Titre: Title:	Collaboration humain-robot: modèles mentaux et apprentissage de la rationalité pour une assistance adaptative
Auteur: Author:	Cyrille Jamabel Tabe
Date:	2025
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Tabe, C. J. (2025). Collaboration humain-robot: modèles mentaux et apprentissage de la rationalité pour une assistance adaptative [Mémoire de maîtrise, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/65823/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/65823/
Directeurs de recherche: Advisors:	Jérôme Le Ny
Programme: Program:	Génie électrique

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Collaboration humain-robot: modèles mentaux et apprentissage de la rationalité pour une assistance adaptative

CYRILLE JAMABEL TABE
Département de génie électrique

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
Génie électrique

Avril 2025

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Ce mémoire intitulé :

Collaboration humain-robot: modèles mentaux et apprentissage de la rationalité pour une assistance adaptative

présenté par **Cyrille Jamabel TABE**
en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
a été dûment accepté par le jury d'examen constitué de :

Roland MALHAMÉ, président
Jérôme LE NY, membre et directeur de recherche
Sarath Chandar ANBIL PARTHIPAN, membre

DÉDICACE

If I have seen further
it is by standing on the
shoulders of Giants.

Isaac Newton

REMERCIEMENTS

Les mots me manquent pour exprimer ma gratitude envers mon directeur de recherche, Jérome Le Ny, dont la rigueur et la présence constante ont guidé mes recherches. Son engagement et sa conviction en mes capacités ont rendu cette maîtrise possible, et sans son soutien financier, je n'aurais pas eu la chance de présenter un tel mémoire.

Je tiens également à remercier sincèrement les professeurs Roland Malhamé et Sarath Chandar d'avoir accepté de faire partie de mon jury et d'avoir partagé leur expertise précieuse dans l'évaluation de mon travail.

Ma reconnaissance s'étend aussi à mes collègues du laboratoire de robotique de Polytechnique Montréal, dont le soutien, les encouragements et nos moments de détente resteront gravés parmi mes souvenirs les plus chaleureux de cette maîtrise.

Enfin, je m'en voudrais de ne pas mentionner ma famille et toutes les personnes qui ont croisé mon chemin depuis le début de cette aventure. Leur présence et leur impact sur ma vie m'ont profondément marqué et ont contribué, à leur manière, à l'aboutissement de ce projet.

RÉSUMÉ

L'interaction humain-machine représente un enjeu central dans le développement de systèmes intelligents destinés à assister l'humain dans des tâches complexes. Ce mémoire explore un cadre de collaboration humain-robot dans lequel le robot aide un humain poursuivant un objectif parmi plusieurs possibles, en adaptant sa stratégie de décision en fonction des actions observées. L'objectif principal est d'optimiser l'assistance du robot en réduisant l'incertitude sur l'intention humaine, sans communication explicite entre les agents. Pour aborder cette problématique, les Processus Décisionnels Markoviens Partiellement Observables (POMDP) sont utilisés. L'hypothèse principale repose sur le fait que l'humain adopte un comportement à rationalité limitée. Ainsi, le robot estime son intention à l'aide d'un modèle mental fondé sur une politique de Boltzmann, ajusté dynamiquement grâce à une méthode d'apprentissage de l'indice de rationalité en fonction des états et actions observées. Deux niveaux de planification sont étudiés : (a) Modèle mental de premier ordre : le robot suppose que l'humain agit indépendamment de ses propres décisions et optimise sa politique en conséquence ; (b) Modèle mental de second ordre : le robot intègre l'idée que l'humain peut adapter ses actions en réponse aux siennes, introduisant une complexité supplémentaire liée à la récursion dans les modèles de décision. L'approche adoptée est validée expérimentalement dans un environnement simulé où le robot doit aider un humain à atteindre un objectif en retirant des obstacles sur son chemin. Les résultats montrent que l'intégration du modèle mental améliore la capacité du robot à anticiper les besoins de l'humain et à agir de manière proactive. Cependant, des limitations persistent, telles que l'estimation précise de la rationalité humaine et la complexité des modèles. Ce travail contribue à l'avancement des méthodes de prise de décision en collaboration humain-robot, ouvrant la voie à des applications en logistique, en santé et en robotique de service.

ABSTRACT

Human-machine interaction represents a central issue in developing intelligent systems designed to assist humans in complex tasks. This thesis explores a framework for human-robot collaboration in which the robot assists a human pursuing one of several possible goals, adapting its decision strategy according to observed actions. The main objective is to optimize robotic assistance by reducing uncertainty about human intention, without explicit communication between agents. To address this problem, Partially Observable Markov Decision Processes (POMDP) are used. The main assumption is that human behavior is boundedly rational. Thus, the robot estimates its intention using a mental model based on a Boltzmann policy, adjusted dynamically through a method of learning the rationality index as a function of observed states and actions. Two levels of planning are studied: (a) First-order mental model: the robot assumes that humans act independently of their own decisions, and optimizes its policy accordingly; (b) Second-order mental model: the robot incorporates the idea that humans may adapt their actions in response to its own, introducing additional complexity linked to recursion in decision models. The approach adopted is validated experimentally in a simulated environment where the robot must help a human to achieve a goal by removing obstacles in its path. The results show that integrating the mental model improves the robot's ability to anticipate the human's needs and act proactively. However, limitations remain, such as accurate estimation of human rationality and model complexity. This work contributes to advancing human-robot collaborative decision-making methods, paving the way for applications in logistics, healthcare and service robotics.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

MDP	Markov Decision Process
POMDP	Partially Observable Markov Decision Process
MOMDP	Mixed Observability Markov Decision Process
HGMDP	Hidden Goal Markov Decision Process
HAMDP	Helper Action Markov Decision Process
I-POMDP	Interactive Partial Observability Markov Decision Process
Dec-POMDP	Decentralized Partial Observability Markov Decision Process
CIRL	Cooperative Inverse Reinforcement Learning

LISTE DES ANNEXES

CHAPITRE 1 INTRODUCTION

Dans ce chapitre, nous introduisons la notion de collaboration lorsqu'un robot est chargé d'assister un humain dans l'exécution de ses tâches, en examinant les problématiques potentielles ainsi que la pertinence d'une telle coopération. Nous définissons ensuite un scénario d'application destiné à éclairer notre sujet de recherche. Enfin, nous exposons les objectifs qui ont orienté les travaux de ce projet de maîtrise.

1.1 Interaction humain machine

1.1.1 Mise en contexte

La collaboration humain-machine [1] est un sujet d'intérêt majeur dans les récentes avancées scientifiques, principalement en raison de l'évolution des capacités des machines. En effet, ces dernières sont dotées de compétences de plus en plus similaires à celles de l'humain, tant en termes de décision et de raisonnement autonomes que de types d'interaction. Il est essentiel de catégoriser les différents types d'assistance au sein d'une telle équipe. Selon [1], l'évolution des systèmes humain-machine se divise en trois catégories distinctes : les systèmes de support à la décision, les systèmes intelligents hybrides-augmentés et les systèmes humains-cyber-physiques. Plus précisément, les systèmes de support à la décision exploitent la puissance de calcul des ordinateurs pour assister les utilisateurs ou opérateurs dans leurs prises de décision. Les systèmes intelligents hybrides-augmentés visent à répartir les tâches entre l'humain et la machine, celle-ci tentant alors de les accomplir de manière autonome. En revanche, pour les systèmes humains-cyber-physiques, la machine exécute une partie significative du travail précédemment réalisé par l'humain. Ce dernier système repose généralement sur les connaissances, l'expérience et les données expérimentales de l'opérateur. Il faut donc comprendre que l'accomplissement d'une tâche peut être réalisé par la combinaison de l'effort entre une machine et un humain moyennée par les capacités cognitives de chaque acteur. Dès lors, la question à se poser est de savoir comment cette collaboration peut être orchestrée afin de surpasser la performance individuelle de l'humain.

À cet effet, la structure de coopération d'égal à égal (*peer-to-peer*) permet aux parties prenantes d'exécuter des rôles complémentaires dans l'accomplissement d'une même tâche [2–4]. Par exemple, dans des situations de catastrophes naturelles, un pompier pourrait se charger

1. Dans ce mémoire, nous désignons par "machine" tout système doté de capacité de prise de décision. Cela inclut les logiciels de supervision, les robots (par exemple, Roomba), l'intelligence artificielle (IA) (par exemple, ChatGPT) et les agents intelligents (par example, Alexa).

de secourir les personnes coincées sous les débris, tandis qu'un robot explorerait le terrain à la recherche d'autres survivants, tout en évacuant les personnes secourues. Quelques aspects majeurs de la coopération sont alors mis en lumière : comment déterminer le but final d'une tâche entamée par un humain, comment collaborer efficacement avec l'humain afin de refléter des performances similaires ou supérieures à celles d'une équipe humain-humain [5].

Dans ce mémoire, nous nous intéressons à la modélisation de la structure de prise de décision du robot, afin d'optimiser la performance humaine.

1.1.2 Spécificité du modèle humain

La conception d'un agent intelligent repose sur une compréhension approfondie de l'environnement avec lequel il interagit. Les objets statiques ou dynamiques présents dans cet environnement peuvent être observés grâce à divers capteurs, tels que les caméras ou les lidars. Cependant, bien que l'humain puisse également être observé, de nombreux facteurs restent difficiles, voire impossibles, à quantifier ou à détecter. Ces facteurs incluent l'influence de la machine (performance face à d'autres agents, incertitude des observations), de l'environnement (multiplicité des stimuli externes) et de l'humain lui-même (niveau d'expérience ou de compétence, confiance envers la machine, simplicité des interactions) [6, 7]. Bien que certains capteurs, comme ceux utilisés en vision assistée par ordinateur, permettent de détecter des traits subtils et d'améliorer la compréhension du comportement humain, leur intégration peut complexifier le modèle et augmenter le risque d'erreurs critiques si elle est mal implémentée dans les décisions du robot.

La modélisation du comportement humain est essentielle pour permettre au robot d'anticiper des décisions influencées par des facteurs, tels que la perception des obstacles ou la charge mentale. Par exemple, un humain peut agir de manière sous optimale en raison d'éléments invisibles pour le robot ou d'un stress important. Ce mémoire intègre ces aspects en représentant la rationalité humaine à l'aide d'un modèle dédié. Cependant, en raison de la complexité inhérente à cette modélisation, des hypothèses simplificatrices sont nécessaires pour garantir un cadre cohérent et sans ambiguïté.

1.1.3 Spécificité du modèle du robot

L'idéal dans une coopération humain-machine est de favoriser une interaction aussi fluide et naturelle que possible, idéalement comparable à celle entre deux humains, voire même plus efficace ou adapté aux spécificités de la machine. L'intégration d'un robot repose sur ses capacités computationnelles et la personnalisation de son processus décisionnel. Pour cela,

la logique décisionnelle du robot doit être adaptée à ses capacités. Par exemple, un système conçu pour analyser le son n'a pas besoin d'une logique dédiée à l'interprétation d'images. La capacité de raisonnement d'un robot dépend fortement de son espace d'observation et de ses ressources limitées, ce qui peut le rendre sous-optimal face à des environnements très vastes. De plus, certaines actions du robot peuvent, au lieu d'aider, nuire à la progression des tâches de l'humain. Le robot doit donc être capable de ne pas agir si cela ne contribue pas à améliorer ses performances ou celles de l'équipe.

Une observation essentielle concernant la dépendance du modèle du robot à celui de l'humain est l'impact de l'inexactitude du modèle humain sur les décisions prises par le robot. Plusieurs solutions peuvent être envisagées, telles que l'apprentissage continu du comportement de l'humain à travers des observations, la modélisation exacte du comportement de l'humain ou même l'incorporation d'incertitude sur le modèle de l'humain.

Ce mémoire tiendra compte de ces aspects en limitant l'espace des états de l'humain et de l'environnement.

1.2 Problématique

Dans cette section, nous présentons des applications intégrant des stratégies de coopération entre humains et robots afin de motiver le sujet, pour ensuite mettre en évidence le scénario spécifique traité dans ce mémoire. Une définition plus formelle de ce scénario est donnée dans le chapitre 3.

1.2.1 Motivation et application

Dans de nombreuses situations, les humains doivent accomplir des tâches exigeantes dans des environnements partiellement inconnus ou dangereux. Par exemple, un ouvrier sur un site industriel peut être concentré sur le maniement précis d'une machine lourde, rendant toute interaction avec un robot difficile, voire risquée. De même, lors d'opérations de secours, un pompier engagé dans la recherche de survivants sous des débris ou dans un bâtiment en feu ne peut détourner son attention pour coordonner ses actions avec un robot. Ces scénarios illustrent des contextes où l'humain est absorbé par une tâche critique et ne dispose ni du temps ni des moyens pour communiquer directement avec d'autres acteurs, qu'ils soient humains ou robotiques. Bien que certains travaux se soient focalisés sur l'intégration d'informations obtenues par divers moyens de communication dans la prise de décision du robot, démontrant ainsi leur efficacité dans des équipes humain-robot ([8, 9]), nous avons choisi de ne pas orienter notre problématique dans cette direction. Cette décision vise à simplifier le

cadre de développement de ce mémoire. Dans ces conditions, un robot doit être capable de comprendre implicitement les actions de l'humain à partir de ses observations, afin d'agir de manière autonome et adaptée.

Pour répondre à ces défis, un robot capable d'observer l'environnement et d'identifier des solutions adaptées peut jouer un rôle clé. De même, dans un contexte médical, un robot pourrait organiser des trajets optimaux dans une salle d'opération encombrée, aidant les chirurgiens à accéder rapidement à des instruments critiques sans gêner leur travail. Ces exemples montrent que la capacité du robot à anticiper les besoins de l'humain, sans nécessiter une interaction explicite, est essentielle pour garantir une coopération fluide et efficace.

Cette problématique soulève une question fondamentale : comment concevoir un système robotique capable d'aider un humain en interprétant ses comportements et son environnement, même en l'absence de communication directe ? La réponse à cette question trouve des applications dans des domaines variés, tels que l'industrie, où des robots peuvent faciliter des processus complexes, ou encore dans des opérations de secours, où la sécurité et la réactivité sont primordiales. Ces applications mettent en évidence l'importance d'une collaboration implicite et adaptative entre humain et robot, dans laquelle le robot agit comme un assistant intelligent capable de s'adapter aux besoins humains en temps réel.

1.2.2 Description du scenario

Considérons un scénario où un humain doit accomplir une tâche dans un environnement complexe. Il peut déterminer le chemin optimal vers chacune des tâches potentielles, mais son objectif est d'atteindre une tâche spécifique. Toutefois, des obstacles l'empêchent de progresser, et il ne peut pas les retirer lui-même, car il est occupé à d'autres activités essentielles, comme la recherche d'informations ou la récupération d'objets nécessaires à l'exécution de sa tâche.

De son côté, le robot observe l'environnement, les obstacles potentiels, ainsi que les objectifs possibles de l'humain. Étant donné que l'humain ne peut pas communiquer avec le robot et ne tient pas compte de sa présence, le robot doit être capable de raisonner sur les comportements probables de l'humain, tout en tenant compte de l'incertitude sur ses actions qui sont par hypothèse encapsulées dans son indice de rationalité. La complexité du problème réside dans le fait que le robot ne connaît pas l'objectif précis de l'humain. Il doit donc, à partir de ses observations, déduire l'objectif le plus probable et les actions pour maximiser l'atteinte de cet objectif. Par exemple, dans un entrepôt logistique, un robot pourrait planifier et dégager un chemin sûr pour qu'un travailleur puisse transporter des matériaux lourds vers une zone spécifique sans avoir à se préoccuper des obstacles, comme le montre la figure 1.1. Dans

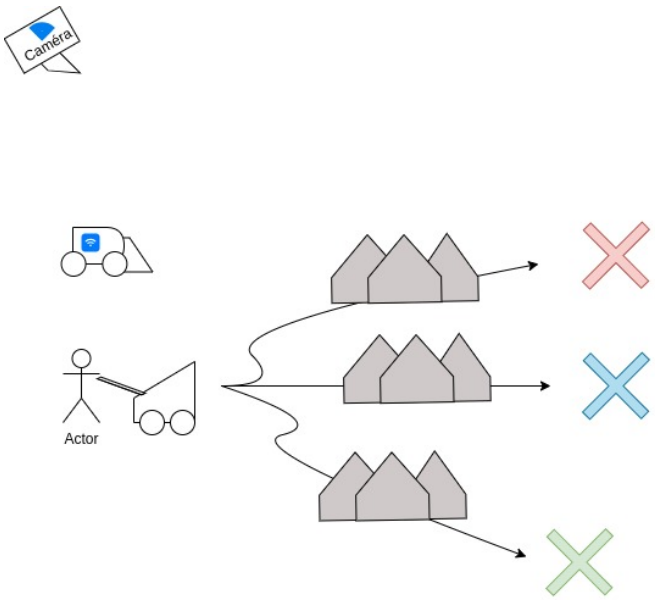


FIGURE 1.1 Exemple d'un entrepôt où un humain transporte des matériaux fragiles, nécessitant toute son attention. Pendant ce temps, un robot mobile doté d'une compréhension globale de l'environnement (grâce à une caméra) est capable de dégager les obstacles présents sur différents chemins menant à plusieurs points de dépôt potentiels.

ce contexte, si l'humain est bloqué, dégager un ou plusieurs obstacles pourrait clarifier son objectif, lui permettant ainsi de se déplacer de manière optimale vers celui-ci. Cependant, comme chaque action d'enlèvement d'obstacles entraîne un coût pour le robot, ces actions doivent être effectuées de manière stratégique.

Ce scénario fera l'objet d'une présentation formelle dans la section 3.1. En résumé, le problème consiste à déterminer la stratégie de décision optimale pour le robot, en tenant compte des comportements possibles de l'humain dans une situation où son objectif est inconnu.

1.3 Plan du mémoire

Ce mémoire s'organise comme suit : le chapitre 2 présente une revue de la littérature, situant notre travail dans son contexte de recherche. Il examine les modèles existants et leur utilisation pour permettre des prises de décision coopératives par les robots. La section 3.1 propose une formulation détaillée du problème, en décrivant le modèle de prise de décision du robot dans des situations où l'humain poursuit un objectif connu ou pas. Ce chapitre introduit également une méthodologie d'apprentissage permettant au robot de corriger les écarts entre son modèle prédéfini du comportement humain et le comportement effectivement observé.

Les résultats obtenus en appliquant notre modèle dans divers scénarios sont présentés et

analysés dans le chapitre 4. Enfin, le chapitre 5 propose une approche pour résoudre un problème d'impasse identifié lors de l'analyse des résultats expérimentaux, ouvrant ainsi la voie à des travaux futurs plus approfondis.

CHAPITRE 2 REVUE DE LITTÉRATURE

Dans un scénario de collaboration humain-robot, les stratégies de prise de décision des deux acteurs contribuent à l'accomplissement de la tâche. Puisque les stratégies du robot peuvent être modifiées, il est essentiel qu'il gère alors l'incertitude liée aux comportements de l'humain, par exemple, l'humain peut ne pas partager les mêmes objectifs initiaux que le robot. Selon l'étude de [10], une collaboration humain-robot réussie dépend de la capacité du robot à connaître ou à déduire les objectifs potentiels de l'humain en fonction du contexte. La stratégie de décision du robot doit donc être structurée selon le type de tâche collaborative envisagée.

Dans ce chapitre, les méthodes de représentation de l'objectif d'un humain sont d'abord exposées. Ensuite, la littérature sur les types de collaboration humain-robot est discutée, avec un accent sur les niveaux d'ordres mentaux intégrables au robot. La contribution de ce mémoire de maîtrise est présentée en conclusion.

2.1 Facteurs d'incertitudes du comportement humains

Le comportement humain perçu par le robot est soumis à des incertitudes qui influencent fortement sa stratégie décisionnelle. Ces incertitudes concernent notamment l'identification de l'intention humaine, les actions possibles de l'humain, sa charge mentale [11], son niveau de fatigue pour lequel [12] propose un modèle d'estimation, sa confiance envers le robot, ses préférences et l'environnement dans lequel il évolue. Parmi celles-ci, les incertitudes liées aux intentions et aux actions humaines sont particulièrement importantes, car elles impactent directement les décisions du robot.

2.1.1 Incertitude de l'intention humaine

Définitions

La modélisation de l'intention humaine revient à définir et à quantifier les multiples facteurs influençant celle-ci. En effet, avant de prendre une décision, un humain se base sur son objectif, son état mental, son expérience, sa culture, sa confiance envers l'environnement qu'il observe et les agents intelligents qui s'y trouvent. Nous nous intéressons plus précisément à la modélisation de l'objectif.

Modèles de l'objectif

Selon [13], il existe trois candidats pour la représentation des connaissances antérieures concernant les objectifs humains. Ces modèles peuvent être assimilés à la manière dont un être humain s'attend à observer l'objectif d'une tierce personne. Ainsi, le premier modèle suppose qu'une personne exécute un ensemble d'actions au fil du temps dans le but de compléter un seul objectif fondamental. Par exemple, un robot transportant une bouteille en plastique vide vers le bac de recyclage effectue une série d'actions optimales pour atteindre le point de dépôt. Indirectement, selon ce modèle, toute action déviant du cours d'actions optimal s'explique par un choix involontaire de l'agent.

Le deuxième modèle s'inspire du concept de compréhension de l'action présenté par [14]. En effet, ce modèle décomposable intègre le choix d'un agent de poursuivre un objectif simple, similaire au premier modèle, avec une probabilité k , ou un objectif complexe, incluant la contrainte que l'agent doit passer par un point spécifique dans son chemin optimal, avec une probabilité 1 - k. Dans ce cas, une déviation du chemin optimal vers le but final peut s'expliquer par un objectif intermédiaire à accomplir.

Enfin, le dernier modèle suggère que l'objectif d'un agent peut changer au fil du temps pour des raisons inconnues de l'observateur. Pour un nombre fini de buts, un réseau dynamique bayésien prend forme lorsque l'on considère une connaissance préalable de l'objectif actuel, associée à une distribution conditionnelle de changement de but entre deux instants de temps. Pour [13], les deux derniers modèles sont ceux qui se rapprochent le plus d'une possible structure d'objectif humain.

Le modèle de représentation à but unique est utilisé ici pour définir l'objectif de l'agent humain. Comparé aux autres modèles, celui-ci est simple à interpréter et direct à implémenter, ce qui permet de concentrer l'analyse sur la planification des actions du robot dans un environnement collaboratif. Ce modèle se distingue par sa capacité à fournir une structure claire et précise pour la détermination des objectifs, facilitant ainsi la mise en place de stratégies de collaboration efficaces entre l'humain et le robot.

Quantification de l'objectif humaine

Pour aborder les incertitudes dans l'interprétation des objectifs humains, [15] distingue deux approches : reconnaissance de plan et reconnaissance d'objectif. La reconnaissance de plan identifie à la fois l'objectif et les actions nécessaires pour l'atteindre, tandis que la reconnaissance d'objectif se limite à déterminer la cible finale en se basant sur des observations. Bien que plus générale, la reconnaissance de plan, notamment via les modèles HTN (*Hierarchical*

Task Networks), nécessite des modèles complexes, coûteux en calcul et parfois difficiles à interpréter [16].

D'autre part, [13] propose une planification inverse, déduisant les objectifs les plus probables à partir des comportements observés en supposant des actions rationnelles, sans modéliser explicitement l'incertitude de perception de l'agent. Cela peut limiter la précision de la planification inverse dans les cas où le comportement de l'agent n'est pas optimal, ambigu ou exploratoire, car il peut ne pas refléter une stratégie directe orientée vers un objectif. Afin de combler cette limite, les processus de décision markoviens partiellement observables (*Partially Observable Markov Decision Process*, POMDP) [17] sont utilisés pour modéliser la reconnaissance d'objectif en tant que processus d'actualisation de croyance, où chaque action et observation affine progressivement la probabilité des objectifs potentiels. Ce modèle permet de tenir compte d'un comportement potentiellement moins rationnel en intégrant des transitions probabilistes et un environnement incertain.

2.1.2 Modèle d'action de l'humain

Les approches décrites dans la sous-section précédente mettent en évidence un élément commun : la modélisation des actions humaines. En effet, la déduction d'un objectif est étroitement liée aux actions observées, car une séquence d'actions, ou plan reflète la préférence d'un humain pour atteindre un but spécifique. Pour cela, la téléologie des actions, qui suppose que chaque action vise un but logique en fonction des contraintes environnementales, est un modèle pertinent, notamment dans la recherche infantile [18,19]. Ce principe de téléologie inspire des modèles computationnels, notamment la planification inverse bayésienne, qui interprète les actions observées pour en déduire l'objectif humain [14]. Un modèle plus complexe est présenté par [20, 21], indiquant que le principe de rationalité d'une action dépend à la fois des croyances, déterminées par le niveau d'accès perceptuel de l'agent à l'environnement et par ses connaissances générales du monde, ainsi que d'un objectif fixé par les préférences de l'humain. En raison de sa simplicité computationnelle, le modèle de la téléologie des actions est privilégié pour définir la logique employée par le robot afin de représenter le plan suivi par l'humain. Ce choix s'aligne naturellement avec l'utilisation des processus de décision markoviens (MDP), qui offrent une structure mathématique permettant d'incorporer les informations essentielles à la prise de décision de l'humain.

En général, un agent planifie ses actions pour atteindre un but en suivant une logique parfois non entièrement perceptible pour l'observateur, le robot dans notre cas, ce qui pose la question de la rationalité perçue. Un des objectifs de cette recherche est ainsi de déterminer, à partir d'observations, l'indice de rationalité d'un agent pour mieux adapter la stratégie décisionnelle

du robot.

2.2 Collaboration humain-machine

Dans cette sous-section, nous nous intéressons aux diverses solutions liées à l'introduction de machines dotées d'une intelligence ou d'une autonomie distincte face aux autres agents, notamment les humains, dans un environnement de travail coopératif. Nous citons quelques travaux qui présentent les principales approches de la collaboration au sein d'une équipe humain-machine, en mettant en lumière les hypothèses formulées ainsi que leurs limitations.

2.2.1 Mise en contexte

La collaboration humain-machine met en exergue la problématique de la planification des actions de différents acteurs, notamment l'humain et la machine, dans des environnements statiques ou dynamiques. Ces environnements incluent de nombreux facteurs influençant la prise de décision des acteurs. En conséquence, de nombreux travaux se concentrent sur la modélisation de ces interactions au sein d'une architecture standard et générale, dans le but d'améliorer le processus de prise de décision sans nécessiter de modifications majeures face aux différents types de scénarios de coopération. Plusieurs approches existent, telles que la planification classique, qui correspond à une méthode de planification séquentielle fondée sur des hypothèses spécifiques (modèle du monde déterministe et entièrement observable, état initial unique, etc.) [22]. Cette approche a prouvé son efficacité dans divers domaines, notamment la logistique [23] et la robotique [24]. Toutefois, malgré les recherches approfondies basées sur cette méthode, ce mémoire adopte le processus décisionnel markovien (*Markov Decision Process*, MDP). En effet, la planification classique présente des limites importantes, notamment en raison de ses hypothèses simplificatrices qui ne reflètent pas toujours la réalité et de sa difficulté à modéliser l'incertitude ainsi que les environnements dynamiques. Le MDP, grâce à sa modularité et sa simplicité, permet de pallier ces insuffisances. Cette structure sera adaptée afin d'être compatible avec les problèmes définis. Le processus de décision markovien constitue la structure couramment utilisée dans la plupart des méthodes que nous aborderons dans la suite de cette sous-section. Il est alors essentiel d'en faire une brève introduction.

