'état future de l'agent à chaque pas de temps, et 2. suppose que l'agent agit selon une politique stochastique calculée à partir d'un problème sous-jacent connu, et nous nous appuyons sur le cadre des processus de décision markoviens conscients d'un observateur (OAMDP). Nous considérons différents critères de performance candidats pour la prédictibilité à travers des fonctions de récompense construit sur la croyance de l'observateur concernant la politique de l'agent ; montrons que ces OAMDP prédictibles induits peuvent être représentés par des MDP orientés but ou actualisés ; et analysons les propriétés des fonctions de récompense proposées à la fois théoriquement et empiriquement sur deux types de mondes grilles.

Abstract

In this paper, we are interested in predictability problems, wherein an agent must choose its strategy in order to optimize the predictions that an external observer could make. We address these problems while taking into account uncertainties on the environment dynamics and on the observed agent's policy. To that end, we assume that the observer 1. seeks to predict the agent's future action or state at each time step, and 2. models the agent using a stochastic policy computed from a known underlying problem, and we leverage on the framework of observer-aware Markov decision processes (OAMDPs). We consider several candidate predictability performance criteria through reward functions built on the observer's belief about the agent policy; show that these induced *predictable* OAMDPs can be represented by goal-oriented or discounted MDPs; and analyze the properties of the proposed reward functions both theoretically and empirically on two types of grid-world problems.

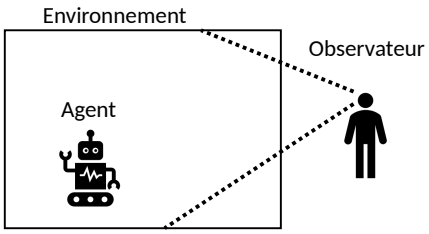


FIGURE 1 – Agent dans son environnement et un observateur passif

1 Introduction

Dans des situations de collaboration homme-robot, certaines propriétés du comportement du robot peuvent être appréciées de l'humain, voire permettre une meilleure collaboration. Divers travaux récents ont porté sur l'obtention automatique de comportements dotés de telles propriétés, en particulier dans le cas où l'humain ne fait qu'observer l'agent dans son environnement, et où l'agent, conscient de cet observateur, cherche à adopter un comportement qui permette de contrôler au mieux les informations acquises par l'humain (cf. figure 1).

CHAKRABORTI, KULKARNI, SREEDHARAN et al. [1] proposent une taxonomie des différents concepts rencontrés dans ces travaux, certains cherchant 1. à transmettre de l'information, tels que la *lisibilité* (lorsque l'agent essaye de communiquer son but à travers ses choix d'actions), l'*explicabilité* (un comportement explicable est conforme aux attentes de l'observateur), et la *predictabilité* (un comportement est prédictible si il est facile de deviner la fin d'une trajectoire en cours), ou 2. d'autres à cacher de l'information, par exemple l'*obscurcissement*, quand le comportement vise à cacher la tâche réelle de l'agent. Ils formalisent aussi ces différents problèmes de manière unifiée sous l'hypothèse que les transitions sont déterministes, raisonnant donc principalement sur des plans (une séquence d'actions

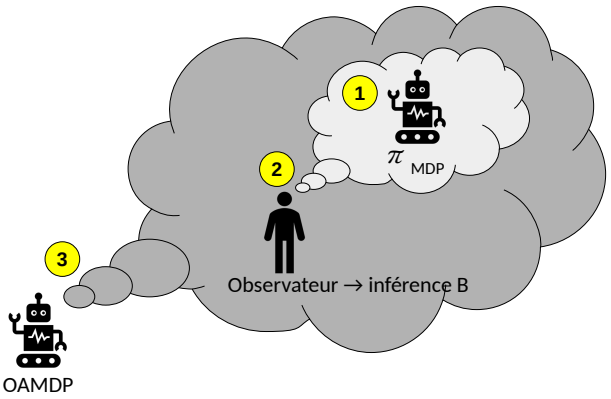


FIGURE 2 – Un agent OAMDP (3) fait l'hypothèse que l'observateur s'attend (2) à ce que l'agent se comporte de manière à accomplir une certaine tâche (1).

induisant une unique séquence d'états). Dans leur approche, le robot modélise l'humain comme ayant un certain modèle du système robot+environnement (y compris de la ou les tâches possibles du robot), et pouvant ainsi anticiper les comportements possibles du robot. Chacune de ces propriétés peut être intéressante dans certaines situations et transmet différentes informations à l'observateur. CHAKRABORTI, KULKARNI, SREEDHARAN et al. [1] expliquent qu'un plan explicable peut être imprévisible, notamment dans le cas où il existe plusieurs plans explicables. FISAC, LIU, HAMRICK et al. [2] suggèrent que, lorsqu'un agent agit de façon lisible, il est possible d'inférer son but mais pas forcément la façon dont il va atteindre ce but (auquel cas il aurait un comportement prédictible).

MIURA et ZILBERSTEIN [3], pour leur part, proposent un formalisme générique analogue (voir figure 2), mais sous l'hypothèse de transitions stochastiques, d'où le nom de *processus de décision markovien conscient d'un observateur* (OAMDP pour *observer-aware Markov decision process*). Entre autres choses, ils travaillent aussi sur l'explicabilité, la lisibilité et la prédictibilité. Comme MIURA et ZILBERSTEIN l'exposent, DRAGAN, LEE et SRINIVASA [4] ont proposé, sous hypothèse de transitions déterministes, de modéliser la prédictibilité d'une trajectoire comme proportionnelle à sa valeur, ce qui peut se traduire dans le cadre OAMDP par la maximisation de la récompense sous-jacente. FISAC, LIU, HAMRICK et al. [2], pour leur part, ont proposé de modéliser des agents t-prédictibles comme maximisant $P(a_{t+1}, \ldots, a_T | a_1, \ldots, a_t)$, ce qui peut être adapté dans le cas stochastique, mais avec un fort coût computationnel [3]. L'objectif de cet article est donc de proposer une nouvelle façon de modéliser la prédictibilité raisonnant non pas sur des séquences d'actions complètes, comme peuvent y inciter les travaux dans des cadres déterministes, mais sur les choix d'actions dans chaque état rencontré. Cela implique de raisonner sur des types dynamiques, ce qui requiert d'introduire une variante du formalisme OAMDP. En outre, nous ne considérons pas que des problèmes avec facteur d'actualisation, mais aussi des problèmes de type "chemin stochastique le plus court" (orientés vers des buts), étendant ainsi le cadre OAMDP.

