Un cadre pour la planification consciente d'un observateur sous observabilité partielle

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 planification où l'agent est conscient de la présence d'un observateur et où cet observateur est en situation d'observabilité partielle. L'agent doit donc choisir sa stratégie dans le but d'optimiser les informations qu'il transmet à travers les observations. Nous proposons un cadre qui permet de traiter ce type de problème et de travailler avec différentes propriétés telles que la prédictibilité, la lisibilité et l'explicabilité. Notre travail s'appuie sur le cadre des processus de décision markoviens conscients d'un observateur (OAMDP). Étendre les OAMDP en observabilité partielle permet d'une part de travailler sur des problèmes plus réalistes (des situations où l'observateur n'aurait pas accès à l'ensemble des données de l'environnement), mais permet aussi de considérer des variables cibles dynamiques. Ces types dynamiques permettent de traiter la prédictibilité telle que présentée dans les pOAMDP (predictable OAMDP) ainsi que des problèmes de lisibilité à objectifs multiples où l'objectif de l'agent pourrait changer au cours du temps.

Abstract

In this article, we are interested in planning problems where the agent is aware of the presence of an observer, and where this observer is in a partial observability situation. The agent has to choose its strategy to optimize the information transmitted by observations. We build a framework to handle those kinds of problems and work with various properties such as predictability, legibility and explicability. Our work is based on the Observer Aware Markov Decision Process (OAMDP) framework. The extension of OAMDPs to partial observability can handle more realistic problems (situations where the observer doesn't have access to all of the environment information) but also allows to consider dynamic types. Those dynamic target variables allow to work with predictability as presented in the pOAMDP (predictable OAMDP) framework and with legibility problems with multiple goals where the goal might change during the task.

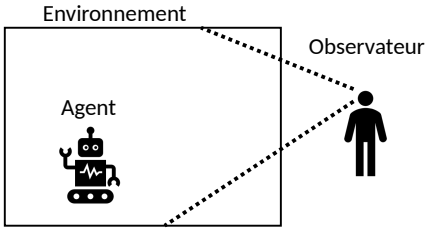

FIGURE 1 : Agent dans son environnement en présence d'un observateur passif

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

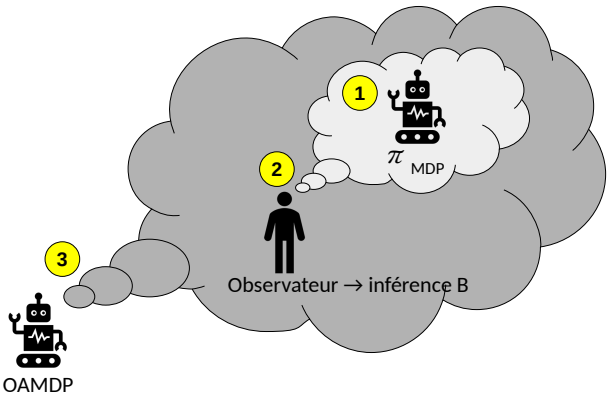

FIGURE 2 : Un agent OAMDP (3) fait l'hypothèse que l'observateur s'attend (2) à ce que l'agent se comporte de manière à accomplir une certaine tâche (1).

sous l'hypothèse que les transitions sont déterministes, raisonnant donc principalement sur des plans (une séquence d'actions 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.

MIURA et ZILBERSTEIN [2], 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*). Ils font l'hypothèse que, du point de vue de l'observateur, l'agent effectue sa tâche en ignorant la présence de l'observateur. C'est une hypothèse réaliste dans un grand nombre de contextes. En outre, faire l'hypothèse contraire (l'observateur suppose que l'agent essaye d'aider son inférence) induirait un problème de poule et d'œuf, les deux cherchant à se modéliser l'un l'autre. Entre autres choses, ils travaillent aussi sur l'explicabilité, la lisibilité et la prédictibilité.

LEPERS, THOMAS et BUFFET [3] ont plus récemment proposé une nouvelle façon de modéliser la prédictibilité en s'inspirant du cadre OAMDP, et en proposant une approche plus adaptée aux environnements incertains. La variable cible n'étant plus un type statique, mais la prochaine action ou le prochain état de l'agent, donc une variable dynamique, il a fallu introduire un nouveau cadre, celui des pOAMDP (predictable OAMDP). L'objectif de cet article est de proposer un modèle qui permette de traiter à la fois des problèmes avec un type statique (lisibilité, explicabilité pour les OAMDP), des problèmes avec des variables cibles dynamiques (prédictibilité des pOAMDP) et des problèmes en observabilité partielle. Dans cette dernière situation, l'observateur n'a alors plus forcement accès à l'état et à l'action de l'agent mais à une observation liée à la transition suivie par le système. L'introduction d'observabilité partielle permet de travailler sur des problèmes plus divers et plus proches de la réalité tels que des situations où l'observateur n'aurait pas accès à l'ensemble des informations de l'environnement. Nous pouvons par exemple considérer des situations où l'agent PO-OAMDP n'est pas toujours visible et doit choisir d'utiliser certains passages pour être vu par l'observateur et lui permettre de mieux inférer la situation courante.