Un MDP est un formalisme mathématique utilisé pour la prise de décision d'un seul agent et se définit comme un tuple $\langle X, A, T, R, \gamma \rangle$ où X et A sont respectivement les ensembles d'états et d'actions, et T est le modèle de transition d'état ou la dynamique de transition d'état telle que $T(x, a, x') := P[x_{t+1} = x' | x_t = x, a_t = a]$ est la probabilité que l'état suivant soit x' étant donné que l'état actuel est x et que l'action a est exécutée. R est une fonction de récompense

telle que $R(x, a)$ est une récompense à valeur réelle obtenue en exécutant l'action a dans l'état x, γ dans l'intervalle $[0, 1]$ est un facteur d'actualisation (*Discount*) qui détermine la valeur relative de la récompense immédiate par rapport aux futures récompenses [25].

Dans les sous-sections à venir, nous discuterons des différentes approches adoptées dans la littérature pour atteindre l'objectif d'une collaboration entre un humain et une machine. Ces approches seront organisées selon trois grands concepts d'adaptation : l'adaptation du robot à l'humain, l'adaptation de l'humain au robot et, enfin, l'adaptation mutuelle [26].

2.2.2 Adaptation du robot à l'humain

Les récentes avancées technologiques se concentrent sur l'amélioration de l'utilité des agents intelligents pour l'homme. Il est crucial d'aligner les valeurs des robots sur celles des humains, qui sont eux aussi des êtres intelligents dotés de leurs propres états internes, objectifs et mécanismes de raisonnement. Ceci s'apparente grandement à une relation meneur-assistant (*leader-assistant*) comme décrit dans [27]. Pour ce faire, [28, 29] propose que, dans les approches de collaboration humain-machine, il est possible de distinguer trois catégories de modèles mentaux de l'agent humain : le modèle mental de premier ordre, le modèle mental de deuxième ordre et le modèle mental partagé. Ces modèles, représentés sur la Figure 2.1 influencent grandement la planification d'action d'un robot, notamment à travers leurs complexités et la variété des manières de les résoudre.

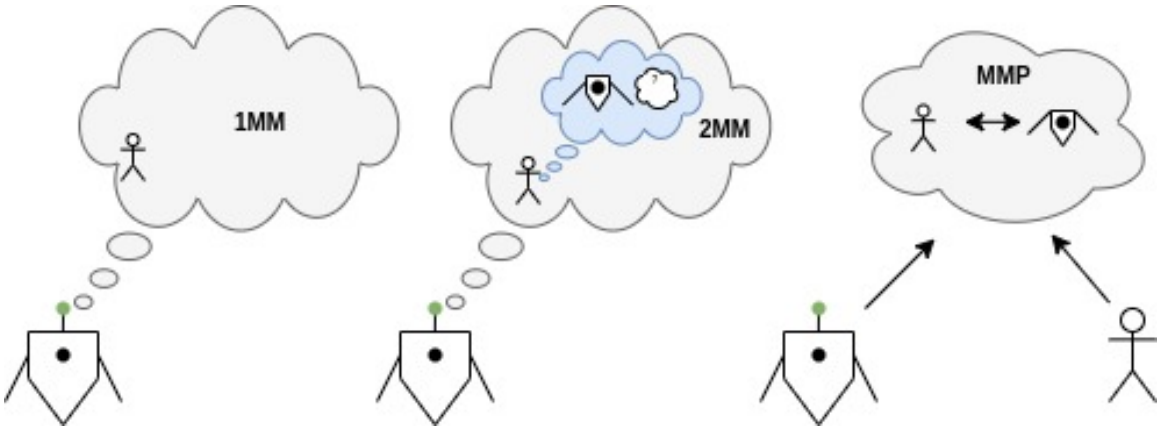


FIGURE 2.1 Représentation des trois modèles mentaux : modèle du premier ordre (1MM), celui du deuxième ordre (2MM) et le modèle mental partagé (MMP)

Modèle mental du premier ordre

Le modèle mental du premier ordre est défini dans [29] par la planification du robot dans un environnement où il considère que l'humain agit de façon indépendante et ne raisonne

pas en fonction des actions possibles du robot. Ainsi, toute action supplémentaire du robot dans son interaction avec l'humain aura pour but d'aider à l'accomplissement de la tâche, de l'entraver ou de n'avoir aucun impact. Plusieurs techniques tentent de résoudre le problème de la collaboration dans une équipe humain-machine en émettant des hypothèses sur les comportements de chaque acteur, telles que la quantité d'informations disponibles pour chacun d'eux. Dans la littérature relative à la collaboration humain-machine, le robot doit accomplir une tâche sans nécessairement connaître la fonction de récompense qui guide les actions de l'humain, autrement dit, son objectif.

Un modèle théorique décisionnel pour l'assistance intelligente peut être formalisé par un processus décisionnel de Markov partiellement observable à objectif caché (HGMDP) [30]. Dans cette structure, l'humain (l'agent dans leurs termes) et le robot (l'assistant) prennent des actions de manière séquentielle, l'assistant n'ayant pas connaissance du but g de l'humain. Le formalisme de ce type de problème tire sa similarité du processus de décision markovien partiellement observable (*Partially Observable Markov Decision Process*, POMDP) [31], avec la différence qu'il s'agit d'une sous-classe de ce dernier. Un HGMDP est défini par un tuple $\langle S, G, A, A', T, R, \pi, IS, IG \rangle$ où S est l'ensemble des états considéré comme totalement observable, G l'ensemble des buts possibles de l'agent, A l'ensemble des actions de l'humain, A' l'ensemble des actions du robot assistant, T la fonction de transition dépendant d'un but $g \in G$, R la fonction de récompense aussi dépendante d'un but, π la politique de l'agent qui associe $S \times G$ à une action sur A, IS la distribution initiale des états et IG la distribution initiale des buts. Autrement dit, un HGMDP est un ensemble de $|G|$ MDP où l'assistant est placé dans l'un de ces MDP sans savoir lequel. Aussi, un modèle restreint du HGMDP est proposé par [30] pour éviter la complexité inhérente liée à sa résolution, nommé l'action d'assistance MDP (*Helper Action MDP*, HAMDP). Ce modèle repose sur plusieurs hypothèses : l'humain est compétent pour maximiser les récompenses sans assistance, il détectera et exploitera toujours l'action d'assistance, et cette action ne nuira jamais à l'humain, même si elle s'avère inutile. Pour faire le lien avec notre travail au chapitre 3, nous adoptons également une architecture suivant le formalisme des POMDP, à la différence que l'assistant dispose de son propre modèle de transition. Cette différence a pour avantage d'ajouter une couche de liberté, facilitant la conception de divers scénarios d'application, tels que la possibilité de laisser le robot agir de manière totalement autonome.

D'autres littératures s'intéressent à la structure des POMDP en y ajoutant divers facteurs influençant la performance dans une équipe humain-machine. L'apprentissage inverse par renforcement [32, 33] résout ce problème en déduisant la fonction de récompense d'un agent à partir de l'observation de son comportement. Cela repose sur l'hypothèse de la rationalité des actions de l'humain ou des agents (hypothèse DBE, *demonstration-by-expert*, démons-

tration par expert), qui n'est souvent pas le comportement le plus optimal dans une tâche coopérative. Cela est démontré dans [34], avec l'introduction de l'apprentissage coopératif par renforcement inverse (*Cooperative Inverse Reinforcement Learning*, CIRL). En effet, cet apprentissage propose une structure de jeu d'information partielle à deux joueurs dans laquelle l'un des joueurs, l'humain (H), connaît la fonction de récompense, tandis que l'autre joueur, le robot (R), ne la connaît pas. La solution à ce jeu est une politique qui maximise la fonction de récompense de l'humain, alignée avec celle du robot. Ce problème peut être résolu en le réduisant simplement à un processus de décision markovien partiellement observable (*Partially Observable Markov Decision Process*, POMDP). Les auteurs étendent CIRL en incluant l'apprentissage par démonstration (*apprenticeship learning*) [35, 36], où l'humain démontre une tâche et le robot apprend à la réaliser. La démonstration de l'humain en isolation, basée sur l'hypothèse DBE, peut s'avérer sous-optimale pour un apprentissage efficace du robot. Une approche possible pour améliorer cette solution peut inclure une instruction active de la part de l'humain et un apprentissage actif de la part du robot, comme le montre la Figure 2.2. L'apprentissage coopératif par renforcement inverse standard n'intègre pas nativement une structure d'approche de solution incluant une anticipation des actions du robot par l'humain [34, 36].

Le modèle mental de premier ordre permet une approche basique aux problèmes de collaboration dans une équipe humain-machine. Cependant, cette approche n'est pas compatible avec l'aspect de coordination, pourtant nécessaire dans certains scénarios. Par exemple, dans un scénario de nettoyage dans un hangar où un agent d'entretien doit ramasser tous les matériaux, dont certains nécessitent l'aide du robot, ce dernier planifiera en considérant qu'il est impossible pour l'humain de finir la tâche dans le temps imparti, ignorant que l'humain prend en compte son assistance. Cela résultera en un comportement passif du robot, car il n'a aucune raison d'agir, sachant que, de son point de vue, l'humain ne sera pas capable de terminer la tâche dans les délais.

Certains travaux basant leur méthode sur le modèle mental de premier ordre (1MM) ont obtenu des résultats satisfaisants d'assistance dans des scénarios ne requérant pas de coordination entre l'humain et le robot en définissant le rôle du robot comme celui d'un assistant qui améliore ou simplifie les tâches à effectuer [37]. En effet, [37] présente un modèle basé sur les POMDP pour un robot collaboratif capable d'estimer les objectifs humains à l'aide des fonctions d'états et d'actions Q dérivées d'un comportement rationnel simulé. Le modèle s'adapte en temps réel aux changements d'objectifs et démontre de bonnes performances. Cependant, le modèle présente des limites, notamment dans l'hypothèse de rationalité de l'agent humain.

Ces travaux montrent que, bien que limité, le modèle 1MM peut-être efficace dans des contextes où le rôle du robot est clairement défini. Cette considération jouera un rôle important dans la résolution du problème développé dans le chapitre suivant, car nous exploiterons les avantages de l'utilisation du modèle 1MM.

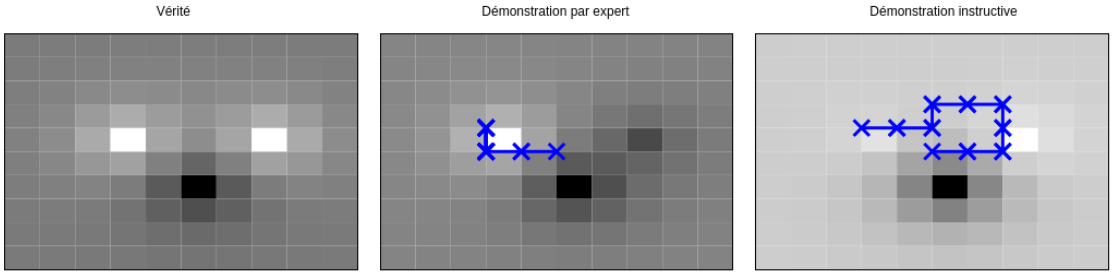


FIGURE 2.2 Différence entre la fonction de récompense déduite par le robot après démonstration par un expert et démonstration instructive . À gauche : la vrai fonction de récompense. Les cellules plus claires de la grille indiquent les zones où la récompense est plus élevée. Au milieu : La trajectoire (en bleu) est la démonstration générée par la politique d'un expert superposée à la fonction de récompense maximale a posteriori que le robot déduit. Le robot apprend où se trouve la récompense maximale, mais pas grand-chose d'autre. Sur l'image de droite : Une démonstration instructive générée par l'algorithme CIRL superposée à la fonction de récompense maximale a posteriori que le robot déduit. Cette démonstration met en évidence les deux cases de forte récompense et le robot apprend ainsi une meilleure estimation de la récompense.

Modèle mental du deuxième ordre

Dans ce modèle, le robot considère que l'humain prend en compte ses actions, ce qui implique que ce dernier doit posséder un modèle du robot, créant ainsi un problème de récursion. Plus précisément, la problématique de ce modèle réside dans la nécessité de modéliser les actions du robot, qui ne sont pas connues en raison des détails de conception des algorithmes de prise de décision ainsi que de la complexité de la récursion. En d'autres termes, une question peut être émise : quel est le modèle mental que l'humain a du robot ?

Plusieurs ouvrages se consacrent à élaborer des structures de prise de décision compatibles avec ce modèle. Ce type de modèle est abordé à travers [38] qui propose une méthode de génération des croyances imbriquées de manière heuristique dans le but de comprendre en langage naturel la pensée de certains agents vis-à-vis d'autres agents sur un sujet.

Plus récemment, [39] propose d'étendre le cadre des POMDPs aux environnements multi-agents en incorporant la notion de modèles d'agent dans l'espace d'état à travers une approche nommée *interactive POMDPs* (I-POMDPs). Les propriétés, telles que la convergence de l'itération de valeur ainsi que la linéarité par morceaux et la convexité de la fonction de valeur,

facilitent les calculs et permettent des solutions approximatives efficaces. Ce concept est repris par [40], qui aborde le problème de planification robuste des robots dans le cadre de la collaboration humain-robot. Comme mentionné précédemment, le modèle mental de second ordre présente une difficulté majeure : l'impossibilité d'accéder au modèle du robot. Cette problématique est contournée par l'utilisation de l'hypothèse selon laquelle l'humain contrôle entièrement le robot. En effet, [40] suppose temporairement que l'humain peut contrôler le robot avec un accès direct à ses observations, ce qui permet de calculer la fonction de valeur optimale d'un processus décisionnel markovien partiellement observable multi-agent (POMDP). Ensuite, en relâchant l'hypothèse de partage d'observations, ils extraient un modèle pour le comportement humain seul à partir de cette fonction de valeur. Les résultats montrent que cette approche améliore l'adaptabilité et la fiabilité des robots dans des scénarios collaboratifs. Cependant, bien que cette hypothèse soit plausible selon le design du robot, elle n'est pas toujours applicable, comme expliqué dans la description du scénario présenté à la section 1.2.2. Par conséquent, ce modèle mental ne sera pas retenu pour la définition du problème étudié dans les chapitres suivants de ce mémoire. Bien qu'il permette la coordination entre plusieurs agents, il est inévitable de traiter le problème de récursion dans le modèle de l'humain.

En raison de la similitude des structures des deux premiers modèles mentaux, nous explorons les limitations de l'utilisation du modèle mental de premier ordre (1MM) dans le scénario 1.2.2 qui fera objet d'une formulation plus précise dans le chapitre 3. Ce scénario n'implique pas la présence d'un robot collaboratif dans la construction du modèle mental de l'humain, de sorte que l'utilisation du 1MM élimine la contrainte de résoudre le problème en se basant sur des hypothèses supplémentaires.

Modèle mental partagé

Un modèle mental partagé est une représentation commune de l'environnement et des objectifs de la tâche, possédée par les membres d'une équipe, en l'occurrence un humain et un robot. Ce modèle réduit les conflits et les incertitudes en garantissant que les actions de chaque agent sont comprises et anticipées par les autres membres de l'équipe. En général, cette interaction s'apparente à celle entre plusieurs humains ayant une connaissance commune de la tâche à accomplir. Un exemple de formalisme pouvant accomplir un tel modèle est le processus de décision de Markov décentralisé et partiellement observable (*Decentralized POMDP*, Dec-POMDP) [41], où une politique commune est connue de chaque agent. Tous les agents optimisent selon une fonction de récompense connue et partagée.

Pour plus de détail, nous aborderons brièvement une méthode fondée sur le modèle mental

partagé (MMP). Afin d'exploiter ce modèle, [42] aborde la résolution du problème de coopération en s'inspirant des pratiques de formation des équipes humaines. Cette approche implique que les membres de l'équipe, qu'ils soient humains ou robots, échangent leurs rôles afin de mieux appréhender les attentes et les actions de chacun. Les résultats montrent que le cross-training entraîne des améliorations significatives des performances de l'équipe, tant en termes quantitatifs que de la perception de la performance du robot et de la confiance des humains envers celui-ci. Malgré les résultats probants obtenus par l'utilisation du MMP, celui-ci présente deux inconvénients majeurs qui justifient l'utilisation du modèle mental de premier ordre en comparaison dans la résolution de notre problème. Premièrement, dans le contexte de la collaboration homme-robot, il est difficile de satisfaire l'hypothèse de partage de politique. Deuxièmement, la résolution de tel problème grâce au formalisme de Dec-POMDP peut s'avérer complexe et requiert l'utilisation d'algorithmes de résolution plus sophistiqués.

2.2.3 Adaptation de l'humain au robot

Dans cette section, nous soulignons l'avantage du robot d'avoir accès à plusieurs types d'informations supplémentaires par rapport à l'humain. En effet, un agent intelligent possède une capacité de traitement de l'information supérieure à celle de l'humain, comme le démontrent les performances des ordinateurs. Cela conduit à l'hypothèse selon laquelle l'agent est capable d'avoir une meilleure interprétation de l'environnement. Par conséquent, la politique d'action suivie par l'humain n'est pas nécessairement la plus optimale, et il devient nécessaire de maximiser les préférences du robot. Cette stratégie de collaboration peut être catégorisée comme une relation d'égal à égal (*peer-to-peer*), intégrant un modèle humain de deuxième ordre signifiant qu'il est dépendant des actions du robot, comme mentionné dans la sous-section précédente.

L'accent est particulièrement mis sur la capacité de l'humain à s'adapter au robot, modifiant ainsi le modèle que le robot doit avoir de ce dernier. Un historique des actions du robot et des états de l'environnement est défini par [26] comme variable intervenant ainsi dans la modélisation de la politique de l'humain. Certains ouvrages profitent de cette structure en soulignant le fait que la taille de cette variable peut devenir très grande [43]. Pour contrer cet inconvénient, les auteurs de [44] utilisent l'hypothèse selon laquelle les êtres humains n'ont pas une mémoire parfaite et proposent de simplifier la conception du modèle de l'humain en utilisant le modèle d'adaptation humain à mémoire limitée (*Bounded-memory human Adaptation Model*, BAM). Ceci permet d'être utilisé dans des scénarios où le robot estime que l'humain observe une séquence de ses actions pour s'y adapter.

Dans certains scénarios, l'humain peut s'adapter à un certain degré ou, dans le pire des cas,

ne pas s'adapter du tout aux actions du robot. Si l'humain peut s'adapter aux actions du robot, une approche encore plus efficace consiste à concevoir une adaptation mutuelle. Cette dynamique bidirectionnelle est explorée dans la section suivante.

2.2.4 Adaptation mutuelle entre l'humain et le robot

L'adaptation mutuelle entre l'humain et le robot implique une couche de complexité supérieure à celle vue dans les sous-sections précédentes. En plus du fait que le robot essaie de discerner l'objectif de l'humain à partir de ses actions, il raisonne également sur le possible changement de cet objectif en fonction de son degré d'adaptation au robot [26].

Cette dynamique ajoute une dimension supplémentaire où le robot doit non seulement interpréter les actions de l'humain, mais aussi anticiper les ajustements de l'humain à ses propres actions. Contrairement à l'hypothèse sur la connaissance de l'adaptabilité de l'humain, les auteurs de [43, 44] relaxent cette hypothèse en traitant l'adaptabilité comme une variable inconnue dans une architecture de processus de décision de Markov à observabilité mixte (*Mixed Observability MDP*, MOMDP). Les résultats obtenus de ces articles montrent que l'adaptation mutuelle améliore la performance dans une équipe humain-machine comparativement à l'adaptation unilatérale du robot à l'humain. Cette approche permet de mieux synchroniser les actions des deux acteurs, optimisant ainsi l'efficacité et la réussite des tâches collaboratives.

Pour la suite de ce mémoire, cette stratégie ne sera pas adoptée, car, bien que puissante, elle est particulièrement adaptée aux scénarios nécessitant une coordination étroite entre les acteurs. Dans notre contexte, une approche moins complexe sera privilégiée pour répondre aux exigences spécifiques de notre problématique sans surcharger le modèle avec des mécanismes de coordination avancés.

2.3 Conclusion de la revue de littérature

La revue de littérature a permis de souligner les approches actuelles en matière de collaboration humain-machine, mettant en évidence les modèles mentaux utilisés pour la planification d'actions de robots dans des environnements collaboratifs. L'étude des différents modèles mentaux a montré que le modèle mental de premier ordre est un choix pertinent pour des tâches nécessitant un soutien, sans qu'il soit nécessaire de modéliser la réaction humaine aux actions du robot. Ce choix réduit considérablement la complexité computationnelle par rapport aux modèles mentaux de deuxième ordre ou partagés, tout en offrant un cadre d'assistance performant pour des tâches d'interaction simples à modéliser. La technique principale

qui sera utilisée dans ce mémoire se base sur les POMDPs pour des cas où l'objectif de l'humain est caché. Ceci permet de modéliser des robots avec des modèles mentaux du premier ordre et du deuxième ordre, en intégrant le modèle humain dans la dynamique du robot.

À titre de rappel de la sous-section 2.2.2, [37] a intégré un modèle mental de premier ordre dans la dynamique de décision du robot et a démontré l'efficacité de cette approche à travers des expérimentations. En s'appuyant sur ce travail et sur le scénario qui sera formalisé a la prochaine section, on étudie :

- La modification des techniques appliquées par [37] pour résoudre le problème défini à la section 1.2.2.
- L'incorporation d'un modèle d'apprentissage du paramètre de rationalité de l'agent humain dans le processus de décision du robot.
- La proposition d'un modèle plus complexe de décision du robot basé sur le modèle mental de second ordre.

CHAPITRE 3 PLANIFICATION DES DÉCISIONS DU ROBOT D'ORDRE 1

Dans ce chapitre, nous aborderons une approche de solution au problème d'assistance introduit dans la section 1.2.2. Dans ce scénario de collaboration humain-machine, nous considérons deux agents, un robot contrôlé à travers sa planification de décision et un humain exécutant des actions dans le but d'atteindre un ou plusieurs objectifs qui agissent en temps discret. Pour aider l'humain, on choisit de doter le robot d'un modèle mental de ce dernier afin de prédire ses actions et, par conséquent, planifier sa stratégie. Nous optons pour le modèle mental de premier ordre. Le modèle mental de second ordre est plus adapté aux tâches complexes de coordination, tandis que le modèle mental partagé repose sur une hypothèse avancée de collaboration humain-machine qui dépasse les besoins de notre scénario.

De plus, nous développons une structure algorithmique de planification d'assistance pour la prise des décisions du robot au sein d'une équipe humain-machine dans le cadre du scénario décrit à la section suivante 3.1.

Pour ce faire, nous modélisons le modèle de premier ordre de l'humain par une politique dérivée de la résolution d'un problème spécifique, inspiré de [37] et de [45] un concept très similaire à celui du HGMDP. Cette politique sera intégrée dans le calcul de la politique du robot sous deux hypothèses :

1. L'humain agit dans un environnement où un seul objectif peut être identifié.
2. L'humain agit dans un environnement ayant plusieurs objectifs possibles.

La section 3.1 introduit la formulation mathématique du problème, suivie de la discussion du modèle de planification de l'humain dans le modèle mental de premier ordre du robot en section 3.2. La structure de prise de décision du robot est abordée en section 3.3. Enfin, la section 3.4 présente une méthode d'apprentissage pour ajuster le modèle décisionnel du robot face aux erreurs de modélisation.

3.1 Formulation du problème en MDP

La compréhension du problème à résoudre repose sur la définition d'un formalisme mathématique capable d'intégrer la structure de prise de décision séquentielle. Un rappel du processus décisionnel markovien (MDP) est présenté dans la section 2.2.1, et la résolution du problème sous cette méthodologie est abordée dans les sections 3.2.1 et 3.3.3.

3.1.1 Description du scénario en MDP

Deux agents interviennent successivement dans un environnement, comme le montre la figure 3.1. Le robot observe l'humain, puis prend une décision d'assistance en exécutant une action, ce qui peut modifier l'état de l'environnement. Par la suite, l'humain agit à son tour, entraînant son déplacement d'un état à un autre, sans toutefois altérer l'environnement.

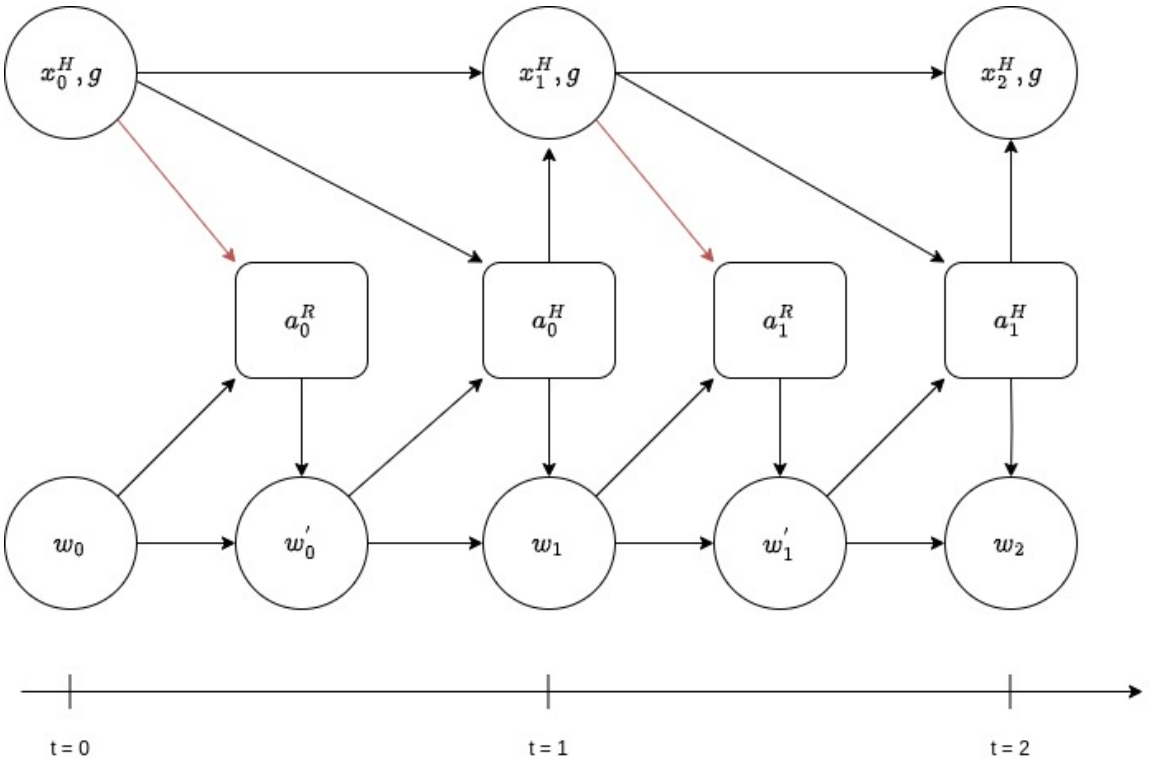


FIGURE 3.1 Diagramme des états représentant le scenario MDP sur une période de $t = [0, 2]$. Les flèches indiquent une dépendance directe entre variables. Les flèches rouges signalent que les éléments de l'état ne sont pas totalement observables, notamment la variable cachée g, qui représente l'objectif réel de l'humain. Les états x_t^H et g définissent l'état de l'humain. L'état de l'environnement après l'action a_t^H de l'humain est noté w_t, tandis que w_t' représente l'état de l'environnement après l'action a_t^R du robot à chaque instant.