La section 2 introduit des pré-requis sur le processus de décision markoviens (MDP) et les MDP conscients d'un observateur. Notre approche général et quelques fonctions de récompense candidates l'implémentant sont décrites en sec. 3. Des expérimentations sont décrites en sec. 4 pour illustrer et analyser davantage les comportements obtenus dans différentes configurations avant de conclure en sec. 5.

2 Pré-requis

Nous présentons d'abord brièvement les processus de décision markoviens avant de passer au cadre des MDP conscients d'un observateur.

2.1 Processus de décision markoviens

Un *processus de décision markovien* (MDP) est un 6-uplet $\langle \mathcal{S}, \mathcal{A}, T, R, \gamma, \mathcal{S}_f \rangle$ où :

- $\mathcal{S}$ est l'ensemble des états ;
- $\mathcal{A}$ est l'ensemble des actions ;
- $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0; 1]$, la fonction de transition, donne la probabilité $T(s, a, s')$ d'aller dans un état s' depuis un état s en exécutant l'action a ;
- $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, la fonction de récompense, donne la récompense reçue $R(s, a, s')$ lors d'une transition (s, a, s') ;
- $\gamma \in [0, 1]$ est le facteur d'actualisation ; et
- $\mathcal{S}_f \subset \mathcal{S}$ est l'ensemble des états terminaux : pour tout $s, a \in \mathcal{S} \times \mathcal{A}$, $T(s, a, s) = 1$ et $R(s, a, s) = 0$.

Une politique $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ détermine un comportement en associant à chaque état une action à effectuer. Elle peut éventuellement être stochastique, $\pi(a|s)$ étant alors la probabilité d'effectuer a dans l'état s. Considérant un *MDP actualisé*, c'est-à-dire tel que $\gamma < 1$, la valeur d'une politique π en un état s est l'espérance de la somme des récompenses actualisées sur un horizon infini :

$$V^\pi(s) \stackrel{\text{def}}{=} \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r(S_t, A_t) | S_0 = s \right].$$

Il existe toujours au moins une politique π^*, dite optimale, telle que, pour tout s, $V^{\pi*}(s) = \max_\pi V^\pi(s)$. L'algorithme d'*itération sur la valeur* (VI) calcule cette fonction de valeur optimale, notée V^*, en itérant le calcul suivant jusqu'à atteindre une précision suffisante (où k désigne l'itération courante) :

$$V_{k+1}(s) \leftarrow \max_a \sum_{s'} T(s, a, s') \cdot (R(s, a, s') + \gamma V_k(s')) .$$

On peut alors dériver une politique déterministe optimale en agissant de "manière gourmande" dans tout état s avec :

$$\pi^*(s) \leftarrow \arg\max_a \sum_{s'} T(s,a,s') \cdot (R(s,a,s') + \gamma V^*(s')) .$$

On interrompt les calculs quand le *résidu de Bellman* est inférieur à un seuil fonction de l'erreur ϵ souhaitée et de γ :

$$\underbrace{\max_s |V_{k+1}(s) - V_k(s)|}_{\text{résidu de Bellman}} \leq \frac{1-\gamma}{\gamma}\epsilon.$$

Les propriétés ci-dessus restent valident avec $\gamma = 1$ si

1. $\mathcal{S}_f$ non vide ; et
2. R est telle qu'il existe des politiques atteignant $\mathcal{S}_f$ avec probabilité 1 depuis tout état s, et que la valeur des autres politiques diverge vers $-\infty$ dans les états depuis lesquels on ne peut pas être sûr de pouvoir atteindre un état terminal.

On parle alors de problème de type *chemin stochastique le plus court* (SSP). On a un SSP en particulier, si, pour tout $(s,a,s') \in (\mathcal{S} \setminus \mathcal{S}_f) \times \mathcal{A} \times \mathcal{S}$, $r(s,a,s') < 0$, c'est-à-dire si on cherche à atteindre un état terminal à "moindre coût" (en moyenne).

Note : On peut transformer tout MDP actualisé en un SSP dans lequel, à chaque instant, on a une probabilité $1-\gamma$ de transiter vers un état terminal. Le cas SSP est donc plus général.

2.2 MDP conscient d'un observateur

Un *MDP conscient d'un observateur* (OA-MDP pour *observer-aware MDP*) [3] décrit une situation dans laquelle un agent interagit avec son environnement en ayant conscience de la présence d'un observateur, et en cherchant à maximiser un critère de performance lié aux croyances de cet observateur. Il est défini formellement par un 8-uplet $\langle S, A, T, \gamma, \mathcal{S}_f, \Theta, B, R\rangle$ où :

— $\langle \mathcal{S}, \mathcal{A}, T, \gamma, \mathcal{S}_f \rangle$ est un MDP sans fonction de récompense ;

— Θ est un ensemble fini de *types* possibles de l'agent, représentant une caractéristique de celui-ci telle que sa tâche réelle ou ses capacités ;

— $B : H^* \to \Delta^{|\Theta|}$ donne la croyance que l'observateur a sur le type de l'agent en fonction de l'historique des états et des actions ($H = \mathcal{S} \times \mathcal{A}$) ;

— $R : \mathcal{S} \times \mathcal{A} \times \Delta^{|\Theta|} \to \mathbb{R}$ est la fonction de récompense.

Dans la plupart des cas considérés par Miura et Zilberstein, B est obtenue en s'appuyant sur la définition de la mise-à-jour de croyance bayésienne BST de Baker, Saxe et Tenenbaum, c'est à dire en considérant que, du point de vue de l'agent, l'observateur modélise le comportement de l'agent pour une tâche donnée à travers un MDP en :

1. utilisant une fonction de récompense R_{MDP} approprié
2. résolvant le MDP $\langle \mathcal{S}, \mathcal{A}, T, \gamma, \mathcal{S}_f \rangle$ (où tout les composants excepté la fonction de récompense R_{MDP} émane de la définition de l'OAMDP) pour obtenir V^*_{MDP} ;
3. construisant une politique 'softmax' tel que pour chaque couple (s,a),

$$\pi_{\text{MDP}}(a|s) = \frac{e^{\frac{1}{\tau}Q^*_{\text{MDP}}(s,a)}}{\sum_{a'} e^{\frac{1}{\tau}Q^*_{\text{MDP}}(s,a')}}, \text{ où}$$
$$Q^*_{\text{MDP}}(s,a) = \sum_{s'} T(s,a,s') \cdot \left(r(s,a,s') + \gamma V^*_{\text{MDP}}(s')\right),$$

avec $\tau > 0$ représentant le niveau de rationalité de l'agent (considéré par l'observateur) pouvant être utilisé pour travailler avec des politiques plus ou moins optimale.

La croyance de l'observateur sur les types peut ensuite être obtenue par inférence bayésienne en utilisant π_{MDP}.