La section 2 introduit des pré-requis sur le processus de décision markoviens (MDP) et les MDP conscients d'un observateur. Notre approche général est décrite en section 3, la résolution des PO-OAMDP est décrite en section section 4 avant de conclure en section 5.

2 Pré-requis

2.1 Processus de décision markovien

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 $s, a \in \mathcal{S}_f \times \mathcal{A}$, $T(s, a, s) = 1$ et $R(s, a, s) = 0$.

Une politique $\pi_{\text{OBS}} : \mathcal{S} \to \Delta(\mathcal{A})$ détermine un comportement en associant à chaque état une action à effectuer. Elle peut éventuellement être stochastique, $\pi_{\text{OBS}}(a|s)$ 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_{OBS} en un état s est l'espérance de la somme des récompenses actualisées sur un horizon infini :

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

Il existe toujours au moins une politique ${\pi_{\text{OBS}}}^*$, dite optimale, telle que, pour tout s, $V^{{\pi_{\text{OBS}}}^*}(s) = \max_{\pi_{\text{OBS}}} V^{\pi_{\text{OBS}}}(s)$. 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 interrompt les calculs quand le *résidu de Bellman* est inférieur à un seuil fonction de l'erreur ϵ souhaitée et de γ :

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

une politique déterministe ϵ-optimale étant alors obtenue en agissant de "manière gourmande" dans tout état s avec :

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

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 $(s,a,s') \in (\mathcal{S} \setminus \mathcal{S}_f) \times \mathcal{A} \times \mathcal{S}$, on a $r(s,a,s') < 0$, 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 Processus de décision markovien conscient d'un observateur

Un *MDP conscient d'un observateur* (OA-MDP pour *observer-aware MDP*) 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 $\langle \mathcal{S}, \mathcal{A}, T, \gamma, \mathcal{S}_f, \Theta, B, R \rangle$ 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 (la *croyance* sur une variable aléatoire est la distribution de probabilité sur ses valeurs possibles étant données les informations disponibles) en fonction de l'historique des états et des actions ($H \stackrel{\text{def}}{=} \mathcal{S} \times \mathcal{A}$) ;
- $R : \mathcal{S} \times \mathcal{A} \times \Delta^{|\Theta|} \to \mathbb{R}$ 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 :

1. en utilisant une fonction de récompense R_{OBS} appropriée ;
2. en résolvant le MDP $\langle \mathcal{S}, \mathcal{A}, T, \gamma, \mathcal{S}_f \rangle$ (où tous les composants, exceptée la fonction de récompense R_{OBS}, émanent de la définition de l'OAMDP) pour obtenir V^*_{OBS} ; et
3. en construisant une politique "softmax", c'est-à-dire telle que, pour chaque couple (s,a),

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

$\tau > 0$ représentant le niveau de rationalité de l'agent (considéré par l'observateur) afin de pouvoir raisonner sur des politiques plus ou moins proches de la politique optimale.

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

Miura et Zilberstein formalisent ainsi, entre autres, des problèmes de lisibilité, d'explicabilité, et de prédictibilité.

Note : Comme déjà fait ci-dessus, on indicera souvent par "OBS" les quantités liées au point de vue de l'observateur (tel que perçu par l'agent). Entre autres, certaines probabilités seront calculées du point de vue de l'observateur, et notées P_{OBS}. Aussi, on écrira parfois une fonction $f(X,Y)$ décrivant une distribution de probabilité conditionnelle sous la forme $f(Y|X)$ pour faire ressortir les dépendances entre variables.

3 Contribution : MDP conscient d'un observateur en observabilité partielle

Comme vu en introduction, on souhaite proposer un modèle OAMDP en observabilité partielle, lequel permettrait à la fois de traiter plus de scénarios (situations où l'observateur ne voit pas toujours le robot) et traiter des propriétés qui nécessitent l'utilisation de variables cibles dynamiques.

3.1 Formalisme

Dans le cadre PO-OAMDP, l'agent a accès à l'état complet du système, et l'observateur n'en a désormais qu'une perception partielle. Sa construction part de l'ajout au formalisme OAMDP d'un ensemble d'observations et d'une fonction d'observation. Nous allons expliquer cette construction avant de donner une définition formelle. Dans ce contexte, le type est désormais nommé *variable cible* et peut évoluer au cours du temps, contrairement au type statique des OAMDP. Afin de rester souple et générique dans la définition de ce qu'est la variable cible, sa valeur à chaque pas de temps est le résultat d'une fonction prenant en entrée la transition suivie par le système. La variable cible peut donc être juste une sous-partie de l'état du système (par exemple une variable non observable par l'observateur), mais elle peut aussi être liée à l'action émise par l'agent (pour des problèmes de prédictibilité) ou à l'évolution de l'état plus qu'à l'état lui même. Évidemment, cette variable cible peut regrouper en son sein plusieurs variables différentes. Nous ne parlons que d'une unique variable que par commodité mais sans perte de généralité.

En outre, on suppose que, en plus de l'état complet du système, l'agent a accès aux observations reçues par l'observateur (ce qui reste réaliste dans de nombreuses situations, en particulier si le processus d'observation est déterministe, auquel cas les observations reçues par l'observateur sont facilement prédictibles). L'agent peut ainsi construire l'état interne de l'observateur au fur et à mesure de l'exécution de son comportement.