3.1.2 Définition des états

Dans notre cadre, nous considérons deux agents distincts : l'humain et le robot. Chaque période de temps t se décompose en deux phases successives :

1. **Intervention du robot :**
   - Le robot effectue une action a_t^R appartenant à l'ensemble discret des actions A^R. On précise que le robot est modélisé comme une entité sans présence physique

propre et, par conséquent, ne possède pas d'état physique ; il ne conserve qu'un *état informationnel*.

- L'action du robot modifie l'état de l'environnement. Avant son action, l'environnement est dans l'état w_t (avec $w_t \in W$, l'ensemble discret des états possibles de l'environnement). Après l'action, il passe à l'état $w'_t \in W$.
- À chaque période, le robot observe l'état partiel de l'humain, c'est-à-dire son état $x_t^H \in X^H$. Cependant, il n'observe pas l'objectif g de l'humain (avec $g \in G$, G étant un ensemble discret).
- L'ensemble des informations accessibles au robot à l'instant t est résumé par l'état informationnel I_t, défini par

$$I_t = \{x_{0:t}^H,\, w_{0:t},\, a_{0:t-1}^H,\, a_{0:t-1}^R\}.$$

2. **Intervention de l'humain :**
   - L'état de l'humain est défini par le couple (x_t^H, g), où $x_t^H \in X^H$ représente son état à l'instant t et $g \in G$ est son objectif, supposé constant pendant toute la tâche.
   - L'humain observe l'état de l'environnement après l'intervention du robot, à savoir $w'_t \in W$.
   - En se basant sur cette observation, il exécute une action a_t^H appartenant à l'ensemble discret A^H.

Chaque transition entre les états de l'environnement et ceux de l'humain est dictée par des dynamiques de transition présentées dans la section 3.1.4. Pour éviter d'alourdir le schéma, la transition de l'état informationnel du robot a été omise.

Hypothèse 1. *Pour simplifier la modélisation, on suppose que l'humain planifie ses actions en résolvant un processus de décision markovien (MDP) avec un modèle du monde incomplet. Ainsi, il part du principe que l'état de l'environnement w'_t restera constant dans le futur. Cette hypothèse rend la planification de l'humain moins rationnelle.*

Le scénario est analysé sous l'hypothèse 1 selon deux approches. La première examine le comportement du robot lorsque l'humain accomplit une trajectoire avec un objectif connu, c'est-à-dire avec un objectif unique et identifié. La seconde étudie son comportement en présence de plusieurs objectifs potentiels, ce qui complexifie la structure du problème en la rapprochant d'un POMDP. Des stratégies adaptées à ces situations seront développées et leurs limitations évaluées.

3.1.3 Définition des actions

Les actions du robot, notées a_t^R, appartiennent à l'ensemble des actions possibles A^R. De même, les actions de l'humain, notées a_t^H, appartiennent à A^H. Ces actions sont déterminées par des politiques de décision qui guident chaque agent en fonction des informations disponibles.

Définition 1. *Dans un processus de décision markovien (MDP), une politique π définit une règle de décision associant un état $x \in X$ à une action $a \in A$, soit $\pi : X \to A$.*

Plus précisément :

- *La politique de l'humain π_H est une fonction qui associe à tout tuple d'état (x^H, g, w') une distribution de probabilité sur l'ensemble discret des actions A^H.*
- *La politique du robot π_R est une fonction qui détermine a^R à partir de l'état informationnel I_t, de l'état de l'humain x^H et de l'environnement w.*

Ainsi, la politique π_R du robot dépend de l'état informationnel I_t, qui synthétise l'historique des observations, ainsi que de x_t^H et w_t. L'humain, en revanche, prend ses décisions selon π_H, en fonction de son propre état (x_t^H, g) et de l'état de l'environnement w_t'.

L'hypothèse 1 impose que l'humain planifie sans tenir compte des actions du robot. Toutefois, cette simplification ne compromet pas l'efficacité de ses trajectoires. En effet, les actions du robot modifient l'environnement, et l'humain, en observant le nouvel état w_t', ajuste sa stratégie en conséquence. Cette replanification, inhérente aux MDP, lui permet d'atteindre son objectif même sans modéliser explicitement l'intervention du robot.

3.1.4 Définition des fonctions de transition

La fonction de transition décrit l'évolution des états en fonction des actions effectuées. Elle définit la probabilité de passer d'un état x_t à un état x_{t+1} après l'exécution d'une action a_t, comme introduit en section 2.2.1. Chaque composante du système suit une dynamique spécifique, détaillée ci-dessous.

Transition de l'environnement

L'évolution de l'environnement w_t dépend des actions du robot et de l'humain. Le robot peut modifier l'environnement via son action a_t^R, ce qui lui permet d'assister l'humain. La transition de w_t sous l'effet du robot est considérée dans la suite de ce mémoire comme déterministe mais dans un contexte générale s'écrit :

$$\mathbb{P}[w_t^{'} = w^{'}|w_t, a_t^R] = T_{w^{'}}(w_t, a_t^R, w^{'}) \tag{3.1}$$

Ceci représente la probabilité pour l'environnement de passer de w_t à un état $w^{'}$ apres une action a_t^R. L'humain ajuste alors sa stratégie en fonction du nouvel état $w_t^{'}$. En revanche, lorsqu'il agit, l'environnement reste inchangé. Dans certains travaux, l'humain peut également influencer l'environnement, notamment dans des scénarios de coopération bilatérale [46, 47], où les deux agents collaborent pour atteindre un objectif commun. Toutefois, dans ce cadre, l'environnement ne change pas sous l'action de l'humain, on suppose donc :

$$w_{t+1} = w_t^{'} \tag{3.2}$$

Dans des travaux futurs, on pourra supposer des fonctions de transition plus générales, $\mathbb{P}[w_{t+1} = w|w_t^{'}, a_t^H] = T_w^H(w_t^{'}, a_t^H, w)$.

Transition de l'état de l'humain

L'état de l'humain x_t^H évolue de manière déterministe selon :

$$\mathbb{P}[x_{t+1}^H = x^H|x_t^H, w_t^{'}, g, a_t^H] = T_x^H(x_t^H, w_t^{'}, g, a_t^H, x^H) \tag{3.3}$$

Ce choix d'une transition déterministe vise à faciliter l'interprétation des actions de l'humain et à rendre le comportement du robot plus explicable. Si la transition était stochastique, le robot pourrait rencontrer des difficultés à inférer l'objectif réel de l'humain. Toutefois, des travaux futurs pourraient explorer des stratégies plus robustes face à cette stochasticité.

Transition de l'état du robot

L'état du robot I_t est purement informationnel et consiste en l'historique des observations des états et actions de l'humain et du robot. Il est défini par :

$$I_t = \{x_{0:t}^H, w_{0:t}, a_{0:t-1}^H, a_{0:t-1}^R\} \tag{3.4}$$

$$I_{t+1} = I_t \cup [x_{t+1}^H, w_{t+1}, a_t^H, a_t^R] \tag{3.5}$$

3.1.5 Définition de la fonction de récompense

La fonction de récompense R associe à une paire état-action (x, a) une récompense obtenue par l'agent lorsqu'il exécute l'action a à l'état x. Il est important de noter que les fonctions de récompense des deux acteurs sont distinctes. Dans certains travaux sur l'apprentissage par renforcement inverse, comme celui présenté dans [34], l'hypothèse sous-jacente à leur méthode de résolution suppose que la fonction de récompense apprise est similaire à celle démontrée. Leur objectif est donc de reproduire le comportement de l'agent démonstrateur. Dans notre cas, l'objectif n'est pas de répliquer ce comportement, mais plutôt de favoriser l'atteinte de comportements aidants, poussant donc à la modélisation de deux fonctions de récompenses distinctes.

Ainsi, la fonction de récompense de l'humain à l'instant t est donnée par :

$$R_t^H = R^H(x_t^H, w_t^{'}, a_t^H, g) \tag{3.6}$$

D'autre part, la fonction de récompense du robot à l'instant t est définie comme suit :

$$R_t^R = R^R(x_t^H, w_t, a_t^R) \tag{3.7}$$

Les fonctions de récompense définissent les objectifs à optimiser afin de déterminer le comportement de chaque agent. Chaque agent cherche ainsi à prendre des actions maximisant les récompenses accumulées sur l'ensemble de la trajectoire, c'est-à-dire la séquence d'états traversés. Pour assister efficacement l'humain, le robot doit être capable d'anticiper son comportement. Cela implique une connaissance, de la fonction de récompense de l'humain afin d'optimiser une fonction objective spécifique, définie à la section 3.3.1. À cette fin, une nouvelle hypothèse est formulée :

Hypothèse 2. *Le robot a accès à la fonction de récompense de l'humain* R_t^H.

3.1.6 Rappel des méthodes de resolution d'un MDP

Définition de l'objectif

En général dans un MDP (voir définition dans la section 2.2.1), l'objectif est d'identifier la politique optimale π^* qui maximise le retour cumulatif de récompenses. La récompense immédiate ne suffit pas pour évaluer cette politique. Par exemple, bien que l'on puisse être tenté de maximiser la récompense immédiate, il est parfois préférable de sacrifier un gain

instantané pour obtenir des récompenses plus élevées à long terme.

On introduit alors la fonction de valeur de l'état $V^{\pi}(x)$, qui évalue une politique π à partir d'un état x :

$$V^{\pi}(x) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^{t} R(x_t, \pi(x_t)) | x_0 = x\right] \tag{3.8}$$

Le facteur d'actualisation γ (*discount factor*) représente la valeur relative des récompenses futures par rapport aux récompenses immédiates. Ainsi, un γ proche de 0 favorise les récompenses immédiates, tandis qu'une valeur proche de 1 met davantage l'accent sur les récompenses futures. Dans notre cas, résoudre le MDP dans un horizon infini, comme indiqué dans l'équation (3.8), nécessite un γ dans l'intervalle $[0, 1)$, garantissant la convergence de l'équations de Bellman. Cette équation se réécrit alors ainsi, avec $a = \pi(x)$:

$$V^{\pi}(x) = R(x, a) + \gamma \sum_{x' \in X} T(x, a, x') V^{\pi}(x') \tag{3.9}$$

Pour évaluer la valeur spécifique d'un état x lorsque l'agent choisit une action a, on définit la fonction de valeur de l'action à l'état (*state-action value function*) comme suit :

$$\begin{aligned} Q^{\pi}(x, a) &= \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^{t} R(x_t, a_t) | x_0 = x, a_0 = a\right] \\ &= R(x, a) + \gamma \sum_{x' \in X} T(x, a, x') V^{\pi}(x') \end{aligned} \tag{3.10}$$

La politique optimale π^* doit maximiser la fonction de valeur :

$$\pi^*(x) = \arg\max_{\pi} V^{\pi}(x_0 = x) \tag{3.11}$$

On a :

$$V^*(x) = \max_{a}[R(x, a) + \gamma \sum_{x' \in X} T(x, a, x') V^*(x')] \tag{3.12}$$

De manière équivalente, la fonction de valeur d'action de l'état optimale Q^* découle de l'équation de Bellman :

$$Q^*(x,a) = R(x,a) + \gamma \sum_{x' \in X} T(x',a,x) \max_{a'} Q^*(x',a') \tag{3.13}$$

Grâce à ces équations, nous garantissons l'obtention de la politique optimale en raison du principe d'optimalité, qui stipule que, pour obtenir une politique optimale, chaque sous-problème de décision doit également être optimal. Ce principe assure qu'en trouvant la valeur optimale pour chaque état, on obtient la politique optimale globale. Ceci s'observe par le fait que la valeur de la fonction à l'état suivant x' est optimale.

Résolution d'un MDP

Pour résoudre un problème de MDP, plusieurs algorithmes d'optimisation existent, mais une distinction importante repose sur la manière de résoudre le MDP : en ligne ou hors ligne.

Dans une planification hors ligne, la politique est pré-calculée avant l'interaction de l'agent avec l'environnement, offrant une optimalité globale si l'environnement est statique et parfaitement connu. En revanche, la planification en ligne adapte la politique en temps réel à chaque interaction avec l'environnement, convergeant progressivement vers une solution optimale en environnement dynamique. En termes de ressources computationnelles, la planification en ligne utilise moins de ressources au moment de l'établissement de la politique, mais en nécessite davantage pendant l'exécution [48]. À l'inverse, la planification hors ligne consomme des ressources importantes pour calculer une politique optimale couvrant tous les états possibles.

Dans ce mémoire, la résolution du problème repose sur une approche de planification hors ligne. La programmation dynamique, introduite par [49], est une méthode permettant de résoudre un problème complexe en calculant de manière itérative la fonction de valeur optimale. Bien que d'autres méthodes, telles que l'algorithme de Dijkstra [50], soient plus efficaces pour les problèmes déterministes en les traitant comme un problème de plus court chemin, la programmation dynamique a été retenue dans notre cas en raison de sa capacité à gérer les incertitudes. Dans le contexte des MDP, la programmation dynamique se résoud principalement à travers deux algorithmes classiques : l'itération de valeur et l'itération de politique [51].

Itération de valeur : L'algorithme 1 procède par des itérations successives sur les valeurs $V(x)$ en appliquant l'équation de Bellman (voir (3.12)), jusqu'à ce que la différence entre les valeurs de tous les états entre deux itérations consécutives soit négligeable. Ce processus permet d'obtenir V^*, à partir duquel on déduit π^* grâce à l'équation (3.11). La complexité d'une itération de cet algorithme s'évalue en $\mathcal{O}(|X^2||A|)$ pour X et A definit en section 2.2.1.

Algorithm 1: Value Itération

Input: Un MDP $\langle X, A, T, R\rangle$, facteur d'atténuation γ, un paramètre de précision ϵ
Output: Politique approximative optimale π^* et la fonction valeur approximative optimale V^*

Initialiser arbitrairement $V^*(x)$, $\forall x \in X$;
repeat
  for *chaque* $x \in X$ **do**
    $temp \leftarrow V^*(x)$
    $V^*(x) \leftarrow \max_{a \in A}[r(x,a) + \gamma \sum_{x' \in X} T(x,a,x')V^*(x')]$
until $\max_{x \in X}|temp - V^*(x)| \leq \epsilon$;
$\forall x, \pi(x) = \arg\max_{a \in A}[r(x,a) + \gamma \sum_{x' \in X} T(x,a,x')V^*(x')]$

Itération de politique : L'itération de politique repose sur deux étapes majeures : l'évaluation et l'amélioration de la politique. Ces étapes sont répétées jusqu'à ce que la politique ne change plus (politique stable). Sa complexité computationnelle est de l'ordre de $\mathcal{O}(|X|^3 + |X|^2|A|)$ dans le pire cas.

Les algorithmes de planification en ligne sont brièvement abordés ici. Toutefois, ils ne sont pas utilisés dans ce mémoire, bien qu'ils soient plus adaptés à l'humain, qui doit replanifier à partir de son nouvel état courant. Les méthodes basées sur l'échantillonnage constituent des solutions palliatives aux problèmes de dimensionnalité. Comme vu dans les méthodes de programmation dynamique, le coût de calcul de l'espérance sur tous les états futurs augmente de manière exponentielle avec le nombre d'états [52]. La recherche arborescente de Monte Carlo (*MCTS*) est une méthode populaire qui ne nécessite pas une connaissance préalable de tous les états. Elle utilise un générateur prenant l'état actuel et une action pour construire un arbre de recherche et choisir la meilleure décision [53].

D'après l'analyse de cette sous-section, la méthode d'itération des valeurs sera utilisée et adaptée pour notre problème en raison de sa complexité computationnelle réduite. Puisque tous les états de l'environnement sont connus initialement, il n'est pas pertinent de développer des méthodes de planification en ligne. Toutefois, dans des travaux futurs, ces méthodes pourraient s'avérer essentielles en raison de l'augmentation de la dimensionnalité du problème.

3.2 Modèle mental du premier ordre

On rappelle la définition du modèle mentale du premier ordre donnée à la section 2.2.2. L'humain ne suit pas une stratégie parfaitement rationnelle. Pour modéliser son comportement,

le robot adopte alors une approche en deux étapes :

1. **Modèle intermédiaire (MDP)** : Le robot résout un MDP (en considérant l'hypothèse 2) en supposant que l'humain agit de manière parfaitement rationnelle selon un objectif donné g. Chaque MDP défini ainsi permet de calculer une fonction de valeur d'état et d'action $Q^H(x^H, a^H, g)$. Comme illustré dans la figure 3.2, une liste de MDP est associée à l'ensemble des objectifs possibles de l'humain.
2. **Politique finale (Boltzmann)** : L'humain ne suit pas strictement la politique MDP optimale. Le robot suppose plutôt que l'humain adopte une politique stochastique de type Boltzmann basée sur Q^H. Cette politique reflète une prise de décision non parfaitement rationnelle, où les actions sont choisies de manière probabiliste en fonction de leur valeur estimée.

Ainsi, bien que la résolution du MDP soit une étape essentielle, elle ne représente pas directement la politique finale de l'humain. Le robot intègre ensuite cette fonction de valeur dans un modèle Boltzmann pour estimer la distribution réelle des actions humaines. Cette approche permet au robot d'interpréter et d'anticiper les décisions humaines avec une meilleure robustesse face aux variations de comportement.

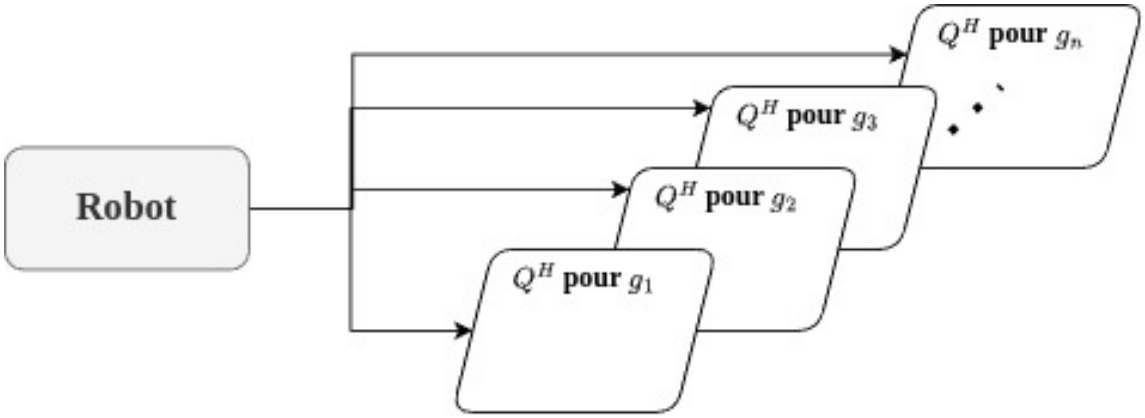


FIGURE 3.2 Fonction Q^H calculable par le robot pour un ensemble d'objectifs g_i , $i = (1, 2, .., n)$ en résolvant un MDP.

3.2.1 Modèle intermédiaire de l'humain (MDP)

Le MDP de l'humain est représenté par le tuple $(X^H, W, A^H, T_x^H, R^H, G)$, avec chaque élément détaillé dans les sections 3.1.2 et 3.1.4. Notons que l'état x^H de l'humain appartient à X^H, peu importe son objectif. L'objectif de l'humain, du point de vue du robot, consiste à maximiser sa fonction de récompense à chaque instant t. Cette perspective permet au robot de déduire que l'humain choisira une action qui optimise sa satisfaction dans l'état actuel de la façon suivante :

$$\arg\max_{a_t^H} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R^H(x_t^H, w_t^{'}, a_t^H, g) | x_0^H, w_o^{'}, g\right]$$

La fonction de récompense R_t^H est définie à l'équation (3.6). Comme illustré dans la figure 3.1, l'humain intervient après l'action du robot, en réponse à l'état modifié $w_t^{'}$. On développe la fonction de valeur suivante :

$$Q^H(x_t^H, w_t^{'}, a_t^H, g) = R_t^H + \gamma \sum_{x_{t+1}^H \in X^H} T_x^H(x_t^H, w_t^{'}, g, a_t^H, x_{t+1}^H) V^H(x_{t+1}^H, w_{t+1}, g) \tag{3.14}$$

La transition entre $w_t^{'}$ et w_{t+1} n'apparaît pas dans l'équation ci-dessus en raison de sa trivialité (voir section 3.1.2). On considère ainsi :

$$V^H(x_{t+1}^H, w_{t+1}, g) = \max_{a_{t+1}^H \in A^H} Q^H(x_{t+1}^H, w_{t+1}, a_{t+1}^H, g) \tag{3.15}$$

En intégrant cette formulation dans l'équation (3.14), il apparaît qu'il suffit de la fonction Q pour déduire la stratégie de l'humain pour un objectif g donné. Les fonctions Q sont calculées en utilisant l'algorithme d'itération de valeur, adapté pour répondre aux spécificités de notre problème. Cette adaptation permet de tenir compte des particularités de l'environnement et des objectifs de l'humain, comme le montre l'algorithme 2.

Algorithm 2: Itération des valeurs MDP de l'humain

Input: Un MDP $\langle X^H, W, A^H, T_x^H, R^H, G\rangle$, facteur d'atténuation γ, un paramètre de précision ϵ

Output: $Q^H(x^H, w^{'}, a^H, g)$

Initialiser arbitrairement $V(x^H, w^{'}, g)$, $\forall x^H \in X^H, w^{'} \in W, g \in G$

Initialiser arbitrairement $Q(x^H, w^{'}, a^H, g)$, $\forall x^H \in X^H, w^{'} \in W, g \in G, a^H \in A^H$

repeat

  for *chaque* $g \in G, w^{'} \in W, x^H \in X^H$ **do**

    temp $\leftarrow V(x^H, w^{'}, g)$

    for *chaque* $a^H \in A^H$ **do**

      $Q^H(x^H, w^{'}, a^H, g) \leftarrow R^H + \gamma \sum_{x^{H,p} \in X^H} T_x^H(x^H, w^{'}, g, a^H, x^{H,p}) V(x^{H,p}, w^p, g)$

    $V^H(x^H, w^{'}, g) \leftarrow \max_{a^H \in A^H} Q^H(x^H, w^{'}, a^H, g)$

until $\max_{x^H, w^{'}, g} |temp - V(x^H, w^{'}, g)| \leq \epsilon$;

Après convergence de l'algorithme 2, on obtient une fonction optimale Q^H. Par exemple, supposons que $Q^H(x_t^H, w_t, a_t^H, g_1) = y_1$ et que $Q^H(x_t^H, w_t, a_t^H, g_2) = y_2$ pour le même état et action, mais avec des objectifs différents g_1 et g_2. Si $y_1 \geq y_2$, le robot peut en déduire que l'humain est plus susceptible de suivre l'objectif g_1 plutôt que g_2 à ses états.

3.2.2 Politique finale de l'humain à rationalité limitée

La sous-section précédente expose une modélisation de la rationalité humaine telle que perçue par le robot. Cependant, comme le souligne [54], l'humain n'agit pas de manière parfaitement rationnelle. Cela signifie que les fonctions optimales Q^H obtenues à partir de l'algorithme 2 ne suffisent pas à prédire avec précision ses actions. Pour pallier cette limitation, l'hypothèse suivante sera adoptée pour la suite de cette étude :

Hypothèse 3. *L'humain ne se comporte pas de manière entièrement rationnelle et peut être considéré comme un agent à rationalité limitée.*

La rationalité limitée désigne la capacité d'un agent à prendre des décisions rationnelles sous des conditions de limites cognitives, une information imparfaite et des contraintes temporelles, contrairement à la rationalité complète qui suppose qu'un agent prend toujours des actions rationnelles. Cette notion assouplit l'hypothèse de rationalité parfaite pour mieux tenir compte de l'incertitude de l'environnement et des facteurs influençant l'agent humain. Ainsi, les décisions de l'humain peuvent s'écarter de la maximisation stricte des fonctions de récompense, reflétant une prise de décision plus réaliste.

La modélisation de la rationalité limitée est un domaine bien établi, utilisé en intelligence artificielle et en économie pour sa pertinence. La politique de Boltzmann, ou softmax, fait partie de ces méthodes. Cette approche de prise de décision attribue des probabilités aux actions selon leurs valeurs respectives, introduisant ainsi une composante aléatoire qui simule les décisions imparfaites d'agents aux capacités cognitives limitées [55]. D'autres techniques incluent la politique gourmande, qui affecte une probabilité légèrement inférieure à 1 à la prise d'une action, et les fonctions de valeur aléatoire, qui maintiennent une distribution des valeurs $V(x)$ plutôt qu'une seule estimation.

Dans ce mémoire, nous optons pour l'utilisation de la politique de Boltzmann, qui s'avère efficace pour modéliser la rationalité limitée. Elle s'appuie sur les valeurs Q^H, calculées par le robot, tout en intégrant une composante aléatoire. Cette approche permet d'ajuster la probabilité de chaque action en fonction de sa valeur Q^H, tout en laissant une marge de sous-optimalité due aux variations comportementales, ce qui rend le modèle plus flexible et réaliste dans des contextes d'incertitude comportementale de l'humain.

3.2.3 Paramètres et ajustement de la politique de Boltzmann

La politique de Boltzmann [56] se décrit par la formule :

$$\begin{aligned}\mathbb{P}_{a^H,w_t^{'}}(g) = \mathbb{P}[a_t^H \mid x_t^H, w_t^{'}, g] &\propto \exp(\beta Q_t^H(x_t^H, w_t^{'}, a_t^H, g)) \\ &= \frac{e^{\beta Q_t^H(x_t^H, w_t^{'}, a_t^H, g)}}{\sum\limits_{a \in A^H} e^{\beta Q_t^H(x_t^H, w_t^{'}, a, g)}} \\ &= \eta e^{\beta Q_t^H(x_t^H, w_t^{'}, a_t^H, g)} \end{aligned} \tag{3.16}$$

La probabilité $\mathbb{P}[a_t^H \mid x_t^H, w_t^{'}, g]$ indique la probabilité de l'humain à prendre l'action a_t^H lorsqu'il est dans l'état x_t^H de l'environnement $w_t^{'}$ avec pour objectif g. Le paramètre β appelé indice de rationalité est le paramètre de température qui ajuste cette probabilité : quand β est proche de 0 la probabilité de prendre une certaine action devient aléatoire suggérant que l'humain agit de manière non rationnel (voir l'annexe B) et avec une valeur assez élevée de β, l'humain adopte une stratégie parfaitement rationnelle, choisissant systématiquement les actions optimales pour atteindre son objectif. η est le facteur de normalisation.

Algorithm 3: Calcul de la distribution de Boltzmann

Input: $Q^H(x^H, w^{'}, a^H, g)\ \forall x^H, w^{'}, g \in X^H, W, G$
Output: probabilité $\mathbb{P}[a_t^H \mid x^H, w^{'}, g]\ \forall x^H, w^{'}, g \in X^H, W, G$

```
Initialiser a zéro P[a_t^H | x^H, w', g], ∀x^H ∈ X^H, w' ∈ W, g ∈ G, a^H ∈ A^H
for chaque g ∈ G, w' ∈ W, x^H ∈ X^H do
    η ← 0
    for chaque a^H ∈ A^H do
        P[a_t^H | x^H, w', g] ← e^{βQ^H(x^H, w', a^H, g)}
        η ← η + P[a_t^H | x^H, w', g]
    η ← 1/η
    for chaque a^H ∈ A^H do
        P[a_t^H | x^H, w', g] ← ηP[a_t^H | x^H, w', g]
```

L'algorithme 3 décrit la procédure permettant d'obtenir une distribution de probabilité à partir des fonctions Q^H. Grâce au facteur de normalisation η, la somme des distributions de probabilité associées à un état donné pour chaque action possible est égale à 1, soit $\sum\limits_{a^H \in A^H} \mathbb{P}[a^H \mid x^H, w^{'}, g] = 1$. En conséquence, la politique de l'humain est établie comme :

$$\pi_H(x_t^H, w_t^{'}, a_t^H, g) = \mathbb{P}_{a_t^H, w_t^{'}}(g) \propto \exp(\beta Q_t^H(x_t^H, w_t^{'}, a_t^H, g)) \tag{3.17}$$

Avec cette politique, il est possible d'incorporer le comportement de l'humain dans le processus de décision du robot.