Miura et Zilberstein formalisent ainsi, entre autres, des problèmes de lisibilité, d'explicabilité, et de prédictibilité. Pour la prédictibilité, sur laquelle nous nous concentrons maintenant, Miura et Zilberstein proposent deux approches. La première repose sur les travaux de Dragan, Lee et Srinivasa, où la prédictibilité d'une trajectoire est modélisée comme étant proportionnelle à sa valeur (définit comme son coût négatif) [4]. Cela revient à optimiser la fonction de récompense R_{MDP}, donc à agir de manière gloutonne par rapport à Q^*_{MDP} (plutôt que de suivre π_{MDP}. La seconde approche repose sur la t-prédictibilité de Fisac, Liu, Hamrick et al. [2], laquelle maximise $Pr(a_{t+1}, \ldots, a_T | a_1, \ldots, a_t)$ dans des contextes déterministes en utilisant un type pour chaque trajectoire possible, c'est-à-dire un nombre exponentiel de types.

Dans la suite, nous proposons une approche alternative pour la prédictibilité et discutons ses propriétés.

3 Contribution

Kolobov, Mausam, Weld et al. [6] considèrent uniquement des OAMDP actualisés. Comme pour les MDP, on distinguera ici deux classes d'OAMDP : les OAMDP actualisés, et les OASSP (en utilisant $\gamma = 1$). En particulier, on se demandera sous quelles conditions un problème orienté but permet de construire un SSP valide.

3.1 MDP conscient d'un observateur et prédictible

Les deux approches de la prédictibilité proposées par Miura et Zilberstein s'inspirent de travaux dans des situations déterministes où il est naturel de raisonner sur les trajectoires. Parce qu'aussi bien la politique softmax π_{MDP} et la dynamique du système peuvent être stochastiques, on propose d'essayer de prédire soit l'action, soit l'état de l'agent,

chacune des deux alternatives pouvant amener à des résultats différents. Cependant, les types pour les OAMDP sont des variables statiques (comme les types des jeux bayésiens) alors que les actions et les états sont dynamiques. Cela nous amène à introduire les pOAMDP (OAMDP prédictibles), où le type (dynamique) est maintenant une fonction de la transition courante : $\theta_t = \tau(s_t, a_t, s_{t+1})$. Cela 1. ne permet pas d'encoder les problèmes où le type est statique et caché de l'observateur comme pour la lisibilité ou l'explicabilité, 2. mais permet toujours de définir et de résoudre le MDP de l'observateur (car le type n'influence pas la dynamique du système), et d'utiliser la mise à jour de croyance bayésienne (à cause de la nature markovienne des types dynamiques).

La section suivante décrit respectivement, pour la prédictibilité sur les actions et pour la prédictibilité sur les états, 1. comment dériver B et comment résoudre le pOAMDP étant donnée une fonction de récompense R, et 2. différentes fonctions de récompense candidates.

3.2 Fonction de croyance et propriété du pOAMDP

Pour la prédictibilité sur les actions, $\Theta = \mathcal{A}$, $\tau(s, a, s') = a$, et B est :

$$B : \begin{array}{ccc} H^* & \to & \Delta^{|\mathcal{A}|}, \\ (s_0, a_0, \ldots, s_t) & \mapsto & \pi_{\text{MDP}}(A_t | s_t). \end{array}$$

Pour la prédictibilité sur les états, $\Theta = \mathcal{S}$, $\tau(s, a, s') = s'$, et B est

$$B : \begin{array}{ccc} H^* & \to & \Delta^{|\mathcal{S}|}, \\ (s_0, a_0, \ldots, s_t) & \mapsto & \sum_{a'} \pi_{\text{MDP}}(a' | s_t) \cdot T(s, a', S_{t+1}). \end{array}$$

Dans les deux cas, B dépend uniquement de l'état courant, s_t, et on peut alors redéfinir la fonction de récompense du pOAMDP comme $R'(s_t, a_t) \stackrel{\text{def}}{=} R(s_t, a_t, B(s_t))$, et la croyance sur θ en s comme $b_s(\theta)$. Le problème de planification de l'agent peut alors être défini comme un MDP $\langle \mathcal{S}, \mathcal{A}, T, R', \gamma, \mathcal{S}_f \rangle$ qui peut être résolu par un algorithme comme Itération sur la valeur. En conséquence, la complexité de résolution correspond à la complexité de résolution de deux MDP : celui "du MDP de l'observateur", puis celui "du MDP induit par le pOAMDP". Dans le cas général [3], il n'est pas possible d'obtenir un tel MDP, et résoudre un OAMDP demande d'utiliser des algorithmes spécifiques dans lesquels le choix d'action est lié à l'historique état-action entière.

3.3 Possibles fonctions de récompense

Nous présentons ici 4 fonctions de récompenses candidates considérées qui peuvent être définies sur les états ou sur les actions.

[Confiance] $R^{\Theta}_{\textbf{max}}(s, a, s') \stackrel{\text{def}}{=} \max_\theta b_t(\theta)$ **:** Cette première fonction récompense l'agent proportionnellement à la plus grande croyance sur les types possibles dans l'état courant s_t : $\max_\theta b_t(\theta)$. En d'autres termes, elle favorise les états où la règle de décision immédiate prévue par l'observateur est plus déterministe. Comme on le verra dans les expérimentations, cette définition conduit à des comportements qui ne répondent pas à notre besoin, essentiellement parce que le choix de l'action courante n'influe pas sur la récompense immédiate. L'agent a tendance à préférer rester dans des états où la croyance de l'observateur est plus déterministe, ce qui peut conduire en pratique à des comportements peu prévisibles.

[Probabilité] $R^{\Theta}_{\textbf{pr}}(s, a, s') \stackrel{\text{def}}{=} b_t(\tau(s, a, s'))$ **:** Cette deuxième fonction de récompense favorise les états où la prochaine action ou le prochain état est plus prédictible. En d'autres termes, elle favorise le fait d'agir comme le prévoit l'observateur. Toutefois, cette fonction étant à valeurs positives, elle n'est pas appropriée pour les pOASSP, d'où les propositions suivantes.