En ayant accès à toutes les informations du problème (l'état du système, les actions effectuées et les observations perçues par l'observateur), l'agent va planifier ses actions pour chercher à contrôler l'inférence faite par l'observateur sur sa variable cible.

Formellement, un PO-OAMDP est ainsi défini par un n-uplet $\langle \mathcal{S}, \mathcal{A}, T, \gamma, \mathcal{S}_f, R_{\text{OBS}}, \Theta, \Omega, O, B, R_{\text{AG}}, \phi \rangle$, où :

- $\langle \mathcal{S}, \mathcal{A}, T, \gamma, \mathcal{S}_f, R_{\text{OBS}} \rangle$ est un MDP ;
- Θ désigne une *variable cible* (dynamique), mais aussi l'ensemble fini de valeurs qu'elle peut prendre ; nous changeons de terminologie pour souligner la différence entre cette variable (dynamique) et la variable *type* des OAMDP ;
- $\phi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \Theta$ est une fonction qui donne la valeur de la variable cible en fonction de la transition : $\theta_t = \phi(s_t, a_t, s_{t+1})$;
- Ω est un ensemble fini d'observations ;
- $O : \mathcal{A} \times \mathcal{S} \to \Omega$ est la fonction d'observation ; $O(a, s', o)$ est la probabilité d'émission d'une observation o si l'action a conduit dans l'état s' ;
- $B : \Omega^* \to \Delta^{|\mathcal{S}|}$ donne la croyance que l'observateur a sur l'état en fonction de l'historique des observations ; la croyance sur la variable cible pourra en être déduite (voir section 3.2) ; on notera $b_t \stackrel{\text{def}}{=} B(o_1, \ldots, o_{t-1})$, en appelant croyance initiale $b_0 \stackrel{\text{def}}{=} B()$;
- $R_{\text{AG}} : \mathcal{S} \times \Delta^{|\Theta|} \times \mathcal{A} \times \mathcal{S} \times \Delta^{|\Theta|} \to \mathbb{R}$ est la fonction de récompense de l'agent sous sa forme la plus générale : $R_{\text{AG}}(s_t, \beta_t, a_t, s_{t+1}, \beta_{t+1})$.

Alors que dans un POMDP, la mise à jour de la croyance dépend nécessairement des actions effectuées choisies par l'agent lui même (incluses dans l'historique), dans un PO-OAMDP, les actions ne sont pas forcément connues de l'observateur. Les observations peuvent inclure les actions selon le scénario considéré.

Nous ferons ici l'hypothèse que l'observation reçue permet à l'observateur de savoir à chaque instant si un état terminal a été atteint ou non, sans nécessairement indiquer duquel il s'agit (ce qui pourra dépendre du problème).

On notera que le modèle PO-OAMDP repose sur un seul MDP sous-jacent, contrairement au modèle OAMDP, qui associe un MDP à chaque type possible. Dans notre cadre à observabilité partielle, n'employer qu'un seul MDP sous-jacent n'est toutefois pas restrictif. On pourrait montrer formellement que tout OAMDP peut se ré-écrire comme un PO-OAMDP.

On notera aussi que l'introduction de la variable cible n'a rien de nécessaire. On pourrait obtenir un formalisme équivalent en écrivant une fonction de récompense directement sur la croyance sur les états au lieu de la croyance sur la variable cible. L'introduction de la variable cible est une commodité pour afficher le lien avec les OAMDP et pour faciliter la modélisation des problèmes.

La suite de cette section décrit comment la croyance de l'observateur (sur l'état) est mise à jour, et comment en déduire la croyance sur la variable cible, laquelle permet de calculer la récompense de l'agent lors d'une transition. Elle illustre ensuite les usages possibles du cadre PO-OAMDP sur différents scénarios.

3.2 Calcul des croyances sur l'état et la variable cible

Comme dans le cadre OAMDP, l'observateur calcule la politique de l'agent étant donnée la fonction de récompense qu'il connaît, R_{OBS}. L'observateur modélise le comportement de l'agent pour une tâche donnée à travers un MDP :

1. en utilisant une fonction de récompense R_{OBS} ;
2. en résolvant le MDP sous-jacent ; et
3. en construisant une politique softmax.

On peut noter que, étant données la dynamique (transitions+observations) du PO-OAMDP et la politique π_{OBS} de l'agent, l'observateur fait face à un HMM : il résout un

problème de *filtrage* en devant estimer la croyance sur l'état s_t en fonction de l'historique d'observation $o_{1:t}$.