3.3 Planification du robot

Cette section exploite le modèle de comportement semi-rationnel de l'humain, tel que modélisé précédemment, pour formuler le processus de prise de décision du robot. L'observation du comportement du robot, lorsqu'il interagit avec un humain ayant un objectif unique, met en lumière son rôle d'assistance et établit des bases pour développer sa stratégie dans des scénarios où l'humain poursuit un objectif incertain parmi plusieurs objectifs. À cette fin, la première sous-section introduira les POMDPs, clarifiant les concepts nécessaires pour résoudre les problématiques rencontrées par le robot.

3.3.1 Cas à objectif unique

Cette sous-section porte sur le développement d'un algorithme d'assistance dans un scénario où l'humain exécute une tâche unique. Dans ce contexte simplifié, le POMDP initial est réduit à un MDP, ce qui permet de définir le problème sous la forme d'un tuple $(X^H, W, A^H, A^R, T_x^H, T_{w'}, R^H, R^R, G)$. Ici, G représente un seul objectif, permettant ainsi de focaliser la planification et la prise de décision du robot dans un cadre déterministe pour cette tâche spécifique. Dans cette situation, l'objectif du robot est :

$$\arg\max_{a_t^R} \left\{ \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \left(R_t^H + \lambda R^R(x_t^H, w_t, a_t^R) \right) \middle| x_0^H, w_0 \right] \right\} \tag{3.18}$$

Le robot, ayant accès à sa propre fonction de récompense et à celle de l'humain, peut maximiser l'équation (3.18) en ajustant le paramètre $\lambda = \{-1, 1\}$. Ce paramètre contrôle l'attitude du robot envers l'humain : une valeur de -1 indique une volonté de gêner (diminuer la récompense de l'humain), tandis qu'une valeur de 1 indique une volonté d'assister (augmenter la récompense totale) [57]. Dans ce mémoire, étant donné l'objectif d'assister l'humain, λ est

fixé à 1. La fonction Q suit donc l'équation de Bellman :

$$Q_t^R(x_t^H, w_t, a^R) = R^R(x_t^H, w_t, a^R) + \sum_{w' \in W} T_{w'}(w_t, a^R, w') \Bigg[\sum_{a^H \in A^H} \mathbb{P}_{a^H, w'} R_t^H + \gamma \sum_{x_{t+1}^H \in X^H} \sum_{a^H \in A^H} \mathbb{P}_{a^H, w'} T_x^H(x_t^H, w', g, a^H, x_{t+1}^H) V^R(x_{t+1}^H, w') \Bigg] \tag{3.19}$$

avec $w' = w_{t+1}$. $\mathbb{P}_{a^H, w'}$ représente la distribution de probabilité sur les actions possibles de l'agent humain (voir équation (3.16)), tandis que R_t^R est définit par l'équation (3.7). Pour des détails supplémentaires sur les étapes menant à cette équation, se référer à la section A de l'annexe. Les équations ci-dessous définissent la politique optimale à suivre :

$$V^R(x_t^H, w_t) = \max_{a^R \in A^R} Q^R(x_t^H, w_t, a^R) \tag{3.20}$$

$$a_t^R = \underset{a^R \in A_R}{argmax} \; Q^R(x_t^H, w_t, a^R) \tag{3.21}$$

Pour résoudre ce problème, une variante de l'algorithme d'itération des valeurs a été utilisée (algorithme 4), légèrement modifiée par rapport à l'algorithme 2. La figure 4.8 montre le processus général de préparation du robot pour interagir avec un humain. Ce processus débute par l'inférence de la politique de l'humain, modélisée selon une politique de Boltzmann. Cette politique sert ensuite de base pour établir la politique du robot, conformément à l'équation (3.21).

Algorithm 4: Itération de valeur MDP du robot (but unique)

Input: Un MDP $(X^H, W, A^H, A^R, T_x^H, T_{w'}, R^H, R^R, G)$, $\mathbb{P}_{a^H,w'}(.)$ obtenu à partir du modèle mental de premier ordre, facteur d'atténuation γ, un paramètre de précision ϵ

Output: Politique optimale π_R

Initialiser arbitrairement $V^R(x^H, w)$, $\forall x^H \in X^H, w \in W$
Initialiser arbitrairement $Q^R(x^H, w, a^R)$, $\forall x^H \in X^H, w \in W, g \in G, a^R \in A^R$
repeat
  for *chaque* $w \in W, x^H \in X^H$ **do**
    temp $\leftarrow V^R(x^H, w)$
    for *chaque* $a^R \in A^R$ **do**
      for *chaque* $a^H \in A^H$ **do**
        Appliquer l'équation (3.19)
    Appliquer l'équation (3.20)
until $\max_{x^H,w} |temp - V^R(x^H, w)| \leq \epsilon$;
$\forall x^H, w, g, \pi_R(x^H, w) = \arg\max_{a^R \in A^R} Q^R(x^H, w, a^R)$

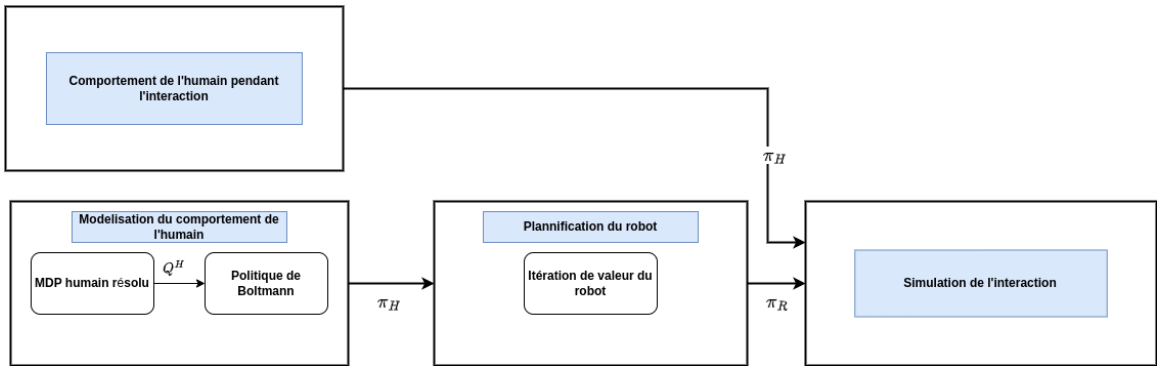


FIGURE 3.3 Phase de préparation du robot, illustrant la planification de ses décisions (π_R) en s'appuyant sur un modèle du comportement humain (π_H), supposé fidèle à celui que l'humain adoptera lors de l'interaction.

La figure 4.8 met en évidence un point essentiel : la stratégie de décision du robot est définie à partir d'un modèle préétabli du comportement humain. Cette approche repose sur l'hypothèse que le modèle utilisé est exact à cette phase. La section 3.4 introduira une méthode visant à réduire les erreurs de modélisation lorsque le comportement réel de l'humain pourrait différer de celui anticipé par le robot lors de la planification.

Pour comprendre la signification d'une simulation complète dans le contexte d'une assistance où une seule tâche précise doit être accomplie par l'humain, il est essentiel de considérer les étapes algorithmiques impliquées (algorithme 5). Cela inclut la planification des actions du robot en fonction de la gestion des états de l'environnement et la prise en compte des réactions humaines.

Algorithm 5: Algorithme d'interaction entre le robot et l'humain

Input: $(X^H, W, A^H, A^R, T_x^H, T_{w'}, R^H, R^R, G)$

Calculer la politique de Boltzmann avec l'algorithme 3
Calculer la politique π_R avec l'algorithme 4
g est connu par le robot
$t \leftarrow 0$
Initialiser l'environnement a w_0
Initialiser l'agent humain a l'état x_0^H
repeat
    Robot agit avec une action a_t^R selon l'équation (3.21)
    Mettre a jour l'environnement $w_t \leftarrow w_t^{'}$
    Recevoir la récompense du robot R_t^R
    Humain agit avec une action a_t^H selon l'équation (3.17)
    Recevoir la récompense du robot R_t^H
    Mettre a jour l'environnement $w_t \leftarrow w_{t+1}$ et l'état de l'humain $x_t^H \leftarrow x_{t+1}^H$
    $t \leftarrow t + 1$
until $x_t^H == g$;

Les détails de l'implémentation, ainsi que les résultats et la discussion relatifs à ce scénario, sont abordés dans la section 4.3.2.

3.3.2 Rappel des problèmes à observation partielle

Dans de nombreux cas, l'agent n'a pas accès à l'état dans notre cas l'état g, mais uniquement à une observation partielle ou altérée de celui-ci. Un POMDP est défini par un tuple $(X, A, T, \Omega, O, R, b_0)$, où (X, A, T, R) prennent les définitions d'un MDP comme présentées à la section 2.2.1, et :

- Ω est un ensemble fini d'observation.
- O est la dynamique d'observation représentant la probabilité d'observer o en allant à l'état $x^{'}$ en prenant l'action $a : O(x, a, o) := \mathbb{P}[o|x, a]$
- $b_0(x) = \mathbb{P}[x_0 = x]$ est la probabilité que l'agent soit à x à l'instant $t = 0$

Généralement, l'état informationnel I_t regroupe toutes les observations et actions passées, constituant une statistique suffisante pour prendre des décisions. Toutefois, sa taille croît

avec le temps, rendant son exploitation complexe. Pour pallier ce problème, on définit l'état de croyance $b_t(x)$, qui est la distribution de probabilité sur l'ensemble X conditionnée par l'historique et la croyance initiale :

$$b_t(x) = \mathbb{P}[x_t = x | I_t, b_0]$$

Ce qui est essentiel ici, c'est que b_t constitue une statistique suffisante qui contient toute l'information nécessaire pour la prise de décision. Grâce à cette propriété, un problème partiellement observable (POMDP) peut être reformulé en un MDP défini sur l'espace des croyances Cette croyance constitue une statistique suffisante et est mise à jour en appliquant la fonction suivante :

$$\begin{aligned}
b_t(x) &= \mathbb{P}[x_t = x | I_t] \\
&= \mathbb{P}[x_t = x | o_t, a_t, I_{t-1}] \\
&= \frac{\mathbb{P}[o_t | x, a_t, I_{t-1}]\mathbb{P}[x | a_t, I_{t-1}]}{\sum_{x^s}\mathbb{P}[o_t | x^s, a_t, I_{t-1}]\mathbb{P}[x^s | a_t, I_{t-1}]} \\
&= \frac{\mathbb{P}[o_t | x, a_t]\sum_{x^-}\mathbb{P}[x | x^-, a_t, I_{t-1}]\mathbb{P}[x^- | a_t, I_{t-1}]}{\sum_{x^s}\mathbb{P}[o_t | x^s, a_t]\sum_{x^-}\mathbb{P}[x^s | x^-, a_t, I_{t-1}]\mathbb{P}[x^- | a_t, I_{t-1}]} \\
&= \frac{O(x, a_t, o_t)\sum_{x^-} T(x^-, a_t, x) b_{t-1}(x^-)}{\sum_{x^s} O(x^s, a_t, o_t)\sum_{x^-} T(x^-, a_t, x^s) b_{t-1}(x^-)}
\end{aligned} \tag{3.22}$$

L'objectif du POMDP est de maximiser la récompense espérée accumulée par l'agent. L'équation suivante décrit l'équation de Bellman pour la fonction de valeurs optimale V^* des POMDP :

$$V^*(b) = \max_{a \in A}\left[R(b, a) + \gamma \sum_{o \in \Omega}\mathbb{P}[o | a, b] V^*(b')\right] \tag{3.23}$$

La politique optimale qui maximise $V^*(b)$ à tout état de croyance b :

$$\pi^*(b) = \arg\max_{a \in A}\left[R(b, a) + \gamma \sum_{o \in \Omega}\mathbb{P}[o | a, b] V^*(b')\right] \tag{3.24}$$

La fonction $R(b, a)$ est la récompense immédiate attendue de l'exécution de l'action a sous la croyance b :

$$R(b,a) = \sum_{x \in X} b(x) R(x,a) \tag{3.25}$$

La notation $\mathbb{P}[o|a,b]$ représente la probabilité d'obtenir l'observation o lorsque l'agent agit selon l'action a et possède une croyance b :

$$\mathbb{P}[o|a,b] = \sum_{x \in X} O(o,a,x) \sum_{x' \in X} T(x,a,x')b(x) \tag{3.26}$$

La résolution d'un POMDP requiert la recherche d'une politique optimale qui associe les états de croyance aux actions. Cependant, cette résolution est complexe et les solutions exactes sont souvent inaccessibles pour les problèmes de grande taille, d'où le recours à des méthodes d'approximation. Comme pour les MDP, les algorithmes de résolution des POMDP peuvent être classés en deux catégories : la planification en ligne, qui ajuste les décisions en temps réel, et la planification hors ligne, qui optimise la politique avant l'exécution.

3.3.3 Cas à objectifs multiples

Dans ce scénario, l'humain agit en fonction d'un objectif précis que le robot ne perçoit pas parmi un ensemble de buts G. Pour estimer cet objectif, il est nécessaire de maintenir un état informationnel ou, de manière plus efficace, une statistique suffisante pour l'estimation de g. Cette structure correspond au modèle des processus de décision de Markov observable et mixte (MOMDP), une variante des POMDP. Les MOMDP distinguent les variables d'état totalement observables des variables partiellement observables, réduisant ainsi la dimensionnalité de l'espace de croyance en maintenant une croyance seulement pour les variables cachées. Dans ce contexte, les variables observables sont x^H et w, tandis que g reste non observable.

Bien que peu de bibliothèques logicielles offrent une prise en charge directe des MOMDP, il est possible de les traiter comme des POMDP pour utiliser des solutions existantes. Néanmoins, cette approche élimine l'avantage de l'espace de croyance réduit. Cette adaptation implique de considérer toutes les variables comme partiellement observables, avec g influençant uniquement la probabilité de l'observation. On rappelle que l'objectif de l'agent humain demeure constant au fil du temps. Plus formellement $\mathbb{P}[g_{t+1} = g | g_t = g] = 1$. Au final, le POMDP du robot est représenté par le tuple $(X^H, W, A^H, A^R, T_x^H, T_{w'}, R^H, R^R, b_o, \pi^H)$ avec π^H définit à l'équation (3.17). Cette considération garantit que la probabilité de transition de l'objectif demeure constante, ce qui réduit la complexité de modélisation et permet au robot de mieux anticiper les actions de l'humain en se basant sur un but fixe. Il serait intéressant, dans des travaux futurs, de traiter des cas où l'humain a un objectif susceptible de changer

afin de généraliser la solution proposée.

Modèle d'observation

Le modèle d'observation du scénario peut être simplifié en raison de la présence d'états totalement observables, permettant ainsi une modélisation déterministe. En effet, le robot dispose d'une perception parfaite de l'état de l'environnement ainsi que de l'état de l'humain et de ses décisions, ce qui conduit à une observation définie par :

$$\mathbb{P}[o_t|x_t^H, w_t, a_{t-1}^H, a_t^R] = 1 \text{ si } o_t = \{x_t^H, w_t, a_{t-1}^H\} \text{ et } 0 \text{ sinon} \tag{3.27}$$

Représentation de la croyance

La croyance du robot concernant la variable non observable g est directement liée à la reconnaissance d'intention. L'agent doit donc estimer la distribution de probabilité des objectifs possibles de l'humain. Pour cela, il s'appuie sur un état informationnel I (voir équation (3.4)), que l'on réécrit de la manière suivante :

$$I_t = \{I_{t-1}, x_t^H, w_t, a_{t-1}^H, a_{t-1}^R\} \tag{3.28}$$

L'objectif est alors de calculer $b_t(g)$ à partir de cet état informationnel en appliquant la règle de Bayes, comme en (3.22).

$$\begin{aligned} b_t(g) = \mathbb{P}[g|I_t] &= \mathbb{P}[g|I_{t-1}, x_t^H, w_t, a_{t-1}^H, a_{t-1}^R] \\ &= \frac{\mathbb{P}[x_t^H \mid I_{t-1}, a_{t-1}^H, a_{t-1}^R, g]\mathbb{P}[g \mid I_{t-1}, a_{t-1}^H]}{\sum_{g' \in G} \mathbb{P}[x_t^H \mid I_{t-1}, a_{t-1}^H, a_{t-1}^R, g']\mathbb{P}[g' \mid I_{t-1}, a_{t-1}^H]} \\ &= \frac{\mathbb{P}[x_t^H \mid I_{t-1}, a_{t-1}^H, a_{t-1}^R, g]\mathbb{P}[a_{t-1}^H \mid I_{t-1}, g]\mathbb{P}[g \mid I_{t-1}]}{\sum_{g' \in G} \mathbb{P}[x_t^H \mid I_{t-1}, a_{t-1}^H, a_{t-1}^R, g']\mathbb{P}[a_{t-1}^H \mid I_{t-1}, g']\mathbb{P}[g' \mid I_{t-1}]} \\ &= \lambda \mathbb{P}[a_{t-1}^H \mid I_{t-1}, g]\mathbb{P}[g \mid I_{t-1}] \end{aligned} \tag{3.29}$$

De plus, $\mathbb{P}[g \mid I_{t-1}]$ représente la croyance à l'instant précédant $t-1$, λ est un facteur de normalisation, la probabilité de transition $\mathbb{P}[x_t^H \mid I_{t-1}, a_{t-1}^H, a_{t-1}^R, g]$ est déterministe et

$$\mathbb{P}[a_{t-1}^H \mid I_{t-1}, g] = \mathbb{P}[a_{t-1}^H \mid x_{t-1}^H, w_{t-1}^{'}, g]$$

.

La distribution de probabilité des actions possibles de l'humain dépend directement de son objectif. Puisque le robot ne connaît pas cet objectif avec certitude, il est essentiel de pondérer ces actions en fonction de sa croyance quant aux différents objectifs possibles. Cette pondération s'applique également à la fonction de récompense, qui doit refléter cette incertitude. Ainsi, on peut redéfinir la probabilité d'action de l'agent humain $\mathbb{P}_{a^H,w'}[b_t(g)]$ et la fonction de récompense associée $R^H(b_t(g))$, intégrant la distribution de croyance du robot.

$$\mathbb{P}_{a^H,w'}[b_t(g)] = \sum_{g\in G} \mathbb{P}[a^H_{t-1} \mid x^H_{t-1}, w'_{t-1}, g] b_t(g) \tag{3.30}$$

$$R^H(b_t(g), a^H_t) = R^H(x^H_t, w'_t, a^H_t, b_t(g)) = \sum_{g\in G} R^H_t(x^H_t, w'_t, a^H_t, g) b_t(g) \tag{3.31}$$

$$\begin{aligned} R^H(b_t(g)) &= \mathbb{E}_{a^H}[R^H_t \mid b_t(g)] \\ &= \sum_{g\in G} \sum_{a^H\in A^H} \mathbb{P}_{a^H,w'}[b_t(g)] * R^H_t(x^H_t, w'_t, a^H, g) \end{aligned} \tag{3.32}$$

La fonction objective qui régit le comportement du robot dans le scénario de l'objectif incertain parmi plusieurs est la suivante :

$$\arg\max_{a^R\in A_R} \left\{ \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \left(R^H_t(b_t(\mathbf{g})) + R^R_t \right) \middle| x^H_0, w_0, b_0 \right] \right\} \tag{3.33}$$

L'optimisation de cet objectif permet de déterminer la meilleure stratégie que le robot doit adopter face à l'incertitude, qu'il s'agisse de la sous-optimalité des actions humaines ou de l'objectif inconnu de l'humain. La résolution de ce type de problème peut être abordée à l'aide d'algorithmes tels que SARSOP [58], qui approxime la fonction de valeur $V(b)$, ou de méthodes basées sur une discrétisation de l'espace des croyances. Ce dernier reformule le POMDP comme un MDP à espace discret, où l'état discret correspond à l'espace des croyances discrétisé. L'idée d'approximation par un espace discrétisé a été introduite par [59, 60], et des approches plus sophistiquées pour résoudre les MDP continus ont ensuite été proposées par [61, 62].

Sur cette base, nous développons l'équation (3.33) afin d'obtenir l'équation de Bellman suivante :

$$Q_t^R(x_t^H, w_t, a_t^R, b_t(g)) = \Bigg[\sum_{w' \in W} \sum_{a^H \in A_H} \mathbb{P}_{a^H, w_t'} T_{w'}(w_t, a^R, w_t') R^H(b_t(g), a^H) + R_t^R$$

$$+ \gamma \sum_{g \in G} \sum_{x_{t+1}^H \in X_H} \sum_{a^H \in A_H} \mathbb{P}_{a^H, w_t'}[b_t(g)] \times T_x^H(x_t^H, w_t', g, a^H, x_{t+1}^H) V_{t+1}^R(x_{t+1}^H, w_{t+1}, b_{t+1}(g)) \Bigg] \quad (3.34)$$

et

$$a_t^R = \underset{a^R \in A_R}{argmax} \;\; Q^R(x_t^H, w_t, a_t^R, b_t(g)) \quad (3.35)$$

La logique de développement de cette équation reste similaire à celle de l'équation (3.19), qui est détaillée dans l'annexe A. Nous proposons aussi une méthode de discrétisation de l'espace des croyances pour l'algorithme d'itération des valeurs afin de tenir compte de la variable continue $b_t(g) \in [0, 1]$:

Algorithm 6: Itération de valeur basée sur la méthode de discrétisation de l'espace des croyances pour POMDP

Input: Un POMDP $(X^H, W, A^H, A^R, T_x^H, T_{w'}, R^H, R^R, b_0, \pi^H)$, Nombre d'éléments dans l'espace des croyances discrétisé N, facteur d'atténuation γ, un paramètre de précision ϵ

Output: Politique optimale π_R

Diviser l'intervalle des croyances $[0, 1]$ en N éléments pour avoir B Initialiser arbitrairement $V^R(x^H, w, b)$, $\forall x^H \in X^H, w \in W, b \in B$;

Initialiser arbitrairement $Q^R(x^H, w, a^R, b)$, $\forall x^H \in X^H, w \in W, g \in G, a^R \in A^R, b \in B$;

repeat

  for *chaque* $b \in B, w \in W, x^H \in X^H$ **do**

    temp $\leftarrow V^R(x^H, w, b)$

    for *chaque* $a^R \in A^R$ **do**

      for *chaque* $a^H \in A^H$ **do**

        Mettre à jour la croyance b' avec la nouvelle action prise grâce à (3.29)

        Approximer la croyance b' a son élément le plus proche dans l'espace discrétisé : $b' \leftarrow \arg\min_{l \in B} |l - b'|$

        Appliquer l'équation (3.34)

    Appliquer $V^R(x^H, w, b) = \max_{a^R \in A^R} Q^R(x^H, w, a^R, b)$

until $\max_{x^H, w, b} |temp - V^R(x^H, w, b)| \leq \epsilon$;

$\forall x^H, w, b, \pi_R(x^H, w, b) = \arg\max_{a^R \in A^R} Q_t^R(x^H, w, a^R, b)$;

Cette méthode par une discrétisation de l'espace des croyances peut réduire l'efficacité de la

stratégie de décision du robot, car elle approxime l'espace des croyances en le discrétisant. En particulier, à la ligne 9, une approximation est effectuée entre la mise à jour réelle de la croyance et sa valeur discrétisée la plus proche appartenant à l'espace discrétisé. Cela entraîne une perte d'information cruciale, surtout si la discrétisation est insuffisante pour capturer les subtilités de la dynamique des croyances selon les différents comportements de l'humain (voir annexe C pour plus d'information). L'interaction entre le robot et l'humain suit l'algorithme d'interaction 5, avec la particularité que l'action a_t^R est sélectionnée selon l'équation (3.35). De plus, la croyance est mise à jour conformément à l'équation (3.29) avant de passer à l'instant suivant. Des expérimentations sur cet aspect sont présentées dans le chapitre suivant, à la section 4.3.3.

3.4 Méthode d'apprentissage de l'indice de rationalité

Le processus de décision du robot prend en compte la politique d'action de l'humain, modélisée par l'équation (3.17) avec un indice de rationalité fixé à une valeur β_0, supposée correcte dans les sections précédentes. Cependant, si ce facteur est incorrect ou repose sur une hypothèse inexacte, cela peut entraîner des actions inadaptées du robot, réduisant l'efficacité de la collaboration. Cette section introduit une méthode d'apprentissage de ce facteur pour améliorer la phase de préparation du robot (voir figure 4.8) et optimiser ses interactions.

L'objectif est d'estimer le paramètre $\beta_{réel}$ qui dicte le comportement réel de l'humain, en observant les actions et les états de ce dernier durant une interaction. Cette estimation est faite en utilisant une méthode de maximisation de vraisemblance qui dépend d'un ensemble de K trajectoires. Pour ce faire, on définit une trajectoire τ, comme l'ensemble des observations du robot de l'instant 1 jusqu'à l'instant final T, soit $\tau = I_T$ (voir l'équation (3.4)). Cette trajectoire est obtenue à l'issue de l'exécution de l'algorithme 5, en considérant que le robot agit avec une politique qui se base sur un indice de rationalité β_0. Bien que le robot ne puisse pas observer directement l'objectif de l'humain, il peut néanmoins estimer une croyance à son sujet, comme discuté en section 3.3.3. L'apprentissage de l'indice de rationalité est d'abord effectué en supposant l'objectif connu, avant d'être généralisé au cas où celui-ci est incertain.

Connaissance de l'objectif de l'humain

On appelle $\Omega = \{\tau_1, \dots \tau_K\}$ l'ensemble de K trajectoires. Par indépendance conditionnelle des trajectoires τ_i et des variables contenues dans cette trajectoire, la probabilité que Ω soit effectuée, sachant que l'agent humain suit une politique softmax inconnue β, est :

$$\mathbb{P}[\Omega|\beta] = \prod_{i=1}^{K} \mathbb{P}[\tau_i|\beta] = \prod_{i=1}^{K} \prod_{t_i=1}^{T_i} \mathbb{P}[x_{t_i+1}^H | x_{t_i}^H, w_{t_i}^{'}.a_{t_i}^H] \mathbb{P}[a_{t_i}^H \mid x_{t_i}^H, w_{t_i}^{'}, g; \beta]. \tag{3.36}$$

Dans la sous-section 3.1.4, la fonction de transition de l'état de l'humain est définie comme déterministe, ce qui signifie que la probabilité de suivre une trajectoire τ dépend uniquement du modèle d'action de l'humain et de son objectif g.

$$\begin{aligned} \mathbb{P}[\Omega|\beta] &= \prod_{i=1}^{K} \prod_{t_i=1}^{T_i} \mathbb{P}[a_{t_i}^H \mid x_{t_i}^H, w_{t_i}^{'}, g; \beta] \\ &= \prod_{i=1}^{K} \prod_{t_i=1}^{T_i} \frac{e^{\beta Q_{t_i}^H(x_{t_i}^H, w_{t_i}^{'}, a_{t_i}^H, g)}}{\sum\limits_{a_{t_i}^H \in A^H} e^{\beta Q_{t_i}^H(x_{t_i}^H, w_{t_i}^{'}, a_{t_i}^H, g)}}. \end{aligned} \tag{3.37}$$

Plusieurs algorithmes ont été développés pour estimer des paramètres de distributions de probabilité, notamment l'algorithme de maximisation d'espérance (*Expectation Maximization*, EM), qui est une méthode itérative pour l'estimation du maximum de vraisemblance d'un paramètre, comme décrit par Dempster et al. [63]. Aussi, l'algorithme de montée de gradient, quant à lui, vise à minimiser une fonction de coût de manière itérative, ce qui est bien expliqué par Boyd et Vandenberghe [64] et [65]. L'algorithme de montée de gradient est choisi ici pour son efficacité et sa simplicité d'implémentation dans l'estimation de paramètres avec des données incomplètes.