[Regret] $R^{\Theta}_{\textbf{regret}}(s, a, s') \stackrel{\text{def}}{=} b_t(\tau(s, a, s')) - \max_\theta b_t(\theta)$ **:** Cette fonction de récompense repose sur le concept de regret (de faire un choix sous-optimal C à la place du choix optimal C^*). On notera que cette fonction de récompense est strictement négative, sauf quand $b_s(\theta)$ est maximale pour l'état s courant, auquel cas elle est nulle. Dans le cas de la prédictibilité sur les actions, les solutions optimales résultant de cette fonction de récompense sont les solutions optimales du MDP résolu par l'observateur. Ce résultat correspond à l'adaptation par Miura et Zilberstein de l'approche de Dragan, Lee et Srinivasa. Dans le cas de la prédictibilité sur les états, il existe des situations où les solutions optimales sont différentes. Un tel cas est illustré dans la partie expérimentale. Une question ouverte reste de savoir si R^{S}_{regret} induit toujours un SSP valide avec $\gamma = 1$

[Coût] $R^{\Theta}_{\textbf{cost}}(s, a, s') \stackrel{\text{def}}{=} b_t(\tau(s, a, s')) - 1$ **:** Cette seconde fonction de récompense négative est équivalente à R^{Θ}_{pr} dans le cadre des pOAMDP actualisés (qui sont, comme les MDP actualisés, invariants à l'ajout d'une constante à la fonction de récompense, et ont donc les mêmes solutions optimales).

Une première observation, détaillé dans les propositions suivante, est que cette fonction de récompense induit des pOASSP valides.

Proposition 1. *Supposons que (i) $\gamma = 1$, (ii) le MDP considéré par l'observateur est un SSP valide, et (iii) R^{A}_{cost} est la fonction de récompense du pOAMDP. Alors le pOAMDP est un problème bien défini car il induit un SSP valide.*

Démonstration. Raisonnons par l'absurde en supposant qu'il existe une politique optimale π^* pouvant atteindre

un sous-ensemble d'états $\mathcal{S}' \subset (\mathcal{S} \setminus \mathcal{S}_f)$ et y rester indéfiniment à "coût nul" (la fonction de récompense étant à valeurs négatives ou nulles). Or, agir à coût nul signifierait ici choisir, dans tout état $s \in \mathcal{S}'$, une action a telle que, pour tout s' possible, $R^A_{\text{cost}}(s,a,s') = 0 = \pi_{\text{MDP}}(a|s) - 1$, c.-à-d. $\pi_{\text{MDP}}(a|s) = 1$. Cela signifierait que, à l'intérieur de $\mathcal{S}'$, l'agent n'effectue que des actions optimales pour le MDP de l'observateur. Pourtant, comme le MDP de l'observateur est un SSP valide, ces actions optimales devrait faire sortir l'agent de $\mathcal{S}'$ (qui ne contient pas d'états terminaux) avec probabilité 1. □

Le même résultat peut être obtenu pour la prédictibilité sur les états. La preuve diffère en ce que la propriété $\pi_{\text{MDP}}(a|s) = 1$ est remplacée par le déterminisme de la fonction de transition T pour les actions échantillonnées.

Proposition 2. *Supposons que (i) $\gamma = 1$, (ii) le MDP considéré par l'observateur est un SSP valide, et (iii) R^S_{cost} est la fonction de récompense du pOAMDP. Alors le pOAMDP est un problème bien défini car il induit un SSP valide.*

Pour la prédictabilité sur les actions, nous pouvons aussi donner une interprétation du critère de performance obtenu. En effet, cette fonction de récompense est *l'opposé de la probabilité de* ne pas *prendre l'action qui serait échantillonnée si l'on suivait* π_{MDP}, ce qui s'écrit formellement $R^A_{\text{cost}}(s,a,s') = -P(A_{\text{MDP}} \neq a|S_t = s)$. Dans le cas $\gamma = 1$, la somme des récompenses sur une trajectoire s'écrit donc :

$$\sum_t R^A_{\text{cost}}(s_t, a_t, s_{t+1}) = -\sum_t P(A_{\text{MDP}} \neq a_t | S_t = s_t).$$

Ainsi, pour une politique π donnée qui atteint un état terminal avec probabilité 1 et pour un état s, $-V_\pi(s)$ est, quand on exécute π de s à un état terminal, *l'espérance du nombre d'actions échantillonnées en utilisant π_{MDP} (c.-à-d. les prédictions de l'observateur) en désaccord avec les actions réellement effectuées.*

Pour résumer cette section, les quatre fonctions de récompenses considérées sont

$$\begin{aligned}
R^A_{\max}(s,a,s') &\stackrel{\text{def}}{=} \max_{a'} \pi_{\text{MDP}}(a'|s),\\
R^S_{\max}(s,a,s') &\stackrel{\text{def}}{=} \max_{s''} \sum_{a'} \pi_{\text{MDP}}(a'|s) T(s,a',s''),\\
R^A_{\text{Pr}}(s,a,s') &\stackrel{\text{def}}{=} \pi_{\text{MDP}}(a|s),\\
R^S_{\text{Pr}}(s,a,s') &\stackrel{\text{def}}{=} \sum_{a'} \pi_{\text{MDP}}(a'|s) T(s,a',s'),\\
R^A_{\text{regret}}(s,a,s') &\stackrel{\text{def}}{=} \pi_{\text{MDP}}(a|s) - \max_{a'} \pi_{\text{MDP}}(a'|s),\\
R^S_{\text{regret}}(s,a,s') &\stackrel{\text{def}}{=} \sum_{a'} \pi_{\text{MDP}}(a'|s) T(s,a',s')\\
&\qquad - \max_{s''} \sum_{a'} \pi_{\text{MDP}}(a'|s) T(s,a',s''),\\
R^A_{\text{cost}}(s,a,s') &\stackrel{\text{def}}{=} \pi_{\text{MDP}}(a|s) - 1, \text{ and}\\
R^S_{\text{cost}}(s,a,s') &\stackrel{\text{def}}{=} \sum_{a'} \pi_{\text{MDP}}(a'|s) T(s,a',s') - 1.
\end{aligned}$$

La section suivante va les étudier sur des exemples simples.

4 Résultats expérimentaux

Le but des expérimentations est d'illustrer et de mieux comprendre les politiques obtenues à partir des fonctions de récompense. En particulier, on souhaite déterminer si ces politiques peuvent être considérées comme prédictible.

4.1 Protocole

Pour décrire les deux types de pOAMDP considérés dans nos expériences, détaillons les MDP correspondant pris en compte par l'observateur :

— un SSP, nommé *labyrinthe*, dans lequel l'agent se déplace dans un monde grille pour atteindre un état but terminal ;

— un MDP actualisé (sans état terminal), nommé *pompier*, dans lequel l'agent utilise pour éteindre des feux.

Pour faciliter les analyses, la plupart des problèmes ont une dynamique déterministe.