La croyance de l'agent peut ensuite être construite à partir de la politique de l'agent, de la fonction de transition du MDP sous-jacent et de la fonction d'observation :

$$\begin{aligned} B(s_{t+1}|o_{1:t+1}) = P(s_{t+1}|o_{1:t+1}) &= \frac{P(s_{t+1}, o_{1:t}, o_{t+1})}{P(o_{1:t+1})} \\ &= \frac{P(s_{t+1}, o_{1:t+1})}{\sum_{s_{t+1}} P(s_{t+1}, o_{1:t+1})} \\ &= \frac{K(s_{t+1}, o_{1:t+1})}{\sum_{s_{t+1}} K(s_{t+1}, o_{1:t+1})}, \text{ avec} \\ K(s_{t+1}, o_{1:t+1}) &\stackrel{\text{def}}{=} \sum_{a_t} O(o_{t+1}|a_t, s_{t+1}) \sum_{s_t} T(s_{t+1}|s_t, a_t) \cdot \\ &\quad \pi_{\text{OBS}}(a_t|s_t) \cdot B(s_t|o_{1:t}). \end{aligned}$$

Croyance sur la variable cible Pour déterminer la récompense reçue lors d'une transition, il est aussi nécessaire de calculer la croyance β sur la valeur que va prendre la variable cible : $\Theta_t = \phi(S_t, A_t, S_{t+1})$. Cela peut être réalisé en partant de la croyance b de l'observateur sur l'état courant, S_t, comme suit :

$$\begin{aligned} \beta(\theta) &= \sum_{s,a,s'} \mathbb{1}_{\theta=\phi(s,a,s')} \cdot P_{obs}(s, a, s'|b) \\ &= \sum_{s,a,s'} \mathbb{1}_{\theta=\phi(s,a,s')} \cdot P_{obs}(s'|s, a) \cdot P_{obs}(a|s) \cdot P_{obs}(s|b) \\ &= \sum_{s,a,s'} \mathbb{1}_{\theta=\phi(s,a,s')} \cdot T(s, a, s') \cdot \pi_{\text{OBS}}(a|s) \cdot b(s). \end{aligned} \quad (1)$$

3.3 Mise en œuvre sur divers scénarios

Le modèle PO-OAMDP permet de construire différents comportements en faisant varier Θ et R et donc d'aborder différents types de problèmes.

Nous allons commencer par montrer comment des OAMDP peuvent être reformulés comme des PO-OAMDP. Dans un OAMDP, l'agent a un *type* statique qui le caractérise. Dans le cadre PO-OAMDP, cela peut se traduire par une variable d'état (cachée) dont la valeur est extraite par $\theta = \phi(s)$.

3.3.1 Expression de B et R' pour différentes propriétés

Lisibilité La lisibilité réduit l'ambiguïté sur les buts possibles de l'agent. Dans cette situation, l'agent a plusieurs buts possibles inconnus de l'observateur. Un comportement lisible transmet le but (ou, plus généralement, le critère de performance) de l'agent à travers ses choix d'actions.

Θ : Dans le cas de la propriété de lisibilité, le type va donc caractériser ce critère de performance parmi un ensemble fini de critères possibles. Cela va se traduire typiquement par le fait que la fonction de récompense dépend de ce type, mais pas nécessairement la fonction de transition ou la fonction d'observation.

R_{AG} : Pour la fonction de récompense, Miura et Zilberstein utilisent l'opposé de la distance euclidienne à la "croyance idéale". La croyance idéale étant définie par : $\beta^*(s) = (0, \ldots, 0, 1, 0, \ldots, 0)$ (avec un 1 en composante $\theta = \phi(s)$), on a alors :

$$R_{\text{AG}}(s, \beta, a, s', \beta') \stackrel{\text{def}}{=} -\sqrt{\|\beta - \beta^*(s)\|_2}.$$

Explicabilité Un comportement explicable est un comportement cohérent avec les attentes de l'observateur.

Θ : Pour traduire cette idée, Miura et Zilberstein (suivant Sreedharan, Kulkarni, Chakraborti et al. [5]) proposent de minimiser la probabilité que le comportement observé soit celui d'un comportement aléatoire, même si plusieurs autres comportements restent probables. Ils introduisent ainsi un type "virtuel" θ_0 qui représente un comportement (une politique) aléatoire en plus des autres types (réels).

R_{AG} : Pour traduire le critère d'explicabilité susmentionné, on va prendre

$$R_{\text{AG}}(s, \beta, a, s'\beta') \stackrel{\text{def}}{=} -\beta(\theta_0).$$

Prédictibilité Un comportement prédictible est un comportement dont la fin de trajectoire est plus facile à prédire pour l'observateur. On propose ici une définition de la prédictibilité inspirée des travaux de Lepers, Thomas et Buffet [3], mais mieux fondée d'un point de vue théorique. On essaye donc de prédire soit l'action, soit l'état de l'agent,

Si nous partons comme Miura et Zilberstein et d'autres de cette définition, nous nous inspirons plutôt des travaux de Lepers, Thomas et Buffet [3], lesquels sont plus adaptés aux problèmes à dynamique stochastique, mais en faisant ici une proposition dont la sémantique est plus claire.