La première étape de l'algorithme est de déterminer la fonction de coût qui représente, dans notre cas la fonction de log-vraisemblance (voir l'équation (3.37)) en tenant compte des données observées et de l'estimation actuelle des paramètres. On obtient alors :

$$\begin{aligned} \ln \mathbb{P}[\Omega|\beta] &= \sum_{i=1}^{K} \ln \mathbb{P}[\tau_i|\beta] \\ &= \sum_{i=1}^{K} \ln \prod_{t_i=1}^{T_i} \mathbb{P}[a_{t_i}^H \mid x_{t_i}^H, w_{t_i}^{'}, g; \beta] \\ &= \sum_{i=1}^{K} \left[\sum_{t_i=1}^{T_i} \beta Q_{t_i}^H(x_{t_i}^H, w_{t_i}^{'}, a_{t_i}^H, g) - \ln \sum_{a_{t_i}^H \in A^H} e^{\beta Q_{t_i}^H(x_{t_i}^H, w_{t_i}^{'}, a_{t_i}^H, g)} \right] \end{aligned} \tag{3.38}$$

Cette fonction de coût est choisie car elle a des propriétés essentielles dans la convergence de

l'algorithme de montée de gradient (voir proposition 1).

Proposition 1. *La fonction de log-vraisemblance* (3.38) *est concave et l'algorithme de montée de gradient converge pour tout pas fixe* $\alpha < \frac{1}{\beta}$ *vers une valeur* $\beta^*_{\tau_{1:K}}$.

En calculant le gradient de (3.38) par rapport à β on obtient :

$$\nabla_\beta\{\ln \mathbb{P}[\Omega|\beta]\} = \sum_{i=1}^{K}\left[\sum_{t_i=1}^{T_i} Q^H_{t_i}(a^H_{t_i}, g) - \eta(\beta)\sum_{a^H\in A^H} Q^H_{t_i}(a^H, g)e^{\beta Q^H_{t_i}(a^H,g)}\right] \tag{3.39}$$

Avec :

$$\eta(\beta) = \frac{1}{\sum\limits_{a^H\in A^H} e^{\beta Q^H_{t_i}(x^H_{t_i}, w'_{t_i}, a^H, g)}}$$

$$Q^H_{t_i}(a^H, g) = Q^H_{t_i}(x^H_{t_i}, w'_{t_i}, a^H, g)$$

On normalise le gradient afin d'obtenir :

$$\nabla_\beta\{\ln \mathbb{P}[\Omega|\beta]\} = \frac{1}{K}\sum_{i=1}^{K}\left[\sum_{t_i=1}^{T_i} Q^H_{t_i}(a^H_{t_i}, g) - \eta(\beta)\sum_{a^H\in A^H} Q^H_{t_i}(a^H, g)e^{\beta Q^H_{t_i}(a^H,g)}\right] \tag{3.40}$$

Preuve 1 (Concavité de la fonction de log-vraisemblance et convergence de la méthode d'apprentissage)**.** *Soit la fonction de log-vraisemblance définie par*

$$\sum_{i=1}^{K}\left[\sum_{t_i=1}^{T_i}\beta Q_{t_i}^{H}(x_{t_i}^{H},w_{t_i}^{'},a_{t_i}^{H},g)-\ln\sum_{a_{t_i}^{H}\in A^{H}}e^{\beta Q_{t_i}^{H}(x_{t_i}^{H},w_{t_i}^{'},a_{t_i}^{H},g)}\right]$$

D'après l'exemple 3.14 de [64], la fonction

$$f(\beta)=\log\left(\sum_{a_{t_i}^{H}\in A^{H}}\exp\Big(\beta\, Q_{t_i}^{H}(x_{t_i}^{H},w_{t_i}^{'},a_{t_i}^{H},g)\Big)\right)$$

est convexe, car pour tout $a^H \in A^H$, la fonction $g_{a^H}(\beta)=\beta\, Q_{t_i}^{H}(x_{t_i}^{H},w_{t_i}^{'},a_{t_i}^{H},g)$ est linéaire et donc convexe. Il en résulte que $-f(\beta)$ est concave.

Par ailleurs, le premier terme de $\ln\mathbb{P}[\tau_i|\beta]$ étant linéaire est donc convexe et on rappelle que la somme de fonctions concave reste concave. La fonction de log-vraisemblance est alors globalement concave par rapport à β. Cette concavité, combinée à la propriété β-Lipschitz de la fonction softmax de l'équation (3.39) *(voir le corollaire 3 dans [66]), garantit la convergence de l'algorithme de montée de gradient pour un pas $\alpha \leq \frac{1}{\beta}$ vers une estimée $\beta^*_{\tau_{1:K}}$.*

Le gradient étant simple à calculer, une implémentation sans l'utilisation de bibliothèque logicielle spécialisée est appropriée. Toutefois, dans des cas plus complexes où le gradient analytique serait difficile à obtenir, des bibliothèques intégrant des méthodes de différentiation automatique, telles que `PyTorch`, peuvent être employées.

Incertitude sur l'objectif de l'humain

L'objectif de l'humain n'étant pas accessible au robot, les équations développées précédemment ne peuvent pas être appliquées directement. On considère ici que le robot conserve à chaque instant de la trajectoire la croyance courante sur l'objectif de l'humain $b_t(g)$. La trajectoire τ contient donc la distribution de probabilité sur tous les objectifs à chaque instant t. En ajoutant une pondération suivant ses croyances on obtient :

$$\mathbb{E}_g[\nabla_\beta\{\ln \mathbb{P}[\tau|\beta]\}] = \mathbb{E}_g \sum_{i=1}^{K} \left[\sum_{t_i=1}^{T_i} \beta Q_{t_i}^H(x_{t_i}^H, w_{t_i}^{'}, a_{t_i}^H, g) - \ln \sum_{a_{t_i}^H \in A^H} e^{\beta Q_{t_i}^H(x_{t_i}^H, w_{t_i}^{'}, a_{t_i}^H, g)} \right]$$

$$= \sum_{g \in G} \sum_{i=1}^{K} \left[\sum_{t=1}^{T_i} b_{t_i}(g) \Big[Q_{t_i}^H(a_{t_i}^H, g) - \eta(\beta) \sum_{a^H \in A^H} Q_{t_i}^H(a^H, g) e^{\beta Q_{t_i}^H(a^H, g)} \Big] \right] \quad (3.41)$$

L'algorithme permettant d'estimer le paramètre β est présenté ci-dessous :

Algorithm 7: Algorithme de montée de gradient pour estimation de β

Input: Trajectoires $\Omega = \{\tau_1, \tau_2, ..., \tau_K\}$, valeur initiale de β_0, un paramètre de précision ϵ, le pas d'apprentissage α

Output: Estimée β_i

```
Initialiser β_i = β_0
repeat
    temp ← β_i
    Calculer le gradient de l'équation (3.41)
    Mettre à jour l'estimation :
    β_i ← β_i + α * (E_g[∇_β{ln P[τ|β]}]/|Ω|)
until |temp − β_i| ≤ ε;
```

Ce mémoire n'aborde pas l'analyse des conditions sur les trajectoires ni les détails théoriques assurant la convergence de $\beta^*_{\tau_{1:K}}$ vers $\beta_{réel}$, ce qui constitue une perspective pour des travaux futurs. Néanmoins, la question de la convergence de la méthode vers $\beta_{réel}$ est examinée empiriquement à travers des simulations dans le chapitre 4.2.2.

CHAPITRE 4 SIMULATIONS ET RÉSULTATS

Ce chapitre a pour but de présenter les simulations et les résultats obtenus de l'implémentation des algorithmes décrits dans le chapitre 3. De plus, les librairies nécessaires aux implémentations ainsi que l'environnement de simulation seront abordés. Enfin, chaque résultat permet d'étoffer la discussion en fin de ce chapitre sur l'efficacité d'un modèle de premier ordre dans la collaboration d'un robot avec un agent humain.

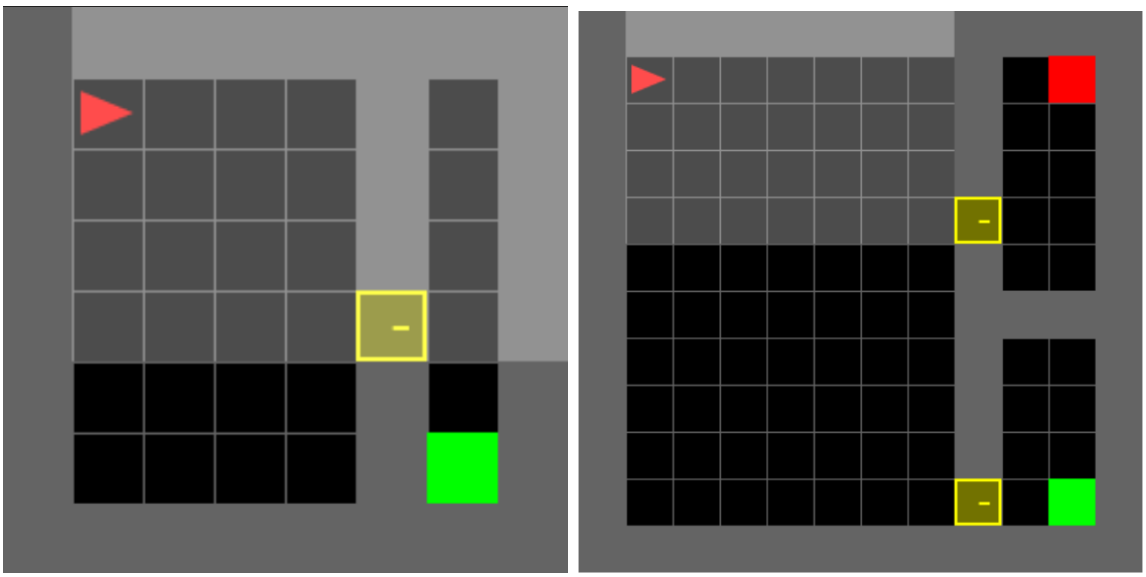
4.1 Présentation des scénarios

Pour tester notre approche, plusieurs scénarios ont été conçus, comme illustré à la figure 4.1.

- **Premier scénario :** L'algorithme d'interaction 5 est évalué dans un contexte simple où un agent humain doit se diriger vers un objectif défini sur une grille. Cependant, un obstacle empêche l'humain d'atteindre sa destination. Cet obstacle, modélisé comme une porte dont l'humain ne possède pas la clé, est impossible à déplacer par lui-même. Le rôle du robot est alors d'observer et d'inférer qu'il doit ouvrir cette porte pour permettre à l'humain de poursuivre son chemin.
- **Deuxième scénario :** Une complexité supplémentaire est introduite avec l'ajout d'un second objectif également bloqué par une porte. Le robot doit alors raisonner sur l'objectif de l'humain pour décider quelle porte ouvrir en fonction du comportement observé.
- **Troisième scénario :** Il reprend la structure du second, mais avec une contrainte supplémentaire : le robot ne peut maintenir qu'une porte ouverte à la fois. Le robot est placé dans une situation d'impasse. Cela force une coopération plus étroite entre les deux agents afin d'atteindre l'objectif, mettant en avant la nécessité d'une prise de décision plus stratégique.

Pour les simulations, le choix se porte sur l'environnement `Minigrid` [67], offrant des mondes en grille modulaires. Ces environnements, adaptés aux besoins de l'étude, font l'objet de modifications spécifiques pour s'aligner sur les scénarios expérimentaux illustrés par la figure 4.1. L'objectif de l'humain est d'atteindre un des objectifs possibles, tandis que le robot doit ouvrir la bonne porte pour faciliter cet objectif. Le modèle s'articule autour des éléments suivants :

- **États :** Les états comprennent la position de l'humain à l'instant t, notée $x_t^H \in X^H$, représentée par une coordonnée (x, y), l'objectif de l'humain g, défini par une couleur associée à une case de coordonnées (x_g, y_g), ainsi que l'état des portes, $w_t \in W$, qui



(a) Scénario 1 (b) Scénario 2

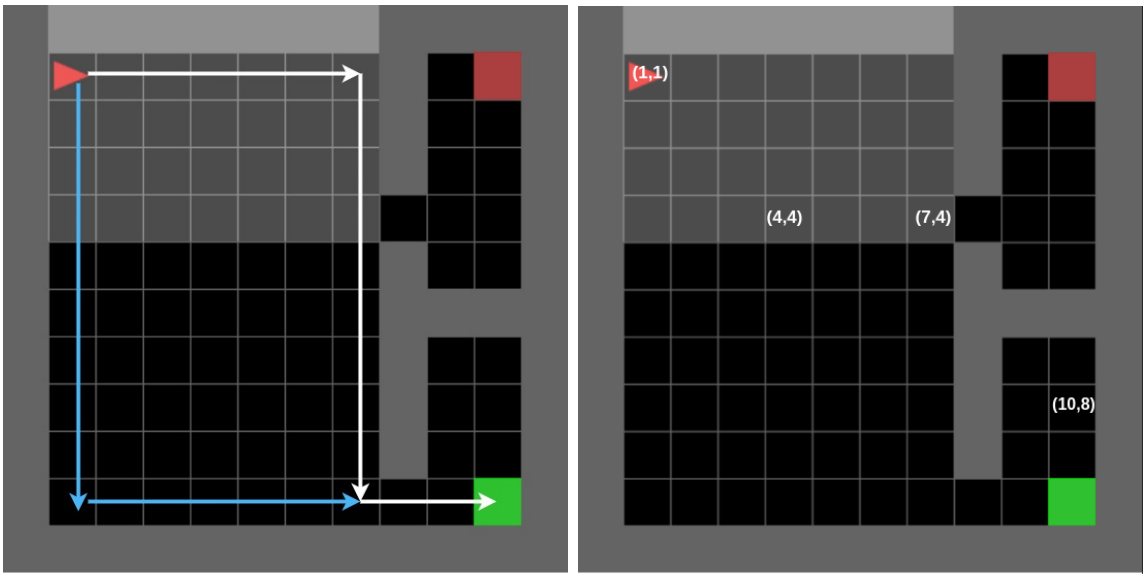
FIGURE 4.1 (a) : Environnement où l'humain (triangle rouge) poursuit un objectif unique. L'agent en déplacement représente l'humain, les cases grises symbolisent les murs et la case jaune représente une porte fermée (grille de taille 8×8). (b) Scénario où l'humain a deux objectifs possibles (en rouge et en vert), tous deux bloqués par des portes (grille de taille 12×12).

peut être "fermée" ou "ouverte" selon le scénario. L'état informationnel du robot est défini par l'équation (3.4).

- **Observation du robot :** Le robot ne peut pas observer directement l'objectif de l'humain, ce qui met en lumière la problématique de reconnaissance d'intention. Toutefois, le robot observe parfaitement l'état actuel de l'humain (position) et le statut des portes (ouvertes ou fermées). L'incertitude provient donc exclusivement de l'objectif de l'humain, inconnu pour le robot.
- **Actions :** L'humain peut effectuer cinq actions : se déplacer vers le haut, le bas, la gauche, la droite ou rester immobile ($A^H = 5$). De son côté, le robot peut ouvrir une porte ou ne rien faire à chaque instant ($|A^R| = 2$ pour le scénario 1 et $|A^R| = 3$ pour le scénario 2). La transition d'un état à un autre en réponse à une action de l'humain ou du robot est déterministe.
- **Fonction de récompense :** L'humain reçoit une récompense de 300 lorsqu'il atteint son objectif. Chaque action de l'humain, sauf l'action attendre, entraîne une pénalité de -1. L'action d'attendre ne génère aucune pénalité. Quant au robot, chaque ouverture de porte lui impose une pénalité de -10.

Pour capturer les comportements spécifiques de l'humain, la fonction de récompense introduit

une préférence claire pour l'objectif de l'humain (récompense de 300 pour un objectif), les autres objectifs étant ignorés (récompense de 0). À l'aide de l'algorithme 2, $|G|$ politiques sont calculées. Un exemple de trajectoire d'un humain lorsque la porte vers son objectif est ouverte est illustré à la figure 4.2a.



(a) (b)

FIGURE 4.2 (a) : Deux trajectoires possibles de l'humain pour atteindre la destination en vert. Les trajectoires en bleu et en blanc sont optimales. (b) : Les numéros sur la deuxième figure indiquent les coordonnées.

Un point important à observer est que la trajectoire optimale n'est pas unique, car le choix entre plusieurs chemins n'influence pas la récompense finale obtenue par l'humain. Pour modéliser le comportement de l'humain, une politique softmax (équation (3.17)) est utilisée plutôt qu'une politique déterministe. Les actions de l'humain suivent alors une distribution de probabilité dépendante de sa position actuelle, de l'état des portes et de son objectif.

Les figures 4.3 et 4.4 illustrent ces comportements pour un humain avec un indice de rationalité ($\beta = 1$). À la position $(1, 1)$, aucune préférence marquée n'est observée entre les objectifs possibles : les actions 'Droite' et 'Bas' présentent des probabilités équivalentes, tandis que les autres actions restent négligeables. En revanche, à la position $(7, 4)$, une préférence se dessine nettement : l'action 'Droite' rapproche de l'objectif en rouge, tandis que l'action 'Bas' oriente vers un autre objectif. Une observation similaire s'applique à la position $(10, 8)$.

Une particularité apparaît à la position $(4, 4)$. Du point de vue de l'objectif en vert, cette position ne joue pas un rôle décisif dans le choix entre aller à gauche ou à droite.

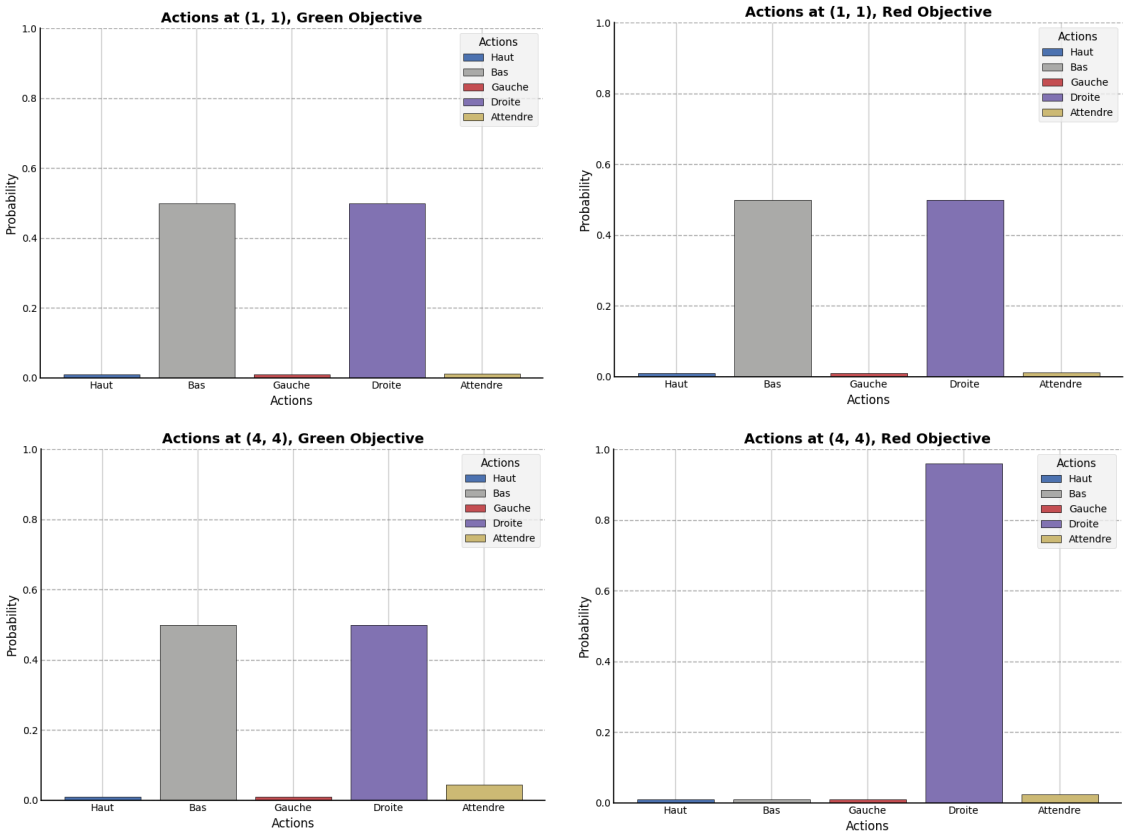


FIGURE 4.3 Distribution de probabilité des différentes actions à différentes positions, $(1, 1)$ et $(4, 4)$ de la figure 4.2b, pour les deux différents objectifs du scénario 2. Les probabilités sont déterminées grâce à (3.16) avec un indice de rationalité $\beta = 1$.

Pour un humain, avec un indice de rationalité $\beta = 0.25$, la figure 4.5 illustre une propension accrue à des décisions moins rationnelles. En particulier, à la position $(7, 4)$, les actions optimales consistent à se déplacer vers le bas pour atteindre l'objectif en vert (voir figure 4.2a). Cependant, avec $\beta = 0.25$, l'humain adopte une action sous-optimale dans 65 % des cas (voir 4.5a) contrairement au comportement avec un indice de rationalité un plus élevé (voir 4.5b) qui effectue des erreurs 40 % des cas ($\beta = 0.8$).

Cette rationalité réduite accentue l'incertitude sur l'objectif poursuivi, influençant directement les décisions du robot et les rendant plus imprévisibles. Ainsi, si le robot élabore sa stratégie en se fondant sur un indice de rationalité bas pour le comportement de l'humain, il risque de prendre des décisions sous-optimales, affectant négativement les performances globales de l'équipe.

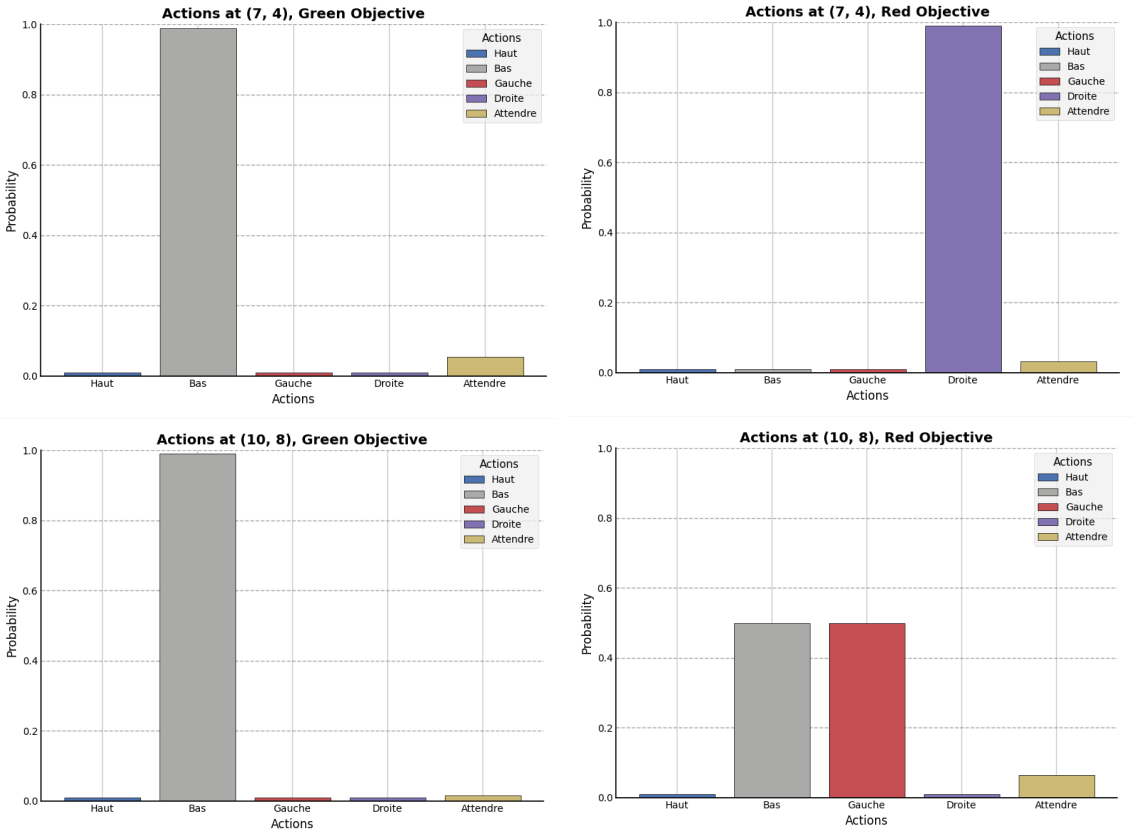
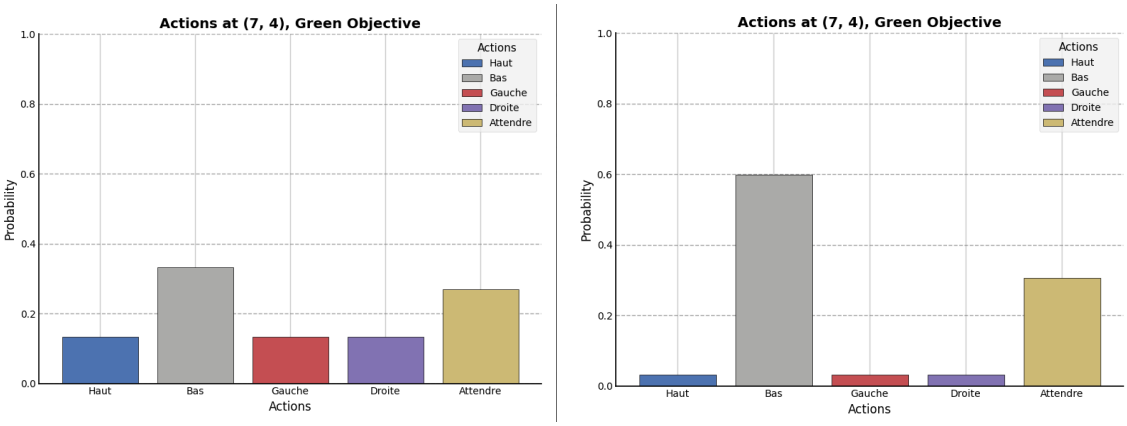


FIGURE 4.4 Distribution de probabilité des différentes actions à différentes positions, $(7, 4)$ et $(10, 8)$ de la figure 4.2b, pour les deux différents objectifs du scénario 2. Les probabilités sont déterminées grâce à (3.16) avec un indice de rationalité $\beta = 1$.