Problème MDP *labyrinthe* : Un *labyrinthe* (voir figure 3.a) est défini par un monde à grille avec des murs (cases grises), des cellules normales (en blanc), des cellules glissantes (en cyan), et des cellules terminales (disques roses). Plus formellement, dans ce SSP :

— chaque état s dans $\mathcal{S}$ indique les coordonnées (x, y) de l'agent dans une case normale, glissante ou terminale ;

— $\mathcal{S}_f$ est un sous-ensemble non-vide (mais aussi éventuellement non-singleton) de $\mathcal{S}$;

— $\mathcal{A} = \{up, down, left, right\}$;

— $T(s, a, s')$ encode les mouvements de l'agent : l'agent dans une cellule normale bouge dans la direction indiquée par son action si aucun mur ne l'empêche ; dans une cellule glissante, l'agent a une probabilité p (0.5 dans nos expérimentations) de faire un mouvement de 2 cellules plutôt qu'une (si possible) ; dans une cellule terminale, l'agent ne bouge pas ;

— R_{MDP}, la fonction de récompense, retourne (i) une pénalité par défaut de $-0,04$ pour chaque action, (ii) -1 si l'agent touche un mur, (iii) $+1$ s'il atteint un état terminal s_f, et (iv) 0 quand l'agent reste dans un état terminal.

Ce problème définit bien un SSP puisque toutes les récompenses ne menant pas un état terminal sont strictement négatives. La politique stochastique π_{MDP} est calculée avec l'algorithme d'itération sur la valeur avec $\gamma = 1$.

Problème MDP *pompier* : Le problème *pompier* utilise des grilles similaires, mais sans états terminaux, et avec des cellules représentant des feux et des sources d'eau (voir figure 8). L'agent a maintenant un réservoir d'eau, qui est vidé quand un feu (inextinguible) est attenit, et rempli quand une source d'eau (jamais vide) est atteinte. Plus formellement, dans ce MDP actualisé :

— chaque état s de $\mathcal{S}$ est représenté par un triplet (x, y, w) avec (x, y) les coordonnées de l'agent et w un booléen indiquant si le réservoir est plein ou vide ;

— $\mathcal{A} = \{up, down, left, right\}$;

— $T(s, a, s')$ est similaire au problème *labyrinthe*, sauf que w devient faux quand un feu est atteint, et vrai quand une source d'eau est atteinte ;

— R_{MDP}, la fonction de récompense, retourne (i) une pénalité par défaut de -0.04 pour chaque action, (ii) -1 quand l'agent touche un mur, et (iii) $+1$ quand l'agent atteint un feu alors qu'il transporte de l'eau (w =vrai).

Les politiques MDP optimales consistent en des allers-retours incessant entre une source d'eau et un feu. La politique softmax π_{MDP} est obtenue en utilisant l'algorithme d'itération sur la valeur avec $\gamma = 0.99$ pour garantir la convergence.

Modèle pOAMDP : Pour les deux types de problèmes, des pOAMDP sont dérivés en utilisant les fonctions de récompense précédemment proposées pour la prédictabilité. Nous construisons 8 pOAMDP (un par fonction de récompense) pour chaque environnement grille. Puisque chaque pOAMDP peut être considéré comme un MDP, les pOAMDP sont résolus en utilisant à nouveau l'algorithme d'itération sur la valeur avec un facteur d'actualisation approprié (détails dans la section suivante). Pour une fonction de récompense R^{Θ}_{X}, la politique solution du pOAMDP est notée π^{Θ}_{X}.

4.2 Résultats

Les figures représentent les politiques softmax π_{MDP} (avec des flèches dont le niveau de gris dépend de la probalitité d'action) et les politiques pOAMDP π^{Θ}_{X} (avec des flèches noires qui indiquent les actions ϵ optimales). Du fait de l'intérêt limité des politiques $\pi^{\Theta}_{\text{max}}$, π^{Θ}_{Pr}, et $\pi^{\Theta}_{\text{regret}}$, celles-ci ne sont montrées que dans le premier problème (labyrinthe).

4.2.1 Problème *labyrinthe* :

grilles utilisées : Les labyrinthes sont principalement constitués de couloirs et de pièces (vides). Pour la prédictibilité sur les actions, nous nous attendons à ce que les politiques pOAMDP préfèrent les couloirs aux pièces (qui permettent plus d'actions optimales possibles).

— Le labyrinthe de la fig. 3 consiste en 1 état terminal qui peut être atteint par 1 couloir ou 1 salle.

— Le labyrinthe de la fig. 4 consiste en 2 états terminaux qui peuvent être atteints respectivement par un couloir de 2 cellules de large et par un couloir d'1 cellule de large.

— Le labyrinthe de la fig. 5 consiste en 1 état terminal qui peut être atteint soit par un couloir suivi d'une salle, soit par une salle suivie d'un couloir. Le but de ce labyrinthe est d'observer l'influence de γ.

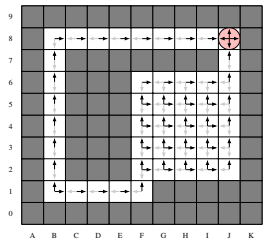
— Le labyrinthe de la fig. 6 consiste en 2 couloirs qui conduisent à un état terminal. Un de ces couloirs contient des cellules glissantes, mais le temps moyen de traversée est le même pour les deux. Le but de ce labyrinthe est d'observer les différences entre R^{A}_{cost} et R^{S}_{cost}.

Softmax MDP policy π_{MDP} Chaque SSP est résolu avec $\gamma = 1$ et $\epsilon = 0.001$. Une politique softmax π_{MDP} est ensuite obtenue en utilisant $\tau = 0.1$. Comme attendu, dans une salle, il y a de multiples actions optimales.

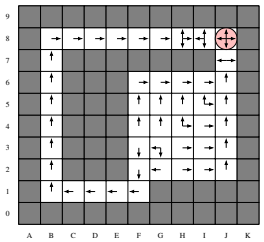
Note : Dans la suite, nous nous concentrons sur la prédictibilité des actions parce que les politiques solutions s'avèrent identiques pour le cas de la prédictibilité sur les états. Ce phénomène est favorisé par les environnements déterministes, où prédire le prochain état est souvent équivalent à prédire la prochaine action.

Analyse de π^{A}_{max} (et π^{S}_{max}) Les deux fonctions de récompenses π^{A}_{max} et π^{S}_{max} sont positives et sont à l'origine de comportements qui empêchent la convergence dans le cas ou $\gamma = 1$. On résout donc le pOAMDP avec $\gamma = 0.99$.