Θ : L'idée de départ est que l'observateur cherche, à chaque instant, à prédire la prochaine action ou le prochain état, d'où deux types de prédictibilité différents. Pour la prédictibilité sur l'action, on pose $\Theta = A$ et $\phi(s, a, s') = a$. Pour la prédictibilité sur l'état, on pose $\Theta = S$ et $\phi(s, a, s') = s'$. Dans les deux cas, pour agir optimalement, l'observateur doit parier sur une des prochaines valeurs cibles les plus probables, et donc choisir une valeur dans l'ensemble

$$\psi_\Theta(\beta_t) \stackrel{\text{def}}{=} \arg\max_\theta \beta_t(\theta).$$

R_{AG} **:** On considère que l'observateur échantillonne sa prédiction de façon uniforme, on peut alors définir :

$$\text{pred}(\theta|\beta_t) \stackrel{\text{def}}{=} \begin{cases} \frac{1}{|\psi_\Theta(\beta_t)|} & \text{si } \theta \in \psi_\Theta(\beta_t), et \\ 0 & \text{sinon.} \end{cases}$$

En définissant

$$R_{\text{AG}}(s, \beta, a, s', \beta') \stackrel{\text{def}}{=} \begin{cases} \text{pred}(a|\beta) - 1 & \text{si } \Theta = A, \\ \text{pred}(s'|\beta) - 1 & \text{si } \Theta = S, \end{cases}$$

la récompense immédiate est l'opposé de la probabilité que le pari (d'un observateur rationnel) échoue : $R_{\text{AG}}(s, \beta, a, s', \beta') = -P(\text{pari perdu})$.

Obscurcissement La problématique inverse peut également être considérée. L'agent essaye alors de cacher des informations telles que son but à l'observateur. Dans cette situation, l'agent a plusieurs buts possibles et essaye de ne pas révéler son "vrai" but à l'observateur. L'obscurcissement avec le modèle PO-OAMDP présente les mêmes difficultés que celles rencontrées par le modèle OAMDP :

- si l'objectif ne porte que sur l'obscurcissement, mais pas sur la réalisation de la tâche, l'agent peut simplement ne rien faire pour dissimuler son but, et
- pour construire la croyance de l'observateur sur les buts, on fait l'hypothèse que l'observateur ne sait pas qu'on essaye de le tromper.

3.3.2 Élargissement des types de problèmes couverts

Les sections précédentes ont montré comment étendre les problèmes déjà modélisés dans les cadres OAMDP et p-OAMDP en considérant l'observabilité partielle de l'observateur. Il faut cependant noter que les PO-OAMDP permettent la formalisation de nouveaux problèmes dans lesquels l'agent ne va pas simplement chercher à exhiber un comportement prédictible, lisible ou explicable, mais pourra chercher à transmettre au mieux de l'information sur l'état du monde partiellement observé par l'observateur.

Scénario 1 : Si on considère un environnement de type bureau (cf. figure 3) avec des portes soit ouvertes, soit fermées à clef, on peut imaginer un agent cherchant à faire comprendre à un observateur extérieur l'état des portes à travers ses actions. Cela peut bien entendu se faire en ouvrant des portes visibles pour l'observateur mais aussi en se montrant dans certaines zones qui ne peuvent être atteintes que par l'ouverture de certaines portes. Ainsi, même si ces portes ne sont jamais vues par l'observateur, la présence de l'agent peut lui permettre d'inférer que certaines portes sont ouvertes. Dans l'exemple représenté en figure 3, en

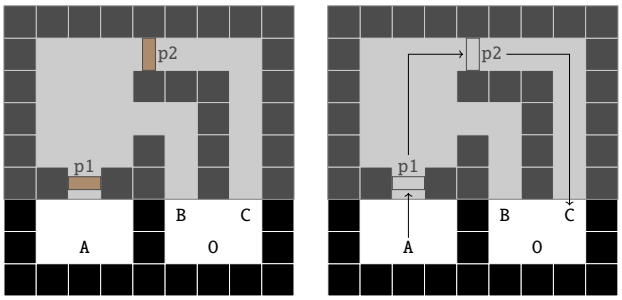

FIGURE 3 : Représentation d'un environnement avec des portes pouvant être verrouillées. Les murs sont représentés pas des cases noires, les portes par des rectangles bruns. La zone grisée correspond à une zone non visible de l'observateur. L'agent débute en `A` et doit se diriger en `O`.

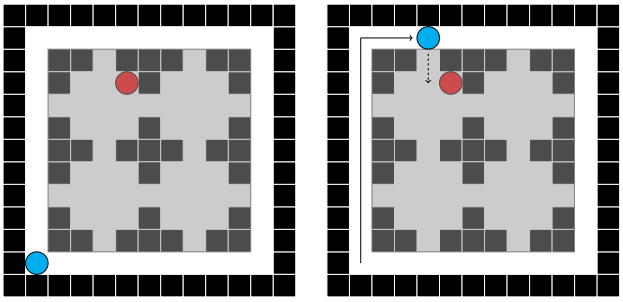

FIGURE 4 : Environnement à grille avec un intrus dont l'observateur cherche à connaître la localisation. Les murs sont représentés par des cases noires. La zone grisée correspond à une zone non visible de l'observateur. L'intrus est représenté par un cercle rouge et l'agent par un cercle bleu.

choisissant un chemin plus long dans la zone cachée et en redevenant visible en `C`, il informe l'observateur que les portes `p1` et `p2` ne sont pas verrouillées. Se rendre visible en `B` permettrait d'atteindre l'objectif et augmenterait un peu la probabilité que la porte `p2` soit fermée. Résoudre ce problème nécessite que l'agent raisonne sur 1. les conséquences de ses actions, 2. sa visibilité en fonction de sa localisation, 3. les inférences que pourra faire l'observateur et 4. les portes dont l'observateur cherche à estimer l'état.