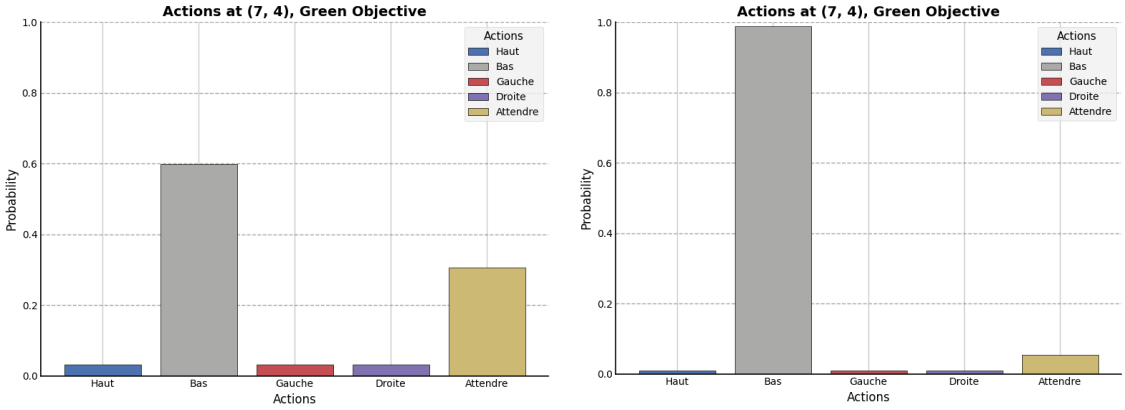


(a) $\beta = 0.25$ (b) $\beta = 0.8$

FIGURE 4.5 Distribution de probabilité des différentes actions à différentes positions, $(7, 4)$ de la figure 4.2b, pour différents β. Les probabilités sont déterminées grâce à (3.16).

D'un autre côté, la figure 4.6 illustre le comportement d'un agent avec un indice de rationalité

de $\beta = 0.8$ face $\beta = 1$. L'humain présente une probabilité de 40 % de choisir une action sous-optimale (voir 4.6a), contre 9 % pour le modèle de la figure (voir 4.6b).



(a) $\beta = 0.8$ (b) $\beta = 1$

FIGURE 4.6 Distribution de probabilité des différentes actions à différentes positions, $(7, 4)$ de la figure 4.2b, pour différents β. Les probabilités sont déterminées grâce à (3.16).

Avec $\beta = 1$, les actions sous-optimales sont plus rares, ce qui limite la distinction de certains cas particuliers face à une politique optimale, surtout sur un nombre limité de simulations. Le modèle avec $\beta = 0.8$ est utilisé pour les simulations afin de définir la politique du robot, conformément à l'hypothèse 3.

4.2 Détails d'implémentation

L'implémentation des méthodes décrites s'effectue en `Python`, avec la bibliothèque `NumPy` pour les opérations numériques.

4.2.1 Analyse de complexité et temps de calcul des politiques

Cette section examine la complexité des algorithmes implémentés et leur applicabilité à des problèmes de grande échelle. Les principaux paramètres influençant le temps de calcul de la politique optimale sont l'espace d'état de l'humain X^H, l'espace des états de l'environnement W, l'espace des actions des deux agents A^H et A^R avec $|A^H| = 5$ et $|A^R| = 3$ (dans le scénario 2), l'espace des états d'objectifs de l'humain G et l'espace des croyances B dans un POMDP.

La taille de X^H dépend de la grille avec

$$|X^H| = m_x m_y,$$

où m_x et m_y représentent les dimensions en longueur et largeur. Si l'environnement contient n_p portes avec deux statuts possibles (ouverte ou fermée), alors :

$$|W| = 2^{n_p},$$

Dans un POMDP, la discrétisation de l'espace des croyances (section 3.3.3) implique que, pour $|G|$ objectifs possibles, cet espace a une taille de $N^{|G|-1}$, avec N représentant le nombre d'éléments dans l'espace discrétisé de la croyance. En résumé, l'espace des états croît linéairement avec le nombre total de cases et exponentiellement avec le nombre de portes et d'objectifs.

Le temps de calcul de la politique de l'humain pour l'algorithme 4 est de l'ordre de :

$$O(|X^H| \times |W| \times |A^H| \times |G|)$$

et celui de la politique du robot pour l'algorithme 6 est :

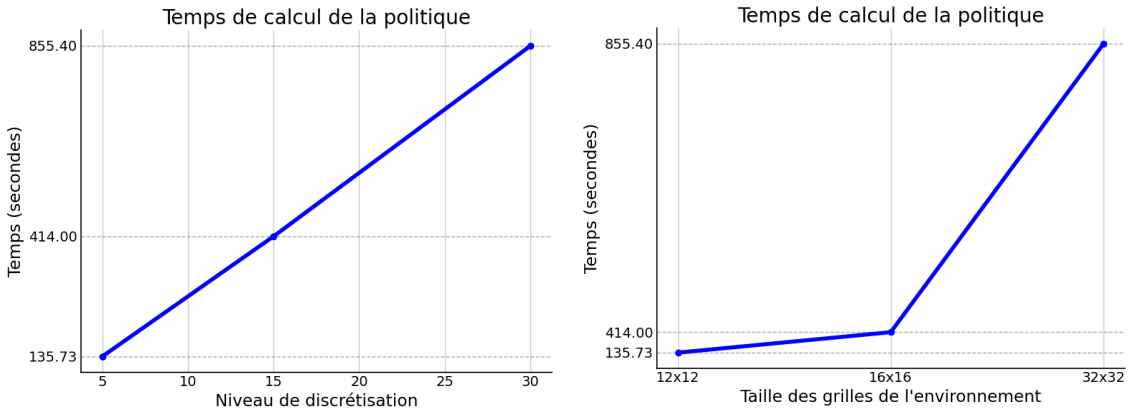
$$O(|X^H| \times |W| \times |A^H| \times |A^H| \times N^{|G|-1}) \tag{4.1}$$

Les paramètres utilisés pour les simulations sont présentés dans le tableau suivant :

Espace d'état de l'humain (X^H)	$(12 \times 12), (16 \times 16), (32 \times 32)$
Nombre de portes (W)	$\{1, 2\}$
Nombre d'objectifs possibles (G)	$\{1, 2\}$
Indice de rationalité β	$\{0.1, 0.8, 2\}$
Discrétisation de l'espace des croyances N	$\{5, 15, 30\}$

Tableau 4.1 Paramètres de simulations

Étant donné que la politique du robot repose sur une discrétisation de l'espace de croyance, il est essentiel de choisir les paramètres clés pour les simulations des sous-sections suivantes, compte tenu des contraintes de temps. La figure 4.7 montre le temps de calcul de la politique du robot déterminée de manière empirique en fonction de la discrétisation et de la taille de l'environnement.



(a) (b)

FIGURE 4.7 (a) : Temps de calcul de la politique du robot (POMDP) pour différentes discrétisation $N = \{5, 15, 30\}$, dans une grille de taille 12×12 et $n_p = 2$ (b) : Temps de calcul de la politique du robot (POMDP) pour différentes tailles de grilles $(12 \times 12), (16 \times 16), (32 \times 32)$

L'analyse révèle que le temps de calcul de la politique croît proportionnellement à l'augmentation de l'espace des états possibles de l'humain ou du paramètre N. Pour des raisons de temps de calcul, l'étude se limite à une variation de N, en fixant l'espace des états de l'humain à une grille de dimensions (12×12). Le paramètre N est susceptible d'influencer la qualité des actions sélectionnées par le robot. L'impact de ce paramètre fait l'objet d'une évaluation détaillée dans la sous-section 4.3.3.

4.2.2 Résultats d'apprentissage de l'indice de rationalité

La méthode d'apprentissage du paramètre β, définie par l'algorithme 7 présenté précédemment, vise à corriger les erreurs de calcul de stratégie du robot en améliorant son modèle de l'humain à partir de ses observations. Pour évaluer ce modèle, K trajectoires sont d'abord collectées trajectoires (avec $K = \{5, 10, 25\}$), dans lesquelles l'humain agit avec un indice de rationalité $\beta_{réel} = 0.1$ en poursuivant son objectif. Toutes les interactions entre l'agent et l'environnement sont enregistrées (algorithme 5) lors de cette phase et l'algorithme de montée de gradient est appliqué pour estimer β. Ceci peut s'illustrer à travers la figure suivante :

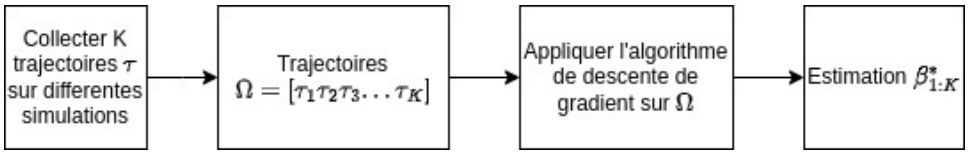
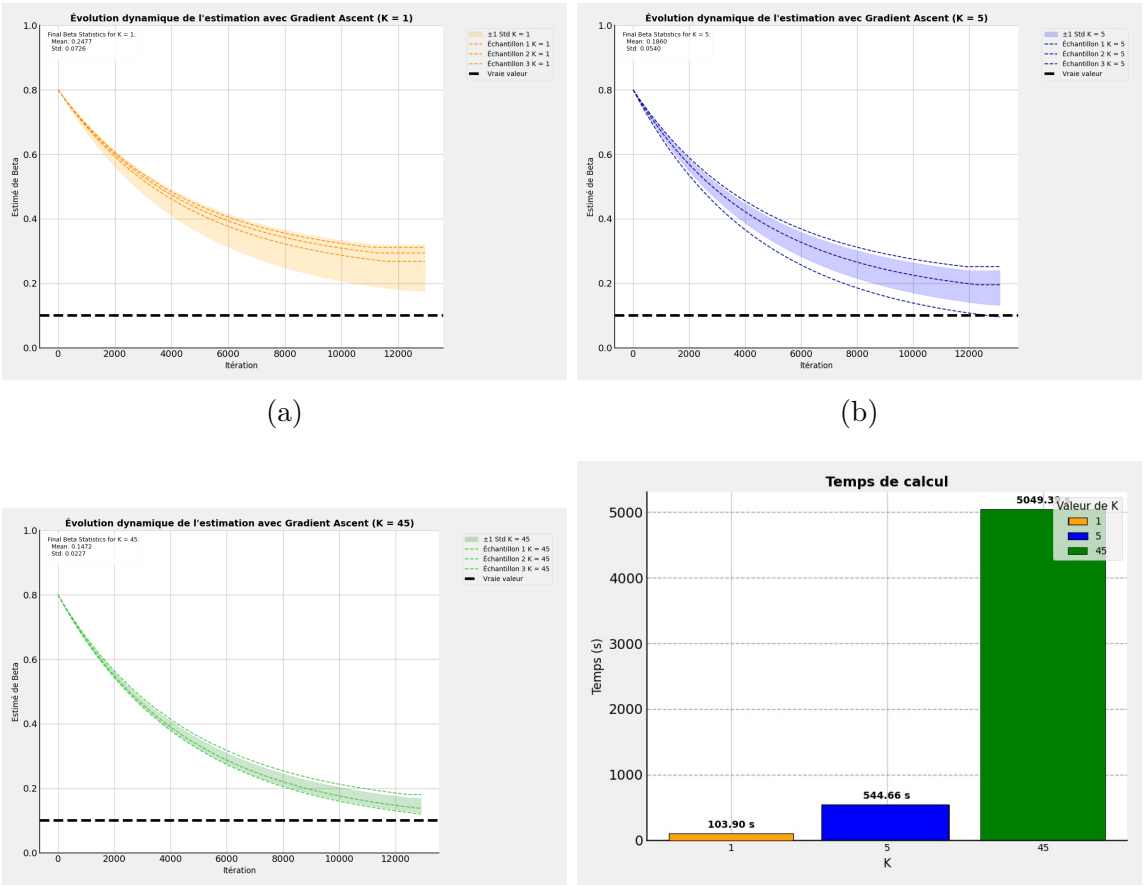


FIGURE 4.8 Étape de collecte des données et estimation hors ligne du paramètre β

Nous effectuons 100 simulations Monte-Carlo hors ligne sur ces données enregistrées pour obtenir une moyenne et variance de la performance de l'estimation. Pour chaque K, nous calculons la trajectoire moyenne des estimations de β (avec l'écart-type correspondant) à chaque itération et déterminons le temps moyen de convergence (moyenne des temps d'exécution sur 100 simulations indépendantes). Ces résultats, présentés respectivement dans les figures 4.9a, 4.9b, 4.9c et 4.9d, permettent d'évaluer l'impact de K sur la précision et l'efficacité de l'estimation du paramètre.



(a) (b)

(c) (d)

FIGURE 4.9 Évolution de l'apprentissage du paramètre β (valeur réelle ($\beta_{réel}$) : 0.1) pour différents nombres de trajectoires collectées $K = \{1, 5, 45\}$. Les zones ombrées indiquent la variabilité des estimations sur 100 simulations. Les courbes en trait discontinues montrent l'évolution de l'estimation dans quelques simulations Monte-Carlo. La figure 4.9d présente le temps d'apprentissage moyen selon la valeur de K.

L'algorithme utilisé repose sur l'équation (3.41), qui dépend des ensembles de trajectoires observées et peut varier en fonction des actions de l'humain. Dans cette étude, le pas d'apprentissage α est fixé à 10^{-3}. Il est observé que l'algorithme converge pour $K = 5$ trajectoires

collectées à une meilleure estimation ($\bar{\beta^*_{1:5}} = 0.1860$) de $\beta_{réel}$ qu'avec $K = 1$ ($\bar{\beta^*_1} = 0.2477$), et une tendance similaire est constatée avec $K = 45$ ($\bar{\beta^*_{1:45}} = 0.1472$). Cette expérimentation empirique confirme que la méthode offre de meilleures performances avec un nombre élevé de trajectoires. Toutefois, une analyse plus approfondie demeure nécessaire et les résultats obtenus constituent une piste pour des travaux futurs.

Bien que l'algorithme proposé semble converger vers $\beta_{réel}$, il nécessite un nombre accru de données pour réduire la variance des estimations. L'analyse des temps de simulations montre que lorsque le nombre de trajectoires est limitée, l'erreur de l'estimation est grande mais avec l'avantage que cela prends moins de temps de calcul. À l'inverse, un nombre élevé de trajectoires augmente le temps de calcul tout en diminuant cette erreur.

Dans le contexte de tâches où le temps de réponse du robot est critique, comme les missions de recherche et de secours, il est essentiel d'équilibrer vitesse d'apprentissage et incertitude sur la valeur estimée (voir figure 4.9d). Pour la suite des simulations, $K = 5$ est retenu.

4.3 Détails des simulations

4.3.1 Spécificités des politiques

Pour le scénario 2 (voir figure 4.1b), les politiques du robot sont calculées en utilisant la méthode basée sur la discrétisation de l'espace des croyances, pour différents comportements de l'humain. Comme rappel, le comportement humain est modélisé à l'aide de l'indice de rationalité β de la politique de Boltzmann. Toutes les simulations débutent avec les états initiaux illustrés à la figure 4.1 avec une taille de grille fixe de 12×12. Sauf mention contraire, la croyance initiale du robot est uniformément répartie sur chaque objectif, soit $\frac{1}{|G|}$.

Quatre politiques ont été calculées :

- Trois politiques provenant de la méthode de discrétisation de l'espace des croyances pour trois niveaux de discrétisation $N = \{5, 15, 30\}$.
- La politique oracle, obtenue en supposant une connaissance parfaite de l'objectif de l'humain.

L'objectif est de comparer les performances de ces politiques par rapport à l'oracle. Les politiques précédentes sont calculées pour indice de rationalité $\beta_R = 0.8$, ce qui signifie que le robot considère l'humain comme ayant une rationalité limitée (hypothèse 3). Pour tester ces politiques, trois comportements distincts de l'humain (voir annexe B pour plus d'information sur les comportements) sont simulés grâce à l'algorithme 2, chacun visant un objectif unique qui ne change pas :

- Comportement quasi aléatoire ($\beta = 0.1$),
- Comportement semi-rationnel ($\beta = 0.8$),
- Comportement quasi rationnel ($\beta = 2$).

Chaque combinaison politique-comportement est testée sur un total de 1000 simulations. De plus, on incorpore l'apprentissage du paramètre β afin d'étoffer notre discussion sur l'importance de compenser les erreurs de modélisation, comme le fait que le robot élabore sa stratégie en considérant toujours $\beta_R = 0.8$.

4.3.2 Résultats et discussion du scénario 1

Le premier scénario évalue la performance d'une équipe humain-machine lorsque l'objectif de l'humain est connu. Ce scénario ne présente que les résultats de l'application des politiques du robot pour différents comportements de l'humain. L'enjeu principal est simplement de vérifier si le robot applique bien sa politique d'assistance en ouvrant la porte. La politique d'assistance est calculée à l'aide de l'algorithme 5. Lorsqu'il est appliqué à ce scénario, le robot ouvre immédiatement la porte afin de permettre à l'humain d'atteindre son objectif, ce qui se traduit par des récompenses cumulatives actualisées positives et un taux de réussite dans l'atteinte de l'objectif de 100 % (voir tableau 4.2). En l'absence de cette intervention, l'humain resterait immobile, n'encourant ainsi aucun coût. Cette observation est valable pour un modèle d'humain rationnel. Toutefois, avec un modèle semi-rationnel, l'humain présente une probabilité non nulle d'adopter des actions sous-optimales (voir figure 4.6a), ce qui l'amène à accumuler des coûts de déplacement inutiles. Cette remarque est aussi observable sur les résultats du scénario 2.

Quel que soit le modèle de décision de l'humain, l'identification de son objectif, qui est unique dans ce scénario, n'a pas d'impact sur la politique adoptée par le robot. Cependant, une observation inattendue mérite d'être soulignée : la récompense cumulée actualisée moyenne d'un humain quasi rationnel, caractérisé par un paramètre $\beta = 2$, se révèle inférieure à celle d'un humain semi-rationnel, avec un $\beta = 0.8$ ((voir tableau 4.2)). Cette différence s'explique par la façon dont le robot interprète la rationalité de l'humain et par les effets de cette perception sur la récompense attendue. En effet, lorsqu'un humain quasi rationnel est perçu par le robot comme semi-rationnel, avec un $\beta_R = 0.8$, le robot sous-estime la probabilité que cet humain atteigne son objectif, ce qui diminue la récompense espérée. À l'inverse, quand l'humain est réellement semi-rationnel et que le robot le reconnaît correctement avec un $\beta_R = 0.8$, cet humain a tendance à effectuer des actions sous-optimales, comme le montre la figure 4.6. Paradoxalement, ces imperfections jouent en sa faveur : elles permettent au robot d'anticiper une variété de récompenses potentielles, ce qui augmente la récompense cumulée

actualisée moyenne (ceci explique les grosses récompenses pour $\beta = 0.1$). Ainsi, bien que la stratégie du robot reste constante, ses attentes évoluent en fonction de sa perception de la rationalité humaine.

Modèle humain	**Récompense actualisée de l'humain**	**Réussite (%)**
$\beta = 0.1$	258 ± 369	100
$\beta = 0.8$	217 ± 88	100
$\beta = 2$	190 ± 20	100

TABLEAU 4.2 Récompense cumulative actualisée pour trois modèles différents de l'humain. Aucune estimation de la valeur réelle du paramètre β n'est effectuée, le robot raisonne avec $\beta_R = 0.8$.

D'un autre côté, l'absence d'ambiguïté sur l'objectif de l'humain élimine toute incertitude, facilitant ainsi une prise de décision optimale pour le robot. Bien que cette approche présente l'avantage d'une faible complexité computationnelle, elle demeure inadaptée aux scénarios plus complexes où l'ambiguïté sur l'objectif humain joue un rôle crucial, rendant chaque action de l'humain déterminante dans la prise de décision du robot.

4.3.3 Résultats et discussion du scénario 2

Dans le scénario 2, les résultats expérimentaux sont synthétisés à travers plusieurs tableaux qui évaluent la récompense cumulative actualisée en fonction de la discrétisation de l'espace des croyances (paramètre N) et de divers choix de modélisation du comportement humain. On distingue notamment deux grandes familles de configurations.

D'une part, les tableaux 4.3 et 4.4 présentent les résultats obtenus lorsque le robot raisonne avec un indice de rationalité fixé ($\beta_R = 0.8$) sans procéder à une estimation dynamique de β. Le tableau 4.3 correspond au cas où l'objectif réel de l'humain est « vert », tandis que le tableau 4.4 le présente pour un objectif « rouge ». Ici, la croyance initiale du robot est symétrique ($b_0(vert) = b_0(rouge) = 0.5$) et l'absence d'ajustement de l'indice de rationalité montre que, malgré un croyance initiale uniforme l'humain arrive à atteindre son objectif.

Modèle humain	**Politique oracle**	**Réussite (%)**			**Discrétisations**		
		5	15	30	5	15	30
$\beta = 0.1$	$\mathbf{281} \pm 326$	100	100	100	278 ± 315	284 ± 329	$\mathbf{288} \pm 325$
$\beta = 0.8$	$\mathbf{221} \pm 99$	100	100	100	211 ± 84	$\mathbf{220} \pm 107$	209 ± 94
$\beta = 2$	$\mathbf{187} \pm 13$	100	100	100	$\mathbf{189} \pm 16$	180 ± 18	180 ± 16

TABLEAU 4.3 Récompense cumulative actualisée par N discrétisations de l'espace des croyances pour trois modèles différents de l'humain, la croyance initiale du robot par rapport aux objectifs est $b_0(vert) = b_0(rouge) = 0.5$ pour chaque objectif et l'objectif de l'humain est la case verte (voir figure 4.1b). Aucune estimation de la valeur du paramètre β n'est faite, le robot raisonne avec un estimé β_R de β égale à 0.8.

Modèle humain	**Politique oracle**	**Réussite (%)**			**Discrétisations**		
		5	15	30	5	15	30
$\beta = 0.1$	$\mathbf{319} \pm 371$	100	100	100	$\mathbf{314} \pm 388$	289 ± 342	278 ± 316
$\beta = 0.8$	$\mathbf{226} \pm 97$	100	100	100	211 ± 109	222 ± 98	$\mathbf{227} \pm 102$
$\beta = 2$	$\mathbf{198} \pm 19$	100	100	100	166 ± 25	186 ± 17	$\mathbf{186} \pm 4$

TABLEAU 4.4 Récompense cumulative actualisée par N discrétisations de l'espace des croyances pour trois modèles différents de l'humain, la croyance initiale du robot par rapport aux objectifs est $b_0(vert) = b_0(rouge) = 0.5$ pour chaque objectif et l'objectif de l'humain est la case rouge (voir figure 4.1b). Aucune estimation de la valeur du paramètre β n'est faite, le robot raisonne avec un estimé β_R de β égale à 0.8.

D'autre part, les tableaux 4.5 et 4.6 illustrent l'impact de l'estimation du paramètre β dans le système de décision du robot. Ce paramètre joue un rôle essentiel en permettant au robot de modéliser précisément la politique de Boltzmann, qui reflète le comportement humain. Pour estimer β de façon fiable, on utilise l'algorithme 7 dans un cadre hors ligne, avant que le robot n'entame ses prises de décision. Cette estimation s'appuie sur des trajectoires humaines collectées au préalable dans un environnement libre, ce qui rend l'apprentissage hors ligne particulièrement adapté. Une fois β déterminé, le robot peut élaborer sa stratégie de décision de manière plus efficace. À l'avenir, une piste intéressante serait de développer une approche en ligne, où le robot ajusterait son modèle en temps réel pour s'adapter dynamiquement. Pour un objectif réel « vert » (tableau 4.5) et pour un objectif « rouge » (tableau 4.6), on observe une amélioration de la récompense cumulative par rapport aux cas sans estimation. Ceci démontre qu'une adaptation dynamique de l'indice de rationalité permet de mieux saisir

le niveau de pertinence du choix d'actions de l'humain et, ainsi, d'optimiser la politique du robot.

Modèle humain	**Politique oracle**	**Réussite (%)**			**Discrétisations**		
		5	15	30	5	15	30
$\beta = 0.1$, $\beta_R = 0.12$	**27** $\pm$ 94	60	0	58	**-43** $\pm$ 40	-65 ± 3	-46 ± 35
$\beta = 0.8$, $\beta_R = 0.79$	**223** $\pm$ 98	100	100	100	216 ± 103	208 ± 95	**218** $\pm$ 111
$\beta = 2$, $\beta_R = 1.98$	**226** $\pm$ 11	100	100	100	225 ± 11	**227** $\pm$ 24	226 ± 15

TABLEAU 4.5 Récompense cumulative actualisée par N discrétisations de l'espace des croyances pour trois modèles différents de l'humain, la croyance initiale du robot par rapport aux objectifs est $b_0(vert) = b_0(rouge) = 0.5$ pour chaque objectif et l'objectif de l'humain est la case verte (voir figure 4.1b). Une estimation du paramètre a été faite afin de mettre à jour le modèle de l'humain dans le système de décision du robot à travers un estimé β_R.

Modèle humain	**Politique oracle**	**Réussite (%)**			**Discrétisations**		
		5	15	30	5	15	30
$\beta = 0.1$, $\beta_R = 0.12$	**4** $\pm$ 87	45	0	23	-40 ± 90	-65 ± 2	-30 ± 117
$\beta = 0.8$, $\beta_R = 0.79$	**227** $\pm$ 98	100	100	100	217 ± 97	**224** $\pm$ 100	215 ± 93
$\beta = 2$, $\beta_R = 1.98$	**237** $\pm$ 22	100	100	100	224 ± 15	224 ± 16	**225** $\pm$ 22

TABLEAU 4.6 Récompense cumulative actualisée par N discrétisations de l'espace des croyances pour trois modèles différents de l'humain, la croyance initiale du robot par rapport aux objectifs est $b_0(vert) = b_0(rouge) = 0.5$ pour chaque objectif et l'objectif de l'humain est la case rouge (voir figure 4.1b). Une estimation du paramètre a été faite afin de mettre à jour le modèle de l'humain dans le système de décision du robot à travers un estimé β_R.

Enfin, les tableaux 4.7 et 4.8 illustrent l'impact d'une croyance initiale erronée du robot sur sa prise de décision. À l'inverse, les tableaux 4.9 et 4.10 montrent les effets d'une croyance initiale mieux informée sur l'objectif réel de l'humain. Ces résultats indiquent que, selon l'objectif réel ("vert" ou "rouge") que ce biais influence les choix du robot. De plus, bien que l'estimation hors ligne de β, réalisée avec l'algorithme 7, soit précise, elle ne corrige pas entièrement cet effet. Une calibration adéquate de la croyance initiale demeure donc essentielle.

Modèle humain	Politique oracle	Réussite (%)			Discrétisations		
		5	15	30	5	15	30
$\beta = 0.1$, $\beta_R = 0.12$	$\mathbf{27} \pm 94$	0	0	0	-65 ± 2	-65 ± 2	-65 ± 3
$\beta = 0.8$, $\beta_R = 0.79$	$\mathbf{223} \pm 98$	100	100	100	205 ± 89	202 ± 83	$\mathbf{210} \pm 93$
$\beta = 2$, $\beta_R = 1.98$	$\mathbf{226} \pm 11$	100	100	100	$\mathbf{214} \pm 19$	213 ± 1	213 ± 1

TABLEAU 4.7 Récompense cumulative actualisée par N discrétisations de l'espace des croyances pour trois modèles différents de l'humain, la croyance initiale du robot par rapport aux objectifs $b_0(vert) = 0.15$ et $b_0(rouge) = 0.85$ pour chaque objectif et l'objectif de l'humain est la case verte (voir figure 4.1b). Une estimation du paramètre a été faite afin de mettre à jour le modèle de l'humain dans le système de décision du robot à travers un estimé β_R.