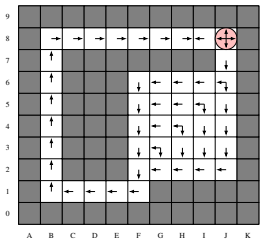
Comme attendu, la politique π^{A}_{max} tente d'atteindre des états dans lesquels l'observateur a une plus grande confiance, c.-à-d. que sa croyance est plus déterministe. L'agent peut choisir une action pour rester dans un état



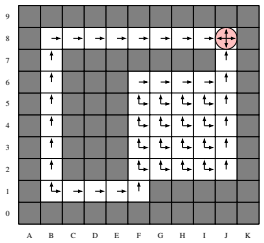
a) π_{MDP} ($\gamma = 1$, $\tau = 0.1$)



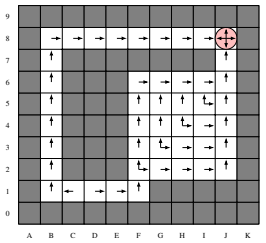
b) π^A_{max} ($\gamma = 0.99$)



c) π^A_{Pr} ($\gamma = 0.99$)

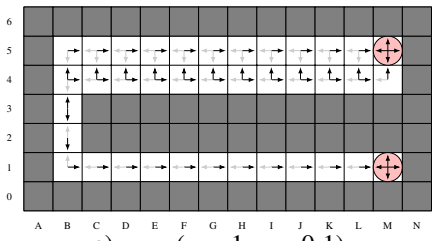


d) π^A_{regret} ($\gamma = 1$)

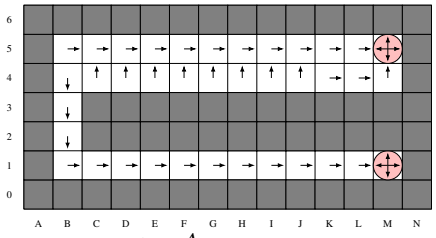


e) π^A_{cost} ($\gamma = 1$)

FIGURE 3 – Résultats de la prédictibilité sur les actions pour le premier labyrinthe, en commençant par la politique stochastique attendue par l'observateur (a) avec une température de $\tau = 0.1$, et en montrant ensuite toutes les actions optimales pour les 4 fonctios de récompenses considérées (b–e), avec $\gamma < 1$ quand le problème n'est pas un SSP valide.



a) π_{MDP} ($\gamma = 1$, $\tau = 0.1$)



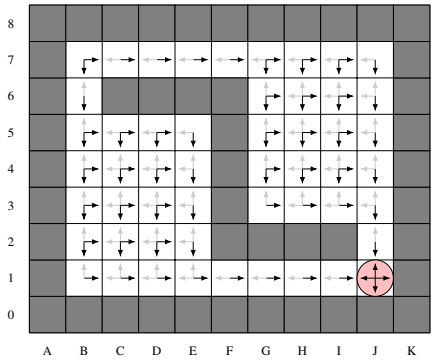
b) π^A_{cost} ($\gamma = 1$)

FIGURE 4 – Résultats pour le deuxième labyrinthe

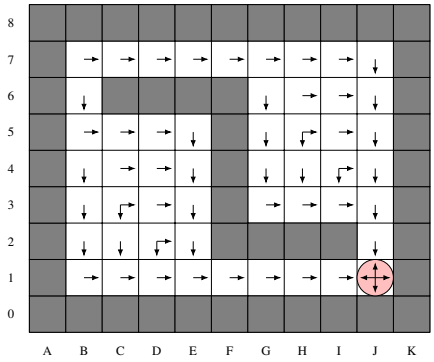
plus déterministe même si cette action n'est pas prédictible (voir fig. 3.b, cellules $(H, 8)$, $(I, 8)$ et $(J, 7)$). Il évite aussi les états terminaux pour continuer d'accumuler des récompenses positives.

Analyse de π^A_{Pr} (et π^S_{Pr}) Pour des raisons identique au cas précédent, on utilise $\gamma = 0.99$ pour résoudre le pOAMDP.

Pour la prédictibilité sur les actions, donc avec R^A_{Pr}, les états terminaux ne sont toujours pas récompensés, ce qui, à nouveau, dissuade l'agent de l'atteindre. π^A_{Pr} diffère de π^A_{max} parce que la récompense encourage à prendre les actions



a) π_{MDP} ($\gamma = 1$, $\tau = 0.1$)



b) π^A_{cost} ($\gamma = 0.99$)

FIGURE 5 – Résultats pour le troisième labyrinthe

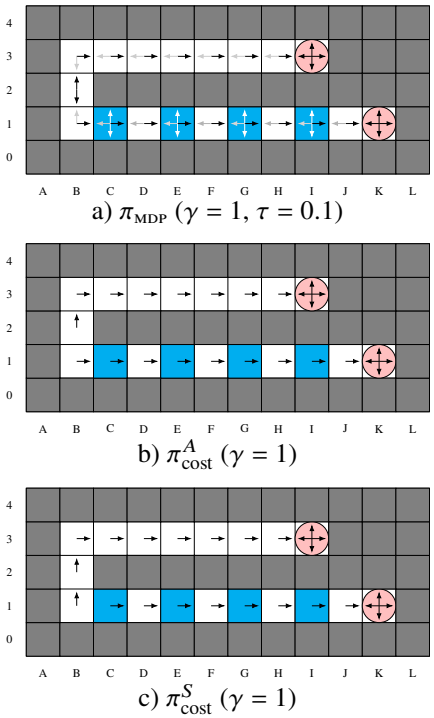


FIGURE 6 – Résultats pour le quatrième labyrinthe

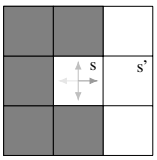


FIGURE 7 – Partie de labyrinthe dans laquelle π^S_{regret} peut différer de π^A_{regret} dans l'état s (et donc de la politique MDP optimale) car il est plus probable pour l'observateur (à cause de π_{MDP}) que le prochain état soit s au lieu de s'.

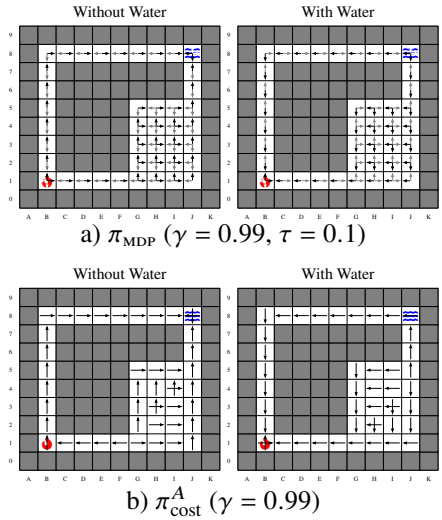


FIGURE 8 – Résultats pour le premier problème pompier

les plus prédites par l'observateur, comme on le voit sur la fig. 3.c, cellules $(H, 8)$, $(J, 8)$, $(J, 7)$, et $(J, 8)$.