Scénario 2 : Dans une seconde situation, on peut considérer un agent en charge de détecter des intrus dans un environnement inaccessible pour l'observateur (cf. figure 4). En utilisant un modèle que l'observateur a de son comportement, cet agent peut tirer parti des attentes de l'observateur pour agir, se rendre visible à certains endroits et faire comprendre à l'observateur la présence effective d'un intrus et sa localisation. Dans l'exemple illustré par la figure 4, l'observateur estime que l'agent cherche à se rapprocher

de l'intrus. L'agent peut ainsi informer l'observateur de la position de l'intrus en choisissant, parmi les trajectoires possibles l'amenant au plus proche de l'intrus, un chemin souvent visible de l'observateur.

Scénario 3 : Enfin, dans des tâches complexes qui requièrent plusieurs étapes intermédiaires, l'agent peut chercher à transmettre l'état de la tâche en cours en empruntant des chemins plus longs mais 1. qui sont partiellement visibles par l'observateur et 2. qui laissent moins d'ambiguïté sur son objectif intermédiaire.

En cherchant à faciliter l'inférence des objectifs intermédiaires qu'il cherche à atteindre, l'agent peut ainsi transmettre l'état de la tâche à réaliser, ce qui peut s'avérer crucial dans une situation collaborative (qui attendrait en retour une action particulière de l'humain).

Ces différentes situations montrent que le cadre des OAMDP permet d'élargir les problèmes à d'autres types d'échanges d'information que simplement des informations sur le comportement de l'agent lui-même. Le cadre permet de modéliser des problèmes proches des problèmes de recherche active d'information tels que formalisés par les ρ-POMDP [6]. Dans le cadre ρ-POMDP, un agent est dans un environnement partiellement observé et doit agir au mieux pour acquérir des observations pertinentes et maximiser une mesure d'information sur des variables cibles (par exemple sa localisation). La principale différence avec le cadre PO-OAMDP est que, dans ce dernier, c'est l'information acquise par un tiers (l'observateur) que l'agent cherche à contrôler, pas la sienne propre (qui est complète). Cela nécessite en particulier un modèle de l'observateur dont l'agent va tirer parti pour chercher à contrôler indirectement les croyances de l'observateur.

4 Résolution

Nous allons maintenant discuter des approches possibles pour la résolution de PO-OAMDP, c'est-à-dire de politiques maximisant (au moins à $\epsilon > 0$ près) la somme des récompenses atténuées. On fait l'hypothèse dans cette section qu'on dispose d'un couple état-croyance initial, les algorithmes discutés en faisant tous usage.

Dans la suite, on considère d'abord les solutions prenant la forme naturelle de politiques dépendant de l'historique, avant de voir que l'on peut aussi raisonner avec des croyances à la place de ces historiques.

4.1 Recherche d'une politique historique-dépendante

4.1.1 Données pertinentes pour la politique

Au premier abord, à un instant donné, le choix d'action de l'agent dépend au plus des données spécifiques à sa trajectoire actuelle, c'est-à-dire l'historique des états, actions et observations.

On a toutefois besoin pour prendre des décisions que d'informations nécessaires à la prédiction des récompenses. Or, par définition, la récompense reçue à un instant t dépend de la transition $s_t, b_t, a_t, s_{t+1}, b_{t+1}$ (la croyance sur la variable cible se déduisant de la croyance sur l'état comme on l'a déjà vu). Comme, dans ce tuple, 1. l'état s_t évolue de manière markovienne (indépendamment de l'historique d'états et d'actions antérieurs), et 2. la croyance sur l'état b_t ne dépend que des observations passées, alors le choix d'action ne dépend que de l'état courant s_t et de l'historique d'observations $o_1, \dots, o_{t-1}$. On va donc pouvoir ne considérer que les politiques de la forme $\pi_{\text{AG}} : \mathcal{S} \times (\Omega)^* \to \mathcal{A}$ (en notant que, dans cet espace "d'états", parmi les politiques optimales, certaines sont déterministes).

4.1.2 Résolution

Dans le cas $\gamma < 1$, étant donné $\epsilon > 0$, on peut trouver une solution ϵ-optimale en se ramenant à un problème à horizon fini et en employant un algorithme tel que la programmation dynamique, AO* [7] ou MCTS [8] (comme l'ont fait Miura et Zilberstein [2] pour les OAMDP). L'opérateur d'optimalité de Bellman va alors s'écrire, pour $t < H - 1$,

$$V_t^*(s, o_{1:t}) = \max_a \sum_{s', o_{t+1}} T(s'|a, s) \cdot O(o_{t+1}|s', a) \cdot [R'(s, b_{[o_{1:t}]}, a, s', b_{[o_{1:t+1}]}) + \gamma V_{t+1}^*(s', b_{[o_{1:t+1}]})],$$

et

$$V_H^*(s, o_{1:H}) = 0.$$

Dans le cas $\gamma = 1$, sauf cas particulier, on ne peut pas se ramener à un horizon temporel fini. Les politiques historique-dépendantes ne paraissent donc pas adaptées.