Modèle humain	Politique oracle	Réussite (%)			Discrétisations		
		5	15	30	5	15	30
$\beta = 0.1$, $\beta_R = 0.12$	$\mathbf{4} \pm 87$	73	55	69	-13 ± 126	-20 ± 153	-8 ± 166
$\beta = 0.8$, $\beta_R = 0.79$	$\mathbf{227} \pm 98$	100	100	100	218 ± 97	$\mathbf{219} \pm 101$	214 ± 94
$\beta = 2$, $\beta_R = 1.98$	$\mathbf{237} \pm 22$	100	100	100	223 ± 11	223 ± 1	$\mathbf{225} \pm 22$

TABLEAU 4.8 Récompense cumulative actualisée par N discrétisations de l'espace des croyances pour trois modèles différents de l'humain, la croyance initiale du robot par rapport aux objectifs $b_0(vert) = 0.85$ et $b_0(rouge) = 0.15$ pour chaque objectif et l'objectif de l'humain est la case rouge (voir figure 4.1b). Une estimation du paramètre a été faite afin de mettre à jour le modèle de l'humain dans le système de décision du robot à travers un estimé β_R.

Modèle humain	Politique oracle	Réussite (%)			Discrétisations		
		5	15	30	5	15	30
$\beta = 0.1$, $\beta_R = 0.12$	$\mathbf{27} \pm 94$	89	94	98	-10 ± 67	8 ± 92	$\mathbf{14} \pm 80$
$\beta = 0.8$, $\beta_R = 0.79$	$\mathbf{223} \pm 98$	100	100	100	216 ± 97	$\mathbf{217} \pm 105$	214 ± 104
$\beta = 2$, $\beta_R = 1.98$	$\mathbf{226} \pm 11$	100	100	100	227 ± 22	226 ± 19	$\mathbf{227} \pm 19$

TABLEAU 4.9 Récompense cumulative actualisée par N discrétisations de l'espace des croyances pour trois modèles différents de l'humain, la croyance initiale du robot par rapport aux objectifs $b_0(vert) = 0.85$ et $b_0(rouge) = 0.15$ pour chaque objectif et l'objectif de l'humain est la case verte (voir figure 4.1b). Une estimation du paramètre a été faite afin de mettre à jour le modèle de l'humain dans le système de décision du robot à travers un estimé β_R.

Modèle humain	Politique oracle	Réussite (%)			Discrétisations		
		5	15	30	5	15	30
$\beta = 0.1$, $\beta_R = 0.12$	$\mathbf{4} \pm 87$	0	0	0	-65 ± 2	-65 ± 2	-65 ± 2
$\beta = 0.8$, $\beta_R = 0.79$	$\mathbf{227} \pm 98$	100	100	100	$\mathbf{228} \pm 107$	221 ± 104	222 ± 95
$\beta = 2$, $\beta_R = 1.98$	$\mathbf{237} \pm 22$	100	100	100	236 ± 16	236 ± 16	$\mathbf{237} \pm 19$

TABLEAU 4.10 Récompense cumulative actualisée par N discrétisations de l'espace des croyances pour trois modèles différents de l'humain, la croyance initiale du robot par rapport aux objectifs $b_0(vert) = 0.15$ et $b_0(rouge) = 0.85$ pour chaque objectif et l'objectif de l'humain est la case rouge (voir figure 4.1b). Une estimation du paramètre a été faite afin de mettre à jour le modèle de l'humain dans le système de décision du robot à travers un estimé β_R.

Dans l'analyse, la performance des politiques discrétisées s'évalue par rapport à celle de la politique oracle, qui sert de référence idéale en disposant d'une connaissance parfaite de l'objectif humain. En général, une discrétisation fine de l'espace des croyances rapproche la performance de celle de l'oracle, même si l'écart reste sensible au calibrage du modèle humain utilisé par le robot.

Performance des politiques du robot

Les politiques discrétisées (N = 5, 15, 30) atteignent systématiquement un taux de réussite de 100 %, égalant ainsi la politique oracle sauf dans les cas où l'humain se comporte de manière quasi aléatoire. Ce résultat souligne la robustesse de l'approche POMDP discrétisée pour résoudre les ambiguïtés d'objectif, malgré une modélisation approximative des croyances. Néanmoins, les récompenses cumulées présentent des différences : bien que certains niveaux de discrétisation atteignent les performances de l'oracle (voir performance pour $\beta = 2$ dans le tableau 4.5) elles sont en moyenne inférieure de 10 à 20 points par rapport à l'oracle, en raison d'actions sous-optimales induites par l'approximation des croyances ou par la croyance uniforme du robot qui est obligé au départ de prendre une action aléatoire au départ pour gagner de l'information (ceci le pénalise d'une récompense de -10).

Par exemple, pour un humain semi-rationnel ($\beta = 0.8$), la politique avec N = 5 atteint 211 ± 84 (tableau 4.3), contre 221 ± 99 pour l'oracle. Cet écart s'explique par des erreurs de discrétisation lors de mises à jour des croyances entraînant la prise d'action inutile par le robot. Alors que d'un autre côté, pour N = 15, le robot atteint une performance de 220 ± 107 équivalent à la performance de l'oracle. Les variances sont influencées par plusieurs facteurs, notamment le modèle probabiliste des actions humaines, la discordance entre le modèle réel

et celui utilisé par le robot, ainsi que la granularité des mises à jour des croyances. Il est essentiel d'observer l'impact d'autres facteurs sur les performances du robot.

Impact de la discrétisation N

La granularité de la discrétisation (N) influence les performances. Bien que $N = 15$ améliore légèrement les récompenses (par exemple pour $\beta = 0.8$, tableau 4.6 : 224 ± 100 contre 227 ± 98 pour l'oracle), une diminution de la performance est observée pour $N = 30$ (215 ± 93). Dans la majorité des cas, une meilleure performance des politiques est constatée pour des discrétisations plus fines. Une discrétisation grossière de l'espace des croyances ($N = 5$) entraîne une perte d'information sur l'évolution réelle de la croyance du robot (voir annexe C), ce qui dégrade la performance de ses stratégies de décision par rapport à une discrétisation plus fine. Cependant, dans certains cas, la différence avec les politiques utilisant des discrétisations plus fines reste négligeable (voir $\beta = 2$ dans le tableau 4.6). Cela suggère que le scénario testé ne met pas suffisamment en évidence une amélioration proportionnelle à un affinage plus poussé de l'espace des croyances.

Impact de l'apprentissage de β

L'apprentissage du paramètre β améliore significativement les performances du robot, en particulier lorsqu'on compare les résultats de la politique oracle et des politiques discrétisées pour un modèle de l'humain appris avec $\beta_R = 1.98$ à ceux obtenus sans estimation préalable (voir tableau 4.6 et tableau 4.4).

Impact de la croyance initiale

Une croyance initiale erronée réduit la performance du robot. En comparant la meilleure performance du robot en interaction avec un humain semi-rationnel ($\beta = 0.8$) ayant une croyance uniforme sur les objectifs (218 ± 111, voir tableau 4.5) à celle obtenue avec une croyance initiale biaisée (210 ± 93, voir figure 4.7), une diminution de performance est observée. Un effet similaire se manifeste avec un humain quasi rationnel. Cependant, les performances rapportées dans les tableaux 4.6 et 4.7 restent comparables, car le robot prend initialement des décisions aléatoires lorsqu'il ne dispose d'aucune information avantageuse sur l'objectif de l'humain (croyance uniforme). En revanche, avec une croyance initiale mieux informée sur l'objectif réel de l'humain (voir tableaux 4.9 et 4.10), les performances du robot s'améliorent significativement par rapport à une croyance uniforme.

Commentaires supplémentaires sur les résultats

Lorsque le robot opère avec un modèle humain semi-rationnel c'est-à-dire en supposant un indice de rationalité élevée ($\beta_R = 0.8$), il interprète les actions observées comme cohérentes et informatives. Dans les tableaux 4.3 et 4.4, en partant d'une croyance initiale neutre ($b_0(vert) = b_0(rouge) = 0.5$), le robot, ne disposant d'aucune connaissance préalable sur les préférences de l'humain, se voit contraint d'ouvrir une porte de manière aléatoire – en général, la porte la plus proche (celle de l'objectif en rouge) – afin d'obtenir des informations supplémentaires. Si le modèle de l'humain est rationnel, le robot parvient à mettre à jour sa croyance sur le véritable objectif. En revanche, lorsque l'humain se comporte de manière quasi aléatoire ($\beta = 0.1$), la situation se complique. Si le robot conserve un modèle basé sur une haute rationalité ($\beta_R = 0.8$), il interprète alors les actions irrationnelles comme des comportements semi-rationnels. Cela peut occasionnellement conduire à des « coups de chance », c'est-à-dire à des actions irrationnelles qui, par hasard, orientent l'humain vers son objectif, générant ainsi une récompense cumulative comparable à celle observée dans les tableaux mentionnés. Par ailleurs, si le robot ajuste son modèle pour tenir compte de l'irrationalité de l'humain (en passant à $\beta_R = 0.12$), il ne considère plus les observations humaines comme suffisamment pertinentes pour modifier sa croyance. Autrement dit, il ne « déverrouille pas » les portes nécessaires pour confirmer l'objectif, ce qui se traduit par une diminution notable de la récompense cumulative (voir tableaux 4.5 et 4.6). Ainsi, bien que la discordance entre le modèle de l'humain utilisé par le robot et le comportement réel puisse, dans certains cas, améliorer la performance de l'équipe humain-robot, elle conduit globalement à une performance sous-optimale par rapport au benchmark oracle.

4.3.4 Résultats et discussion du scénario 3

Comme mentionné à la section 4.1, le robot se retrouve dans une situation d'impasse (voir la figure 4.10). L'expérimentation reproduit les mêmes conditions que le scénario 2. Dans ce contexte, la croyance initiale est entièrement centrée sur l'objectif rouge.

Les simulations montrent que, quelle que soit la politique de décision issue des différents niveaux de discrétisation testés, le robot ne parvient pas à guider l'humain vers son objectif. Ce constat se traduit par exemple par des récompenses cumulatives actualisées de -23 ± 9 (voir tableau 4.11) pour une politique avec un niveau de discrétisation $N = 30$ face à un humain semi-rationnel ($\beta = 0.8$).

Cette limitation s'explique par le fait que le robot ne peut maintenir qu'une seule porte ouverte (celle menant à l'objectif vert), tandis que l'autre reste fermée. Cette configuration

Modèle humain	**Discrétisations**		
	5	15	30
$\beta = 0.1$, $\beta_R = 0.12$	-46 ± 4	-46 ± 1	-47 ± 1
$\beta = 0.8$, $\beta_R = 0.79$	-31 ± 1.5	11 ± 43	-23 ± 9
$\beta = 2$, $\beta_R = 1.98$	-12 ± 0.1	-12 ± 0	-13 ± 0.1

TABLEAU 4.11 Récompense cumulative actualisée par N discrétisation de l'espace des croyances pour trois modèles différents de l'humain, l'objectif de l'humain est $g = rouge$. Une estimation du paramètre a été faite afin de mettre à jour le modèle de l'humain dans le système de décision du robot à travers un estimé β_R.

dissuade efficacement l'humain de se diriger vers l'objectif rouge, créant ainsi une impasse. En effet, l'humain en l'absence d'assistance du robot n'est pas capable d'atteindre son objectif.

Ainsi, aucune solution ne permet de résoudre cette situation avec un modèle mental de premier ordre. Cependant, une solution alternative pourrait émerger en exploitant un potentiel comportement quasi aléatoire de l'humain. En ouvrant une porte, le robot mise sur l'espoir que l'humain, par une décision imprévisible et peu logique, choisisse de quitter la salle. Bien que cette hypothèse repose sur une probabilité faible, elle offre une possibilité : si l'humain franchit la porte et libère l'espace, le robot pourrait alors ouvrir la porte appropriée pour atteindre son objectif.

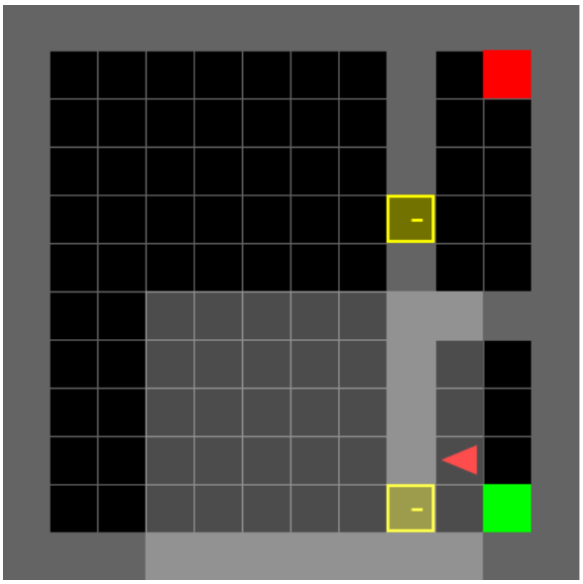


FIGURE 4.10 Environnement où l'humain (agent en rouge) a pour but l'objectif en rouge, mais se trouve coincé dans la salle de l'objectif vert, mais le robot ne peut que maintenir qu'une porte à la fois.

4.4 Conclusion

Ce chapitre a présenté l'environnement de simulation, les défis computationnels rencontrés et les performances des politiques calculées dans divers scénarios. Les résultats montrent que les politiques POMDP permettent au robot d'assister efficacement l'humain, en particulier lorsqu'une estimation de l'indice de rationalité est intégrée. En général, les politiques avec une discrétisation plus fine performent mieux que celle avec une discrétisation grossière. Enfin, le dernier scénario souligne les limites d'un modèle mental de premier ordre, incapable de résoudre certaines situations d'impasse, en l'absence de coordination entre l'humain et le robot.

CHAPITRE 5 PLANIFICATION DES DÉCISIONS DU ROBOT D'ORDRE 2

Dans ce chapitre, il est montré que l'hypothèse 1 est parfois insuffisante pour certains scénarios d'interaction dans une équipe humain-robot (voir 4.3.4). Cette hypothèse a permis de concevoir un modèle mental de premier ordre guidant les décisions du robot (section 3.3.3), qui sont démontrés capable de résoudre les scénarios 1 et 2 (voir 4.3.2 et 4.3.3). Cependant, son incapacité à résoudre l'impasse du scénario 3 met en évidence ses limites. Une extension vers un modèle mental de second ordre (section 2.2.2) est alors proposée pour surmonter ces obstacles et initier des travaux futurs.

5.1 Mise en contexte

Essentiellement, nous formulons une nouvelle hypothèse :

Hypothèse 4. *Dans ce chapitre, il est supposé que l'humain est conscient de la présence d'un robot.*

Cette hypothèse demande la connaissance de la stratégie de décision du robot par l'humain pour qu'il puisse planifier sa politique. Ceci ramène le problème de récursion abordé dans la section 2.2.2, de par le fait que l'humain doit raisonner sur la stratégie du robot, alors que cette stratégie est celle qui doit être déterminée en fonction du modèle de décision de l'humain et vice versa. Comme rappel, afin de régler ce problème, [40] propose un scénario coopératif dans lequel les deux agents, l'humain et le robot, partagent des informations ou des actions. Dans ce cadre, le robot calcule des politiques jointes pour lui-même et pour l'humain. Il extrait ensuite la politique de l'humain à partir de cette solution jointe en la marginalisant, permettant d'obtenir des règles d'action distinctes pour chacun des agents. Le robot peut donc planifier ses propres actions tout en prenant en compte ce que l'humain anticipe de ses décisions. Cependant, cette solution présente un problème : elle suppose que l'humain et le robot coopèrent en considérant que l'humain contrôle le robot. Or, cette hypothèse n'est pas réaliste, car, dans de nombreux cas, l'humain n'est ni en mesure de déterminer la meilleure stratégie de décision du robot ni disposé à coopérer immédiatement avec ce dernier. Cette limitation affecte le calcul de la politique du robot, pouvant ainsi conduire à des politiques sous-optimales.

D'autres méthodes décrites dans la revue de littérature (voir section 2.2.2) s'attaquent à ce problème sous divers aspects. Nous proposons une méthode itérative pour résoudre ce

problème en nous basant sur une hypothèse importante qui nous évite de connaître dès le départ la politique du robot, politique que nous essayons activement de déduire.

5.2 Planifications du robot avec un modèle mental de deuxième ordre

5.2.1 Approche de solution au modèle mental de deuxième ordre

Afin d'apporter une solution, nous définissons une hypothèse permettant à l'humain de construire itérativement sa politique autour de la politique du robot qu'il ne connaît pas au départ.

Hypothèse 5. *Initialement, l'humain suppose que la politique du robot repose sur un modèle mental de premier ordre.*

Cette hypothèse postule que le modèle de décision du robot s'apparente à un modèle mental de premier ordre (voir hypothèse 1). Elle permet, dans un premier temps, de résoudre le problème initial lié à l'intégration de la politique du robot dans la planification de l'humain. Par la suite, elle autorise le robot à déduire une stratégie de décision actualisée à partir de cette première planification, entraînant ainsi un processus itératif entre les deux agents pour optimiser leur collaboration. En général, notre structure de résolution s'appuie sur un modèle mental de premier ordre pour définir la politique de l'humain tenant compte de la présence du robot. Cela permet de construire un modèle mental de deuxième ordre. Ce processus peut être itéré pour raffiner successivement les modèles. Le schéma de la figure 5.1 illustre l'approche adoptée pour résoudre le problème. Il s'appuie sur la modélisation des interactions présentée dans la section 3.3.3, en tirant parti des algorithmes développés pour planifier et optimiser les décisions du robot et de l'humain. Ce processus met en œuvre une progression logique, où la politique de chaque agent est ajustée en fonction des itérations. Soit une fonction, $R_u(\pi^A, \pi^B)$ qui représente le rendement obtenu en utilisant la politique π^A contre π^B. Notre objectif est de déterminer une paire de politiques $<\pi_H^n, \pi_R^n>$, représentant les stratégies respectives de l'humain et du robot, obtenues après n itérations qui donne de meilleur rendement que le couple de stratégie à $n-1$ itérations (voir définition 2). Ce processus repose sur une optimisation itérative, où chaque agent ajuste sa stratégie en fonction de la politique de l'autre, dans le but de converger vers un équilibre d'interaction optimale.

Définition 2. *La meilleure réponse d'un joueur A face à une politique π^B d'un joueur B est une politique π^A_{BR}, telle que $R_u(\pi^A_{BR}, \pi^B) \geq R_u(\pi^A, \pi^B)$ pour toute autre possible stratégie π^A [68].*

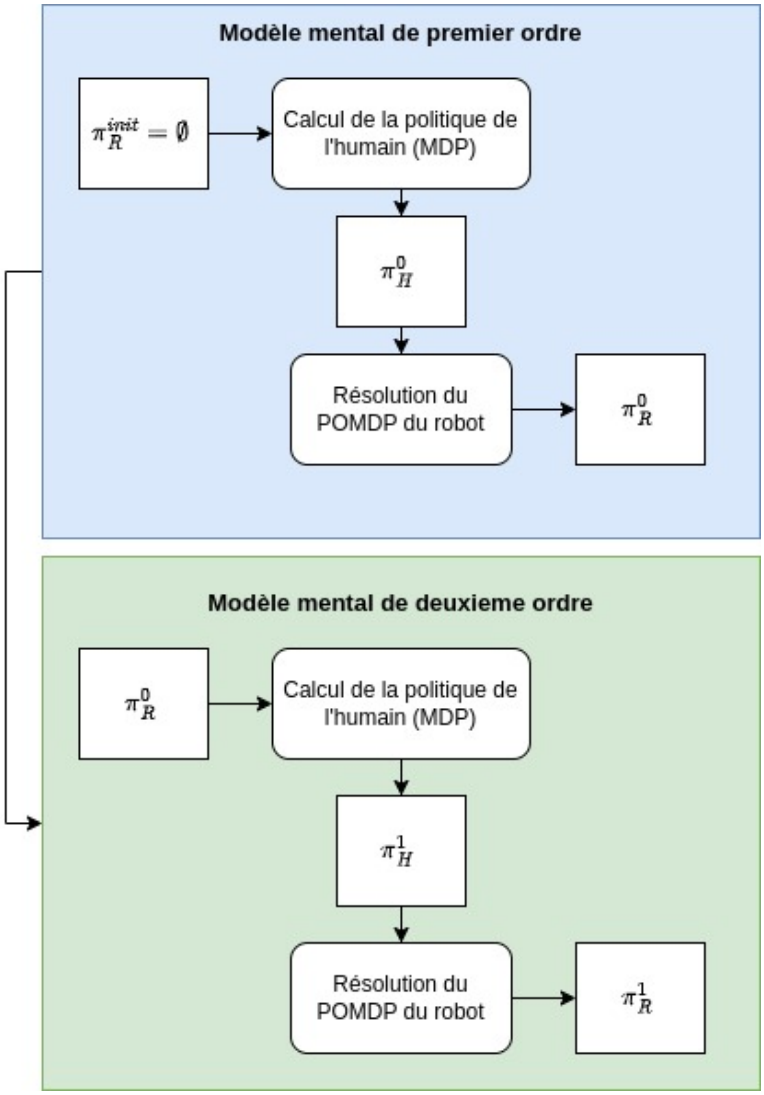


FIGURE 5.1 Observation de l'approche de solution basée sur le modèle mental de second ordre après n = 1 itération pour obtenir $<\pi_H^1, \pi_R^1>$

Dans ce cadre, chaque agent, à son tour, ajuste sa stratégie en fonction de celle de l'autre pour maximiser ses récompenses, ce qui illustre une dynamique itérative d'optimisation stratégique.

5.2.2 Formulation du modèle et algorithme

La formulation du problème est similaire à celle présentée dans la section 3.1, et les équations nécessaires à l'approche de solution sont principalement développées dans la section 3.3.3. La principale différence réside dans le calcul de la politique de l'humain, qui prend en compte l'existence d'un acteur (le robot) capable de modifier le modèle du monde de l'humain via une politique initiale π_R^0 obtenu à l'aide de l'algorithme 6. Ainsi, le MDP de l'humain est décrit par le tuple $(X^H, W, A^H, T_x^H, T_{w'}, R^H, G, B, b_0, \pi_R)$, similaire à celui de la section 3.1, avec la particularité que l'humain est capable de calculer la croyance du robot et de raisonner sur les actions possibles du robot π_R .

La nouvelle fonction de valeur d'actions et d'états est :

$$Q^H(x_t^H, w_t^{'}, a_t^H, g, b_t) = R_t^H + \gamma \sum_{x_{t+1}^H \in X^H} \sum_{w_{t+1}^{'} \in W} T_x^H(x_t^H, w_t^{'}, g, a_t^H, x_{t+1}^H) \times T_{w'}(w_t^{'}, \pi_R(x_{t+1}^H, w_t^{'}, b_{t+1}(g)), w_{t+1}^{'}) \times V^H(x_{t+1}^H, w_{t+1}^{'}, g, b_{t+1}(g)) \tag{5.1}$$

Avec $b_{t+1}(g)$ issue de l'equation (3.29). Quelques modifications sont alors apportées à l'algorithme 8 de la manière suivante :

Algorithm 8: Itération des valeurs MDP de l'humain pour un modèle mental de second ordre

Input: Un MDP $\langle X^H, W, A^H, T_x^H, R^H, G, B, b_0, \pi_R \rangle$, facteur d'atténuation γ, un paramètre de précision ϵ, la politique du robot π_R^0

Output: Politique optimale π_H^1

Initialiser arbitrairement $V(x^H, w, g, b)$, $\forall x^H \in X^H, w \in W, g \in G, b \in B$;
Initialiser arbitrairement $Q(x^H, w, a^H, g, b)$, $\forall x^H \in X^H, w \in W, g \in G, b \in B, a^H \in A^H$;
Initialiser arbitrairement $V(x^H, w, g, b)$, $\forall x^H \in X^H, w \in W, g \in G, b \in B$;
repeat
    for *chaque* $g \in G, b \in B, w \in W, x^H \in X^H$ **do**
        temp $\leftarrow V(x^H, w_t, g, b)$;
        for *chaque* $a^H \in A^H$ **do**
            $a_R = \pi_R^0(x^H, w_t, b)$;
            Appliquer (5.1)
        $V^H(x^H, w, g, b) \leftarrow \max_{a^H \in A^H} Q^H(x^H, w, a^H, g, b)$;
until $\max_{x^H, w, g, b} |temp - V(x^H, w, g, b)| \leq \epsilon$;
$\forall x^H, w, g, b, \pi_H^1(x^H, w, g, b) = \arg\max_{a^H \in A^H} Q^H(x^H, w, a^H, g, b)$;

Les fonctions Q^H obtenues par cet algorithme sont transformées en politiques softmax selon la même logique que l'algorithme 3. Ce résultat implique que l'humain est modélisé comme étant capable de prendre des décisions selon sa connaissance du comportement du robot. Le calcul de la politique du robot π_R^1 suit l'algorithme présenté en 6, en supposant que l'humain peut calculer la croyance du robot en connaissant sa croyance initiale.

La planification de décision d'ordre deux du robot intègre les réactions de l'humain face à sa présence, ce qui permet d'évaluer la capacité de ce modèle à résoudre les situations d'impasse. Par exemple, lorsque l'humain est bloqué dans la salle de l'objectif vert (voir figure 4.10), son modèle anticipe l'intervention du robot, qui consistera à ouvrir la porte. Cette anticipation incite alors l'humain à se déplacer vers la sortie. De son côté, le robot, tenant compte de ce modèle de l'humain, maintiendra la porte de sortie ouverte (tout en sachant que la porte menant à l'objectif rouge est fermée, car une seule porte peut être ouverte à la fois) jusqu'à ce que l'humain quitte la salle. Ensuite, le robot ouvrira la porte menant à l'objectif rouge. Cette logique de résolution du problème d'impasse sera confirmée à la section suivante à l'aide de l'approche développée.

5.3 Expérimentations et discussion de l'algorithme

Les conditions d'expérimentation du scénario 3, présentées dans la section 4.3.4, sont appliquées pour évaluer la méthode développée. La politique du robot testée correspond à celle obtenue après une seule itération ($n = 1$) (voir figure 5.1). Il n'est pas nécessaire d'itérer jusqu'à obtenir un équilibre ou des politiques qui ne varient plus, étant donné la simplicité de la tâche à accomplir. Des travaux futurs pourraient explorer l'impact d'itérations supplémentaires sur la performance ou tester des problèmes nécessitant une coordination plus poussée entre agents.

Modèle humain	**Discrétisations**		
	5	15	30
$\beta = 0.1$, $\beta_R = 0.12$	-46 ± 11	-37 ± 27	-44 ± 26
$\beta = 0.8$, $\beta_R = 0.79$	213 ± 96	207 ± 97	210 ± 100
$\beta = 2$, $\beta_R = 1.98$	122 ± 80	122 ± 81	112 ± 83

Tableau 5.1 Récompense cumulative actualisée évaluée en fonction de N, représentant la discrétisation de l'espace des croyances, pour trois modèles distincts d'humain. L'objectif poursuivi par l'humain est fixé à la case rouge dans le cadre du scénario 3. Une estimation de l'indice de rationalité hors ligne a été réalisée via un estimé β_R.

Les résultats obtenus surpassent généralement ceux du modèle de planification du robot

d'ordre 1. En effet, pour un humain semi-rationnel ou quasi rationnel, les performances indiquent que le robot parvient à assister efficacement l'humain, avec des récompenses cumulatives actualisées positives, supérieures à celles du tableau 4.11. Par exemple, pour une politique discrétisée avec $N = 30$ et un modèle humain $\beta = 0.8$, une récompense de 215 ± 106 est obtenue (voir tableau 5.1), contre -23 ± 9 dans le tableau 4.11.