Analyse de π^A_{regret} et π^S_{regret} Les fonctions de récompense "regret" induisent ici des SSP valides. Comme attendu (voir sec. 3.3), $R^A_{\text{regret}}(s, a, s')$ conduit aux solutions optimales du SSP de l'observateur, comme on le voit en comparant les actions les plus probables de la fig. 3.a et les actions (toutes optimales) de la fig. 3.d.

Comme précédemment, π^A_{regret} et π^S_{regret} s'avèrent identiques sur le premier problème. Mais regardons, sur la fig. 7, un motif de labyrinthe qui conduit à des comportements locaux dans π^S_{regret} différents de ceux de la politique optimale du MDP de l'observateur (et donc différents de π^A_{regret}). Ici, l'action optimale du MDP est de sortir de cette impasse s en allant vers la droite. Pourtant, en supposant une température assez élevée τ et des pénalités assez petites quand on touche le mur, la politique softmax pourrait être telle que $\pi_{\text{MDP}}(left|s) + \pi_{\text{MDP}}(up|s) + \pi_{\text{MDP}}(down|s) > \pi_{\text{MDP}}(right|s)$, de sorte que 1. le prochain état le plus probable est s plutôt que s', et 2. π^S_{regret} choisira n'importe quelle action autre que $right$.

Analyse de π^A_{cost} et π^S_{cost} Nous observons plusieurs comportements intéressants avec $R^A_{\text{cost}}(s, a, s')$:

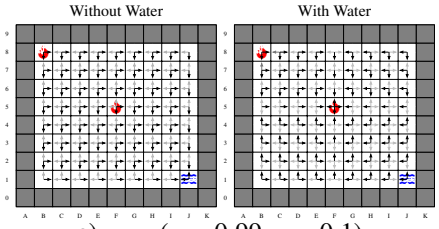
1. L'agent préfère un long chemin à travers un couloir étroit à un chemin plus court passant par une salle (fig. 3.e) ou un couloir large (fig. 4.b). Il y a en effet moins de choix d'actions dans les couloirs étroits, de sorte que les actions sont plus prédictibles.
2. Dans les salles, l'agent va souvent vers le mur le plus proche pour le suivre, comme sur la Fig. 3.e et la fig. 5.b.
3. Dans la fig. 5, l'agent peut choisir entre (i) un couloir conduisant à une pièce, et (ii) une pièce conduisant à un couloir. Quand $\gamma < 1$, l'agent préfère passer par le couloir d'abord parce que le facteur d'actualisation met plus d'importance sur les récompenses proches (voir cellule $(B, 7)$).
4. Dans la fig. 6, cellule $(B, 2)$, π^A_{cost} préfère monter plutôt que descendre. Cela est dû à la dynamique moins régulière le long du chemin du bas, lequel, à travers les Q-valeurs, conduit à de petites différences dans les prédictibilités des actions.

R^S_{cost} conduit à un résultat différent de R^A_{cost} dans le labyrinthe de la figure 6, parce que π^S_{cost} préfère monter dans la cellule $(B, 1)$, ce qui va à l'encontre des prédictions de l'observateur, pour suivre le chemin sans cellules glissantes.

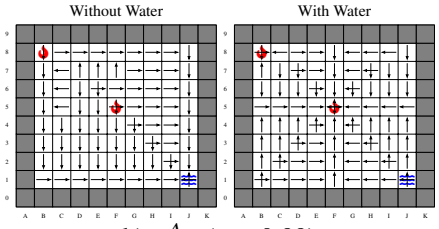
4.2.2 Problème du *pompier*

Grilles utilisées Les grilles suivantes ont été utilisées pour tester les fonctions de récompense :

1. la grille de la fig. 8 contient 1 feu et 1 source d'eau reliés par une salle et un couloir ;

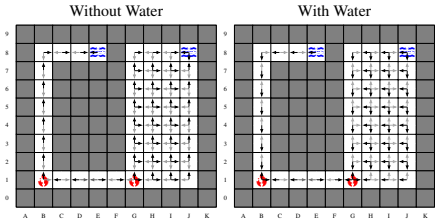


a) π_{MDP} ($\gamma = 0.99$, $\tau = 0.1$)

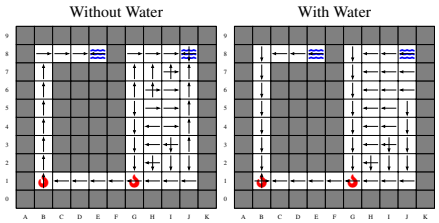


b) π^A_{cost} ($\gamma = 0.99$)

FIGURE 9 – Résultats pour le deuxième problème pompier



a) π_{MDP} ($\gamma = 0.99$, $\tau = 0.1$)



b) π^A_{cost} ($\gamma = 0.99$)

FIGURE 10 – Résultats pour le troisième problème pompier

2. la grille de la fig. 9 est une salle avec 2 feux et 2 sources d'eau ;
3. la grille de la fig. 10 contient 2 feux et 2 sources d'eau ; une partie de la carte est une pièce et l'autre un couloir ; le but de cette carte est d'observer si l'agent pOAMDP préfère apporter de l'eau au feu de la zone déterministe même si le feu situé sur la salle est plus proche.

Comparaison des politiques Les MDP sous-jacents ne sont plus des SSP, de sorte ue nous employons des pOAMDP $\gamma = 0.99$-actualisés. Comme pour le problème du labyrinthe : 1. R^A_{max} et R^S_{max} créeent des politiques qui préfère "boucler" qu'effectuer la tâche ; 2. R^A_{regret} et R^S_{regret} induisent la même politique que la politique optimale du MDP de l'observateur, et ne sont donc pas utiles ; et 3. dans la plupart des cas, les prédictibilités sur les actions et les états donnent des résultats similaires. Aussi, R^A_{Pr} et R^A_{cost} (ou R^S_{Pr} et R^S_{cost}) induisent les mêmes politiques solutions dans ce cadre actualisé. Pour toutes ces raisons, nous n'étudions que R^A_{cost}.

Analyse des politiques π^A_{cost} et π^S_{cost} Un comportement similaire au cas du problème labyrinthe peut être observé. Dans la fig. 8, π^A_{cost} préfère le couloir à la salle vide. Dans de telles salles, l'agent pOAMDP cherche à atteindre un mur pour le longer (figs. 8 and 10). Dans la fig. 9, l'agent pOAMDP tente d'être plus prédictible en marchant le long des murs ou en atteignant la ligne 5 ou la colonne F pour réduire le nombre de chemins optimaux pour atteindre le feu au milieu. Dans la fig. 10, l'agent pOAMDP préfère le feu situé en $(B, 1)$ et la source d'eau située en $(E, 8)$ même si un autre feu ou une autre source d'eau est plus proche. C'est particulièrement visible sur le côté "sans eau" de la figure, où π^A_{cost} va de $(G, 5)$ à $(E, 8)$ pour se remplir.