4.2 Recherche d'une politique croyance-dépendante

Nous allons voir ici que la résolution d'un PO-OAMDP est équivalente à celle d'un MDP dans un espace continu particulier. Cela va permettre d'envisager l'emploi de politiques croyances-dépendantes.

4.2.1 Belief-MDP équivalent

Statistique suffisante Nous avons vu précédemment comment calculer l'état de croyance de l'observateur sur l'état, au vu de la séquence d'observations qu'il a reçues, par filtrage bayésien. En fait, l'état d'information (s, b) (l'état courant couplé avec la croyance de l'observateur sur l'état) constitue une statistique suffisante pour la planification puisqu'elle

1. est markovienne (on peut prédire son évolution sans avoir recours à des informations antérieures) puisque l'état s est markovien par définition, et la croyance b l'est aussi d'après nos calculs (elle peut être mise à jour en ne connaissant que la dernière observation reçue) ; et

2. permet d'estimer la récompense à chaque pas de temps, par définition de la fonction de récompense dans un PO-OAMDP.

Formalisation du belief-MDP On obtient donc un MDP valide $\langle \mathcal{I}, \mathcal{A}, T', R', \gamma, \mathcal{I}_f \rangle$, où :

- $\mathcal{I} \stackrel{\text{def}}{=} \mathcal{S} \times B$ est l'ensemble des états ;
- $\mathcal{A}$ est l'ensemble des actions, identique à l'ensemble des actions du PO-OAMDP ;
- $T' : \mathcal{I} \times \mathcal{A} \times \mathcal{I} \to [0; 1]$ est une nouvelle fonction de transition (voir plus bas) ;
- $R' : \mathcal{I} \times \mathcal{A} \times \mathcal{I} \to \mathbb{R}$ est une nouvelle fonction de récompense (voir plus bas) ;
- $\gamma \in [0, 1]$ est le facteur d'actualisation ; et
- $\mathcal{I}_f \subset \mathcal{I}$ est l'ensemble des éléments $\iota \equiv (s, b)$ de $\mathcal{I}$ tels que $s \in \mathcal{S}_f$; on pourra vérifier ultérieurement que, pour tout $\iota, a \in \mathcal{I} \times \mathcal{A}$, $T(\iota, a, \iota) = 1$ et $R'(\iota, a, \iota) = 0$.

Fonction de transition T' : La fonction de transition du belief-MDP est définie par :

$$\begin{aligned} T'(\iota'|a, \iota) &\stackrel{\text{def}}{=} T(s', b'|a, s, b) \\ &= \sum_o \mathbb{1}_{b'=B(b,o)} O(o|s', a) P(s'|a, s). \end{aligned}$$

Fonction de récompense R' : La fonction de récompense du belief-MDP est définie par :

$$R'(s, b, a, s', b') \stackrel{\text{def}}{=} R_{\text{AG}}(s, \beta(b), a, s', \beta(b')),$$

où $\beta(b)$ dénote la croyance β que l'on peut dériver de b comme vu dans l'équation (1).

4.2.2 Résolution

Ce nouveau MDP est défini sur un espace d'états continu. Sauf cas particuliers, l'ensemble des états accessibles depuis l'état initial est donc infini.

Dans le cas $\gamma < 1$, on pourrait à nouveau se ramener à un problème à horizon fini, avec l'avantage que deux trajectoires différentes peuvent conduire à la même paire (ι, t), ce qui permettrait de réduire la quantité de calculs. L'opérateur d'optimalité de Bellman s'écrit alors (en développant $\iota = (s, b)$)

$$\begin{aligned} V_t^*(s, b) &= \max_a \sum_{s'} \int_{b'} T(s', b'|a, s, b) \cdot \\ &\qquad [R'(s, b, a, s', b') + \gamma V_{t+1}^*(s', b')] db' \\ &= \max_a \sum_{s', o} O(o|s', a) \cdot T(s'|a, s) \cdot \\ &\qquad [R'(s, b, a, s', b^o) + \gamma V_{t+1}^*(s', b^o)], \end{aligned}$$

où b^o est la croyance sur l'état mise à jour après observation de o, et

$$V_H^*(s, b) = 0.$$

On peut aussi se demander si, comme dans le cadre des POMDP résolus via des bMDP, la fonction de valeur optimale a des propriétés de continuité particulières (en l'occurrence de convexité) qui permettraient d'approcher celle-ci. Des résultats préliminaires montrent toutefois qu'il peut y avoir des discontinuités sur les bords de la fonction de valeur d'une politique, autour de points dont l'état de croyance est impossible. On ne pourra donc pas re-transcrire directement les approches POMDP reposant sur des approximateurs généralisants de V^* (tels que HSVI [9], PBVI [10] et SARSOP [11]). Mais des versions spécifiques tenant compte des localisations des discontinuités sont peut-être envisageables.

Dans le cas $\gamma = 1$, un premier problème est de savoir si le problème obtenu est un SSP valide. Il faudrait ainsi vérifier, par exemple, si c'est toujours le cas avec les fonctions de récompenses proposées pour les propriétés de lisibilité, explicabilité et prédictibilité.

À supposer que le SSP obtenu soit valide, on se retrouve sur des problèmes proches des POSSP (ou Goal-POMDP) pour lesquels peu de travaux ont été développés à part, par exemple, ceux de Patek [12] ou, plus récemment, de Horák, Boşanský et Chatterjee [13].

Approche bi-critère Un problème déjà présent dans la résolution des OAMDP est que, si l'on emploie un critère lié à une fonction de récompense "observer-aware", il est possible que la performance liée à la fonction de récompense classique du MDP sous-jacent soit fortement dégradée. Une première approche peut être de combiner linéairement deux tels critères, mais cela soulève la question de la bonne pondération de ceux-ci. Une autre approche, aussi évoquée par Miura et Zilberstein [2], est d'optimiser un critère observer-aware sous la contrainte que le

critère “classique” doit atteindre au minimum une certaine valeur [14]-[18]. Il peut alors être nécessaire que la politique optimale soit stochastique.

5 Conclusion

Nous avons introduit un nouveau formalisme, celui des OAMDP en observabilité partielle (PO-OAMDP), lequel permet de travailler sur des problèmes de lisibilité, d'explicabilité et de prédictibilité en observabilité partielle. La complexité des PO-OAMDP est au moins celle des OAMDP telle qu'étudiée par Miura et Zilberstein [2]. Selon eux, il n'est pas nécessairement bénéfique de se ramener un POMDP pour le résoudre. Différents ensembles de variables cibles et fonctions de récompense ont été proposés pour construire des comportements avec des propriétés différentes. Ce nouveau modèle permet aussi de traiter des problèmes plus proches de la réalité où un agent agit en prenant en compte un observateur qui peut ne pas avoir accès à l'ensemble des informations de l'environnement (porte fermée, champ de vision bloqué). Nous avons également discuté de la résolution des PO-OAMDP.

Comme expliqué dans la partie résolution, nous proposons de transformer les PO-OAMDP en belief-MDP équivalents pour les résoudre avec des approches génériques comme MCTS. Une première perspective est d'étudier les performances de ces algorithmes pour résoudre les PO-OAMDP. Une seconde direction de travail est d'étudier les propriétés théoriques du modèle PO-OAMDP pour proposer des algorithmiques tirant parti des propriétés de ce cadre. Une étape sera de proposer des problèmes pertinents pour tester notre modèle et évaluer les algorithmes proposés.

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)*, 2019. adresse : https://ojs.aaai.org/index.php/ICAPS/article/view/3463.

[2] 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*, t. 161, juill. 2021, p. 610-620. adresse : https://proceedings.mlr.press/v161/miura21a.html.

[3] S. Lepers, V. Thomas et O. Buffet, “Comment rendre des comportements plus prédictibles”, in *JIAF-JFPDA - Journées d'Intelligence Artificielle Fondamentale*, juill. 2023. adresse : https://hal.science/hal-04212452.

[4] 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.

[5] S. Sreedharan, A. Kulkarni, T. Chakraborti, D. E. Smith et S. Kambhampati, *A Bayesian Account of Measures of Interpretability in Human-AI Interaction*, 2020. arXiv : 2011.10920 [cs.AI].

[6] M. Araya-López, O. Buffet, V. Thomas et F. Charpillet, “A POMDP Extension with Belief-dependent Rewards”, in *Advances in Neural Information Processing Systems 23*, Vancouver, Canada, 2010.

[7] N. Nilsson, *Principles of Artificial Intelligence*. Morgan Kaufmann Publishers, 1980.

[8] L. Kocsis et C. Szepesvari, “Bandit based Monte-Carlo Planning”, in *Proceedings of the Sixteenth European Conference on Machine Learning*, 2006.

[9] 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.

[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] 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.

[12] S. Patek, “On partially observed stochastic shortest path problems”, in *Proceedings of the 40th IEEE Conference on Decision and Control*, t. 5, 2001, p. 5050-5055. doi : 10.1109/CDC.2001.981011.

[13] K. Horák, B. Bošanský et K. Chatterjee, “Goal-HSVI : Heuristic Search Value Iteration for Goal-POMDPs”, in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 2018, p. 4764-4770.

[14] E. Altman, *Constrained Markov Decision Processes*. Chapman et Hall/CRC, 1999.

[15] D. Kim, J. Lee, K.-E. Kim et P. Poupart, “Point-based value iteration for constrained POMDPs”, in *IJCAI*, t. 11, 2011, p. 1968-1974.

[16] F. Dufour et T. Prieto-Rumeau, "Stochastic approximations of constrained discounted Markov decision processes", *Journal of Mathematical Analysis and Applications*, t. 413, n° 2, p. 856-879, 2014. doi : 10.1016/j.jmaa.2013.12.016.

[17] F. Trevizan, S. Thiébaux, P. Santana et B. Williams, "Heuristic Search in Dual Space for Constrained Stochastic Shortest Path Problems", *Proceedings of the International Conference on Automated Planning and Scheduling*, t. 26, n° 1, mars 2016. doi : 10.1609/icaps.v26i1.13768.

[18] J. Lee, G.-H. Kim, P. Poupart et K.-E. Kim, "Monte-Carlo tree search for constrained POMDPs", *Advances in Neural Information Processing Systems*, t. 31, 2018.