5 Conclusion et perspectives

Nous avons introduit un nouveau formalisme, celui des OAMDP prédictibles (pOAMDP), lequel permet de dériver des politiques dans lesquelles la prochaine action ou le prochain état est plus prédictible, et proposé de prendre en compte non seulement les problèmes actualisés, mais aussi les chemins stochastiques les plus courts (ce qui requiert de s'assurer que des politiques solutions valides peuvent être trouvées). Différentes fonctions de récompense ont été considérées et analysées à travers leurs propriétés théoriques et des illustrations des politiques résultantes sur deux mondes grilles. La fonction de récompense "coût" $R^\Theta_{\text{cost}}(s, a, s') \stackrel{\text{def}}{=} b_t(\tau(s, a, s')) - 1$ ($\Theta \in \{A, S\}$) s'avère être le meilleur choix, puisqu'elle est valide à la fois pour les problèmes actualisés et orientés-but, et rend effectivement les actions ou états plus prédictibles (mais des alternatives sont possibles). Dans certains cas, des actions contre-intuitives sont sélectionnées pour augmenter la prédictibilité ultérieure. Un propriété remarquable est que le complexité de résolution des pOAMDP est comparable à celle des MDP, et bien moindre que celle des OAMDP.

Une première perspective serait de conduire des expérimentations avec de vrais observateurs humains pour voir si les politiques pOAMDP sont effectivement perçues comme plus prédictibles, et de raffiner les exigences pour un agent prédictible. Par exemple, on pourrait s'attendre à ce que les humains arrêtent de faire confiance à l'agent si son comportement est temporairement non-prédictible.

Par ailleurs, pour revenir au travail précurseur de MIURA et ZILBERSTEIN [3], nous souhaiterions étendre la discussion des problèmes orientés-buts aux OAMDP, par exemple pour déterminer lesquels de leurs scénarios conduisent à des SSP valides. Aussi, nous avons dû nous éloigner de leur formalisme original et de leurs types statiques, mais une perspec-

tive importante est de généraliser les deux formalismes pour obtenir une théorie plus unifiée de la prise de décision séquentielle consciente d'un observateur Nous pensons que, pour se faire, un point clef est de restreindre l'observabilité des états et actions par l'observateur, de sorte que le type, qu'il soit statique ou dynamique, puisse être une variable d'état (même pour la prédictibilité sur les actions). En outre, cette observabilité partielle permettre aussi de couvrir plus de scénarios du monde réel. Dans ce cadre, nous envisageons d'étudier les propriétés de continuité de la fonction de valeur optimale pour éventuellement proposer des approximateurs minorant et majorant, et dériver des solveurs à base de points (comme cela a été fait pour les POMDP et des modèles apparentés [7]-[14]).

Références

[1] T. Chakraborti, A. Kulkarni, S. Sreedharan, D. E. Smith et S. Kambhampati, "Explicability ? Legibility ? Predictability ? Transparency ? Privacy ? Security ? The Emerging Landscape of Interpretable Agent Behavior," in *Proceedings of the Twenty-Ninth International Conference on Automated Planning and Scheduling (ICAPS)*, Berkeley, CA, USA : AAAI Press, 2019. adresse : https://ojs.aaai.org/index.php/ICAPS/article/view/3463.

[2] J. F. Fisac, C. Liu, J. B. Hamrick et al., "Generating plans that predict themselves," in *Algorithmic Foundations of Robotics XII : Proceedings of the Twelfth Workshop on the Algorithmic Foundations of Robotics*, 2020.

[3] S. Miura et S. Zilberstein, "A unifying framework for observer-aware planning and its complexity," in *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, C. de Campos et M. H. Maathuis, éd., sér. Proceedings of Machine Learning Research, t. 161, PMLR, juill. 2021, p. 610-620. adresse : https://proceedings.mlr.press/v161/miura21a.html.

[4] A. D. Dragan, K. C. T. Lee et S. S. Srinivasa, "Legibility and predictability of robot motion," 2013, p. 301-308.

[5] C. L. Baker, R. Saxe et J. B. Tenenbaum, "Action understanding as inverse planning," *Cognition*, t. 113, n° 3, p. 329-349, déc. 2009. doi : 10.1016/j.cognition.2009.07.005.

[6] A. Kolobov, Mausam, D. S. Weld et H. Geffner, "Heuristic Search for Generalized Stochastic Shortest Path MDPs," in *Proceedings of the International Conference on Automated Planning and Scheduling (ICAPS'11)*, 2011.

[7] T. Smith et R. G. Simmons, "Point-Based POMDP Algorithms : Improved Analysis and Implementation," in *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, 2005, p. 542-549.

[8] M. T. Spaan et N. Vlassis, "Perseus : Randomized Point-based Value Iteration for POMDPs," *Journal of Artificial Intelligence Research*, t. 24, p. 195-220, 2005. adresse : http://www.aaai.org/Papers/JAIR/Vol24/JAIR-2406.pdf.

[9] H. Kurniawati, D. Hsu et W. S. Lee, "SARSOP : Efficient point-based POMDP planning by approximating optimally reachable belief spaces," in *Robotics : Science and Systems IV*, 2008.

[10] J. Pineau, G. Gordon et S. Thrun, "Anytime point-based approximations for large POMDPs," *Journal of Artificial Intelligence Research*, t. 27, p. 335-380, 2006.

[11] G. Shani, J. Pineau et R. Kaplow, "A survey of point-based POMDP solvers," *Journal of Autonomous Agents and Multi-Agent Systems*, t. 27, n° 1, 2013. doi : 10.1007/s10458-012-9200-2. adresse : http://dx.doi.org/10.1007/s10458-012-9200-2.

[12] J. Dibangoye, C. Amato, O. Buffet et F. Charpillet, "Optimally Solving Dec-POMDPs as Continuous-State MDPs," *Journal of Artificial Intelligence Research*, t. 55, p. 443-497, 2016. adresse : http://www.jair.org/papers/paper4623.html.

[13] K. Horák, B. Bošanský et M. Pěchouček, "Heuristic Search Value Iteration for One-Sided Partially Observable Stochastic Games," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017, p. 558-564.

[14] K. Horák et B. Bošanský, "Solving Partially Observable Stochastic Games with Public Observations," in *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, 2019, p. 2029-2036. doi : 10.1609/aaai.v33i01.33012029.