En revanche, pour un humain quasi aléatoire, la méthode ne présente pas de bonnes performances, suggérant ainsi que le robot ne parvient pas toujours à générer d'actions d'assistance lorsqu'il perçoit l'humain comme agissant de manière non optimale. On en déduit que le modèle de planification d'ordre 2 du robot est plus efficace que celui d'ordre 1 dans les situations nécessitant une coordination entre l'humain et le robot, à condition que l'humain respecte au moins l'hypothèse 3. Il serait intéressant, dans les travaux futurs, de tester cet algorithme dans des situations où plusieurs objectifs sont possibles en dehors de l'impasse.

5.4 Conclusion

Dans ce chapitre, une solution au problème d'impasse a été proposée, aboutissant à des résultats satisfaisants qui démontrent l'efficacité de cet algorithme par rapport au modèle de décision du robot d'ordre 1 décrit à la section 3.3.3.

CHAPITRE 6 CONCLUSION

Le dernier chapitre de ce mémoire résume les travaux abordés dans les chapitres précédents. Les limites de ces derniers sont identifiées, et des pistes de réflexion pour de futures recherches sont proposées en conclusion.

6.1 Synthèse des travaux

L'objectif de cette recherche a été de concevoir des algorithmes permettant de calculer une stratégie de décision pour un robot interagissant avec un agent humain effectuant une tâche critique, sans possibilité de communication directe. Le robot ne connaît pas l'objectif de l'humain et doit ainsi raisonner sur celui-ci.

Tout d'abord, le scénario d'interaction entre les agents a été défini. Plus précisément, un humain bloqué par des obstacles cherche à accomplir une tâche essentielle. Le robot, sans communication directe avec l'humain, observe ses actions afin d'inférer son objectif et retire stratégiquement les obstacles les plus pertinents tout en minimisant ses coûts. Ces interventions facilitent la progression de l'humain, tandis que le robot ajuste ses hypothèses en temps réel. Ce scénario a été modélisé dans un environnement en grille, où l'agent humain peut exécuter des actions discrètes (bas, haut, gauche, droite, attendre) en fonction d'un objectif précis. De son côté, le robot est considéré comme une entité automatisée capable d'observer l'humain et d'interagir avec l'environnement en ouvrant ou en fermant des portes, considérées comme des obstacles.

La tâche que l'humain cherche à accomplir a été formulée sous la forme d'un Processus Décisionnel Markovien (MDP), un cadre permettant d'intégrer de manière explicite les informations disponibles. Le modèle de l'humain adopté par le robot repose sur un modèle mental de premier ordre, selon lequel l'humain ne raisonne pas en fonction de la présence d'une machine susceptible de l'aider ou de le gêner. Pour prendre en compte les incertitudes et les informations non observables par le robot, une politique de décision basée sur une fonction softmax a été introduite. Ce modèle utilise une fonction de valeur d'action-état Q^H en entrée et génère une distribution de probabilité sur les actions possibles de l'humain. Cette modélisation est ensuite intégrée à la stratégie de prise de décision du robot. Pour corriger les erreurs de modélisation de la politique humaine, une méthode d'apprentissage par montée de gradient a été mise en place.

La planification des décisions d'ordre 1 du robot est représentée sous la forme d'un Processus

Décisionnel de Markov Partiellement Observable (POMDP), une extension du MDP. Ce cadre permet au robot de gérer l'incertitude sur l'objectif de l'humain en maintenant une croyance à son sujet, tout en intégrant le modèle mental de premier ordre qu'il possède de l'humain. En raison de la complexité inhérente à la résolution d'un tel problème, une approche basée sur la discrétisation de l'espace des croyances du robot a été adoptée. Cette méthode discrétise l'espace des croyances en un nombre fini d'éléments, permettant ainsi d'expérimenter son efficacité en la comparant à une politique oracle disposant d'une connaissance parfaite de l'objectif de l'humain. Plus précisément, l'analyse a porté sur l'efficacité de l'assistance apportée par le robot en fonction du niveau de discrétisation et de sa croyance initiale sur l'objectif de l'humain. Les résultats montrent que ce modèle assiste efficacement l'humain, atteignant des performances proches de celles d'une politique oracle (ayant une connaissance parfaite de l'objectif de l'humain), notamment après l'apprentissage de l'indice de rationalité de l'humain. Toutefois, il présente des limites : il échoue lorsque l'humain adopte un comportement irrationnel et obtient des résultats médiocres lorsque la croyance initiale est erronée ou dans des situations d'impasse.

Après avoir exploré les limitations de l'incorporation d'un modèle mental de premier ordre, une solution a été proposée en tirant parti d'un modèle mental de second ordre, construit à partir du modèle de premier ordre. L'objectif de cette approche est de résoudre un problème plus complexe nécessitant un certain degré de coordination entre l'humain et le robot. Toutefois, l'introduction de ce modèle de second ordre a entraîné un défi supplémentaire, celui de la récursion : le robot doit raisonner sur l'humain, qui lui-même raisonne sur le robot pour déduire sa politique. Ce problème de récursion a été résolu par la construction d'un modèle itératif, dans lequel la politique initiale est celle du robot utilisant un modèle mental de premier ordre de l'humain. Cette méthode a montré son efficacité, les récompenses cumulées actualisées obtenues étant plus élevées que celles du modèle de planification de d'ordre 1 du robot. Ces résultats démontrent que l'intégration d'un modèle mental de second ordre permet d'obtenir de meilleures performances, notamment dans des scénarios nécessitant une coordination plus fine entre les agents.

6.2 Limitations de la solution proposée

6.2.1 Modèle de l'humain

La première limitation réside dans la méthodologie de modélisation du comportement humain. Ce modèle a été choisi en raison de l'impossibilité de modéliser tous les facteurs influençant la prise de décision de l'humain. En effet, comme observé dans les situations où plusieurs chemins

optimaux mènent à un objectif, le robot ne parvient pas à distinguer de préférence entre ces chemins. Pourtant, il est plausible que l'humain privilégie l'un d'entre eux spécifiquement, car cela pourrait le rapprocher d'un objectif secondaire.

L'apprentissage de l'indice de rationalité, lorsque l'humain dévie du modèle utilisé par le robot pour sa prise de décision, permet de pallier ces différences. Cependant, cela nécessite la collecte de données en temps réel et une mise à jour rapide du modèle, ce qui requiert une analyse approfondie du temps d'apprentissage et des ressources nécessaires.

6.2.2 Stratégie de décision du robot

La stratégie de décision du robot a conduit à des actions permettant d'aider l'humain. Cependant, il a été conclu que, lorsque le robot raisonne sur un modèle d'humain irrationnel, il échoue dans la plupart des cas à l'assister. Plus l'humain agit de manière irrationnelle par rapport au plan optimal, plus l'incertitude augmente concernant les éléments qui influencent sa prise de décision. Cela nécessite donc une stratégie de décision du robot robuste.

D'autre part, la discrétisation de l'espace de croyance du robot entraîne une perte d'information sur la véritable croyance du robot, car elle l'approxime en utilisant l'élément le plus proche dans l'espace des croyances discrétisées. Par exemple, si l'espace des croyances ne contient que 5 éléments $\{0, 0.25, 0.5, 0.75, 1\}$ et que la croyance actuelle du robot est 0.6, celle-ci sera approximée à 0.5, ce qui entraîne une perte significative d'information. De plus, plus l'espace des croyances est discrétisé, plus l'algorithme mettra de temps à converger en raison de l'explosion de l'espace d'états (comme indiqué dans l'équation 4.1).

6.2.3 Méthode Itérative de calcul du modèle mental de deuxième ordre

Le principal problème de cette méthodologie réside dans la politique initiale utilisée. Les limitations de la stratégie du robot, basée sur un modèle mental de premier ordre, ont été abordées dans les sous-sections précédentes. Ces limitations affectent le calcul de la politique de deuxième ordre. Cette dépendance repose sur une hypothèse forte d'optimalité de la politique initiale, puisque la stratégie repose sur une itération des meilleures réponses à chaque politique.

De plus, il est supposé que l'humain est capable de raisonner sur la croyance du robot pour traiter la politique initiale. Cette hypothèse peut s'avérer incorrecte dans des situations où l'humain est surchargé mentalement ou ne tient pas compte de l'état du robot, rendant impossible l'application de l'algorithme itératif.

6.3 Améliorations futures

6.3.1 Tester le modèle face aux incertitudes

L'une des limitations de notre méthodologie réside dans l'absence de validation expérimentale pour différents modèles stochastiques du comportement humain. L'hypothèse retenue stipule que l'humain change d'état de manière déterministe, tout comme l'environnement. Il serait donc pertinent de tester ces modèles dans des scénarios où les états présentent une dynamique stochastique. De plus, les scénarios utilisés pour les expérimentations ne sont pas suffisamment spécifiques ou spécialisés pour différencier clairement l'impact de chaque niveau de discrétisation de l'espace des croyances. Un tel affinement permettrait de conclure de manière plus précise sur la robustesse de l'implémentation suggérée.

6.3.2 Comparer la solution du robot à d'autres approches

L'implémentation d'autres modèles pour comparaison est relativement complexe et nécessite souvent des adaptations pour les faire fonctionner dans un scénario spécifique. Pour de futurs travaux, il serait pertinent de développer une bibliothèque standardisée intégrant des scénarios spécifiques, accessibles à tous, afin de faciliter les comparaisons entre les différentes méthodes de stratégie de décision dans le cadre d'une collaboration humain-robot.

6.3.3 Modèle amélioré de l'humain et robot

L'élaboration de la stratégie de décision du robot repose principalement sur le modèle de l'humain. Il serait pertinent d'intégrer des méthodes d'apprentissage par renforcement, telles que le Q-learning [25, 69], ou des méthodes basées sur l'apprentissage par renforcement distributionnel [70], en exploitant les fonctions d'action-état Q calculés par les méthodes développées, afin d'apprendre la fonction de valeur d'action et d'état au fur et à mesure de l'interaction. Le robot pourrait également adopter des techniques d'apprentissage profond par renforcement multi-agent pour tirer parti de l'efficacité des réseaux de neurones. Par ailleurs, une hypothèse implicite est que le robot connaît la fonction de récompense de l'humain, ce qui n'est pas nécessairement le cas. Ainsi, il serait pertinent d'intégrer une méthode d'approximation de la fonction de récompense réelle de l'humain, ou de maintenir une distribution de probabilité sur les fonctions de récompense possibles. D'un autre côté, l'utilisation de la méthode itérative pour la planification du robot d'ordre 2 ouvre la voie à une extension vers un formalisme plus rigoureux, fondé sur la théorie des jeux. Cette approche permettrait de mettre en évidence des concepts essentiels, tels que les équilibres de Nash.

RÉFÉRENCES

[1] M. Ren, N. Chen et H. Qiu, "Human-machine collaborative decision-making : An evolutionary roadmap based on cognitive intelligence," *International Journal of Social Robotics*, p. 1101–1114, 2023. [En ligne]. Disponible : http://dx.doi.org/10.1007/s12369-023-01020-1

[2] J. Shah et C. Breazeal, "An empirical analysis of team coordination behaviors and action planning with application to human-robot teaming," *Human factors*, vol. 52, p. 234–245, 2010.

[3] F. Terrence, R. N. Illah, K. Clayton, F. Lorenzo, J. S., R. O. Ambrose, R. R. Burridge, R. G. Simmons, L. M. Hiatt, A. C. Schultz, J. G. Trafton, M. D. Bugajska et J. Schermerhorn, "The peer-to-peer human-robot interaction project," *Space*, 2005. [En ligne]. Disponible : https://api.semanticscholar.org/CorpusID:2188287

[4] M. B. Dias, B. Kannan, B. Browning, E. G. Jones, B. Argall, M. F. Dias, M. Zinck, M. M. Veloso et A. J. Stentz, "Sliding autonomy for peer-to-peer human-robot teams," dans *Intelligent Autonomous Systems 10*. IOS Press, 2008, p. 332–341.

[5] A. Henschel, G. Laban et E. Cross, "What makes a robot social ? a review of social robots from science fiction to a home or hospital near you," *Current Robotics Reports*, vol. 2, 2021.

[6] F. Cantucci et R. Falcone, "Collaborative autonomy : Human–robot interaction to the test of intelligent help," *Electronics*, vol. 11, 2022. [En ligne]. Disponible : https://www.mdpi.com/2079-9292/11/19/3065

[7] K. M. Rabby, M. Khan, A. Karimoddini et S. X. Jiang, "An effective model for human cognitive performance within a human-robot collaboration framework," dans *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, 2019, p. 3872–3877.

[8] V. V. Unhelkar, S. Li et J. A. Shah, "Decision-making for bidirectional communication in sequential human-robot collaborative tasks," dans *2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2020, p. 329–341.

[9] S. Nikolaidis, M. Kwon, J. Forlizzi et S. Srinivasa, "Planning with verbal communication for human-robot collaboration," *J. Hum.-Robot Interact.*, vol. 7, 2018. [En ligne]. Disponible : https://doi.org/10.1145/3203305

[10] A. Bauer, D. Wollherr et M. Buss, "Human–robot collaboration : A survey," *International Journal of Humanoid Robotics*, vol. 5, n°. 1, p. 47–66, 2008. [En ligne]. Disponible : https://doi.org/10.1142/S0219843608001303

[11] J. Sweller, “Cognitive load during problem solving : Effects on learning,” *Cognitive Science*, vol. 12, p. 257–285, 1988. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/0364021388900237

[12] L. Peternel, N. Tsagarakis, D. Caldwell et A. Ajoudani, “Robot adaptation to human physical fatigue in human–robot co-manipulation,” *Autonomous Robots*, vol. 42, p. 1–11, 06 2018.

[13] C. Baker, J. Tenenbaum et R. Saxe, “Goal inference as inverse planning,” dans *Proceedings of the 29th Annual Conference of the Cognitive Science Society*, 01 2007.

[14] C. L. Baker, J. B. Tenenbaum et R. R. Saxe, “Bayesian models of human action understanding,” dans *Proceedings of the 19th International Conference on Neural Information Processing Systems*, ser. NIPS’05. MIT Press, 2005, p. 99–106.

[15] G. Sukthankar, C. Geib, H. H. Bui, D. Pynadath et R. P. Goldman, *Plan, Activity, and Intent Recognition : Theory and Practice*. Morgan Kaufmann Publishers Inc., 2014.

[16] K. Erol, J. Hendler et D. Nau, “Complexity results for htn planning,” *Annals of Mathematics and Artificial Intelligence*, vol. 18, 04 2003.

[17] M. Ramírez et H. Geffner, “Goal recognition over pomdps : inferring the intention of a pomdp agent,” ser. IJCAI’11. AAAI Press, 2011, p. 2009–2014.

[18] G. Gergely, Z. Nádasdy, G. Csibra et S. Bíró, “Taking the intentional stance at 12 months of age,” *Cognition*, vol. 56, n°. 2, p. 165–193, 1995. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/001002779500661H

[19] G. Gergely et G. Csibra, “Teleological reasoning in infancy : The infant’s naive theory of rational action : A reply to premack and premack,” *Cognition*, vol. 63, n°. 2, p. 227–233, 1997. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S0010027797000048

[20] H. M. Wellman, *The Child’s Theory of Mind*. MIT Press, 1990.

[21] N. Goodman, C. Baker, E. Bonawitz, V. K. Mansinghka, A. Gopnik, H. Wellman, L. Schulz et J. B. Tenenbaum, “Intuitive theories of mind : A rational approach to false belief,” dans *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, 2006, p. 1382–1387. [En ligne]. Disponible : http://web.mit.edu/cocosci/Papers/pos785-goodman.pdf

[22] D. Nau, M. Ghallab et P. Traverso, *Automated Planning : Theory & Practice*. Morgan Kaufmann Publishers Inc., 2004.

[23] J. García, J. E. Florez, Álvaro Torralba, D. Borrajo, C. L. López, Ángel García-Olaya et J. Sáenz, “Combining linear programming and automated planning to solve intermodal transportation problems,” *European Journal of Operational*

Research, vol. 227, n°. 1, p. 216–226, 2013. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S0377221712009496

[24] J. C. González, F. Fernández, Á. García-Olaya et R. Fuentetaja, "On the application of classical planning to real social robotic tasks," dans *Proceedings of the 5th Workshop on Planning and Robotics (PlanRob), ICAPS Conference, Pittsburgh, Pennsylvania, USA*, 2017, p. 38–47.

[25] D. P. Bertsekas, *A Course in Reinforcement Learning*. Belmont, MA, USA : Athena Scientific, 2023.

[26] S. Nikolaidis, J. Forlizzi, D. Hsu, J. Shah et S. Srinivasa, "Mathematical models of adaptation in human-robot collaboration," 2017. [En ligne]. Disponible : https://arxiv.org/abs/1707.02586

[27] J. Shah, J. Wiken, B. Williams et C. Breazeal, "Improved human-robot team performance using chaski, a human-inspired plan execution system," dans *2011 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2011, p. 29–36.

[28] S. Sreedharan, A. Kulkarni et S. Kambhampati, "Explainable human-ai interaction : A planning perspective," 2024.

[29] A. Tabrez, M. Luebbers et B. Hayes, "A survey of mental modeling techniques in human–robot teaming," *Current Robotics Reports*, vol. 1, 2020.

[30] A. Fern, S. Natarajan, K. Judah et P. Tadepalli, "A decision-theoretic model of assistance," *Journal of Artificial Intelligence Research*, vol. 50, p. 71–104, 2014.

[31] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, 3e éd. Athena Scientific, 2005, vol. I.

[32] S. Russell, "Learning agents for uncertain environments (extended abstract)," dans *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*. Association for Computing Machinery, 1998, p. 101–103. [En ligne]. Disponible : https://doi.org/10.1145/279943.279964

[33] A. Y. Ng et S. J. Russell, "Algorithms for inverse reinforcement learning," dans *Proceedings of the Seventeenth International Conference on Machine Learning*, ser. ICML '00. Morgan Kaufmann Publishers Inc., 2000, p. 663–670.

[34] D. Hadfield-Menell, A. Dragan, P. Abbeel et S. Russell, "Cooperative inverse reinforcement learning," 2024. [En ligne]. Disponible : https://arxiv.org/abs/1606.03137

[35] P. Abbeel et A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," dans *Proceedings of the Twenty-First International Conference on Machine Learning*,

ser. ICML '04. Association for Computing Machinery, 2004, p. 1. [En ligne]. Disponible : https://doi.org/10.1145/1015330.1015430

[36] S. Arora et P. Doshi, "A survey of inverse reinforcement learning : Challenges, methods and progress," 2020. [En ligne]. Disponible : https://arxiv.org/abs/1806.06877

[37] A. B. Karami, "Modèles décisionnels d'interaction homme-robot," Thèse de doctorat, Université de Caen, 2011, thèse de doctorat dirigée par Mouaddib, Abdel-Illah. [En ligne]. Disponible : http://www.theses.fr/2011CAEN2077

[38] Y. Wilks et A. Ballim, "Multiple agents and the heuristic ascription of belief," dans *Proceedings of the 10th International Joint Conference on Artificial Intelligence - Volume 1.* Morgan Kaufmann Publishers Inc., 1987, p. 118–124.

[39] P. J. Gmytrasiewicz et P. Doshi, "A framework for sequential planning in multi-agent settings," *Journal of Artificial Intelligence Research*, vol. 24, p. 49–79, jul 2005. [En ligne]. Disponible : http://dx.doi.org/10.1613/jair.1579

[40] Y. You, V. Thomas, F. Colas, R. Alami et O. Buffet, "Robust robot planning for human-robot collaboration," 2023. [En ligne]. Disponible : https://arxiv.org/abs/2302.13916

[41] D. Bernstein, R. Givan, N. Immerman et S. Zilberstein, "The complexity of decentralized control of markov decision processes," *Mathematics of Operations Research*, vol. 27, 12 2002.

[42] S. Nikolaidis et J. Shah, "Human-robot cross-training : Computational formulation, modeling and evaluation of a human team training strategy," dans *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2013, p. 33–40.

[43] S. Nikolaidis, A. Kuznetsov, D. Hsu et S. Srinivasa, "Formalizing human-robot mutual adaptation : A bounded memory model," dans *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2016, p. 75–82.

[44] S. Nikolaidis, D. Hsu et S. Srinivasa, "Human-robot mutual adaptation in collaborative tasks," *Int. J. Rob. Res.*, vol. 36, nº. 5–7, p. 618–634, 2017. [En ligne]. Disponible : https://doi.org/10.1177/0278364917690593

[45] A.-B. Karami, L. Jeanpierre et A.-I. Mouaddib, "Partially observable markov decision process for managing robot collaboration with human," dans *2009 21st IEEE International Conference on Tools with Artificial Intelligence*, 2009, p. 518–521.

[46] R. Hakli, "Cooperative human–robot planning with team reasoning," *International Journal of Social Robotics*, vol. 9, 11 2017.

[47] S. Rothfuß, M. Wörner, J. Inga, A. Kiesel et S. Hohmann, "Human–machine cooperative decision making outperforms individualism and autonomy," *IEEE Transactions on Human-Machine Systems*, vol. 53, nº. 4, p. 761–770, 2023.

[48] S. Ross, J. Pineau, S. Paquet et B. Chaib-draa, “Online planning algorithms for pomdps,” *Journal of Artificial Intelligence Research*, vol. 32, p. 663–704, juill. 2008. [En ligne]. Disponible : http://dx.doi.org/10.1613/jair.2567

[49] R. Bellman, *Dynamic Programming*. Dover Publications, 1957.

[50] R. Ahuja, K. Mehlhorn, J. Orlin et R. Tarjan, “Faster algorithms for the shortest path problem,” *Journal of the ACM*, vol. 37, p. 213–223, 1990.

[51] R. A. Howard, “Dynamic programming,” *Management Science*, vol. 12, n°. 5, p. 317–348, 1966.

[52] J. Rust, “Using randomization to break the curse of dimensionality,” *Econometrica : Journal of the Econometric Society*, p. 487–516, 1997.

[53] C. Browne, E. Powley, D. Whitehouse, S. Lucas, P. Cowling, P. Rohlfshagen, S. Tavener, D. Perez Liebana, S. Samothrakis et S. Colton, “A survey of monte carlo tree search methods,” *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 4 :1, p. 1–43, 2012.

[54] J. Rust, “Chapter 51 structural estimation of markov decision processes,” dans *Handbook of Econometrics*, ser. Handbook of Econometrics. Elsevier, 1994, vol. 4, p. 3081–3143. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S1573441205800200

[55] H. A. Simon, “A behavioral model of rational choice,” *The Quarterly Journal of Economics*, vol. 69, n°. 1, p. 99–118, 1955. [En ligne]. Disponible : http://www.jstor.org/stable/1884852

[56] C. Liu, S.-Y. Liu, E. Carano et J. Hedrick, “A framework for autonomous vehicles with goal inference and task allocation capabilities to support peer collaboration with human agents,” *ASME 2014 Dynamic Systems and Control Conference*, vol. 2, 2014.

[57] T. Ullman, C. Baker, O. Macindoe, O. Evans, N. Goodman et J. Tenenbaum, “Help or hinder : Bayesian models of social goal inference,” dans *Advances in Neural Information Processing Systems*, vol. 22. Curran Associates, Inc., 2009. [En ligne]. Disponible : https://proceedings.neurips.cc/paper_files/paper/2009/file/52292e0c763fd027c6eba6b8f494d2eb-Paper.pdf

[58] H. Kurniawati, D. Hsu et W. S. Lee, “Sarsop : Efficient point-based pomdp planning by approximating optimally reachable belief spaces,” dans *Robotics : Science and Systems IV*. The MIT Press, 2009. [En ligne]. Disponible : https://doi.org/10.7551/mitpress/8344.003.0013

[59] “Computationally feasible bounds for partially observed markov decision processes,” *Oper. Res.*, vol. 39, n°. 1, p. 162–175, 1991.

[60] L. P. Kaelbling, M. L. Littman et A. R. Cassandra, “Planning and acting in partially observable stochastic domains,” *Artificial Intelligence*, vol. 101, n°. 1, p. 99–134, 1998. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S000437029800023X

[61] R. Zhou et E. A. Hansen, “An improved grid-based approximation algorithm for pomdps,” dans *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 1*. Morgan Kaufmann Publishers Inc., 2001, p. 707–714.

[62] L. Li et M. L. Littman, “Lazy approximation for solving continuous finite-horizon mdps,” dans *Proceedings of the 20th National Conference on Artificial Intelligence - Volume 3*. AAAI Press, 2005, p. 1175–1180.

[63] A. P. Dempster, N. M. Laird et D. B. Rubin, “Maximum likelihood from incomplete data via the em algorithm,” *Journal of the Royal Statistical Society : Series B (Methodological)*, vol. 39, n°. 1, p. 1–22, 1977.

[64] S. Boyd et L. Vandenberghe, *Convex Optimization*. Cambridge : Cambridge University Press, 2004.

[65] S. Shalev-Shwartz et S. Ben-David, *Understanding Machine Learning : From Theory to Algorithms*. Cambridge University Press, 2014.

[66] B. Gao et L. Pavel, “On the properties of the softmax function with application in game theory and reinforcement learning,” 2018. [En ligne]. Disponible : https://arxiv.org/abs/1704.00805

[67] M. Chevalier-Boisvert, B. Dai, M. Towers, R. de Lazcano, L. Willems, S. Lahlou, S. Pal, P. S. Castro et J. Terry, “Minigrid & miniworld : Modular & customizable reinforcement learning environments for goal-oriented tasks,” *CoRR*, vol. abs/2306.13831, 2023.

[68] K. Leyton-Brown et Y. Shoham, *Essentials of Game Theory : A Concise, Multidisciplinary Introduction*. Morgan and Claypool Publishers, 2008.

[69] A. M. Andrew, “Reinforcement learning : An introduction by richard s. sutton and andrew g. barto.” *Robotica*, vol. 17, n°. 2, p. 229–235, 1999.

[70] B. Mavrin, S. Zhang, H. Yao, L. Kong, K. Wu et Y. Yu, “Distributional reinforcement learning for efficient exploration,” 2019. [En ligne]. Disponible : https://arxiv.org/abs/1905.06125

ANNEXE A FONCTION VALEUR ET D'ACTION DU ROBOT DANS LE CAS BUT UNIQUE

Dans cette partie de l'annexe nous présentons les étapes de développement pour retrouver la fonction Q nécessaire pour la prise de décision du robot. Rappelons la fonction objective 3.18, on peut écrire l'équation de Bellman de la façon suivante :

$$
\begin{aligned}
Q_t^R(x_t^H, w_t, a^R, g) = & \sum_{a^H \in A_H} \sum_{w' \in W} T_{w'}(w_t, a^R, w') \mathbb{P}_{a^H, w'} R^H(x_t^H, w', a^H, g) + R^R(x_t^H, w_t, a^R) \\
& + \gamma \sum_{a^H \in A^H} \sum_{x_{t+1}^H \in X^H} \mathbb{P}_{a^H, w'} T_x^H(x_t^H, w_t', g, a_t^H, x_{t+1}^H) V^R(x_{t+1}^H, w') \\
& = \sum_{w' \in W} T_{w'}(w_t, a^R, w') \Bigg[\sum_{a^H \in A^H} \mathbb{P}_{a^H, w'} R_t^H + R_t^R \\
& + \gamma \sum_{x_{t+1}^H \in X^H} \sum_{a^H \in A^H} \mathbb{P}_{a^H, w'} T_x^H(x_t^H, w', g, a^H, x_{t+1}^H) V^R(x_{t+1}^H, w_{t+1}) \Bigg] \\
& = R^R(x_t^H, w_t, a^R) + \sum_{w' \in W} T_{w'}(w_t, a^R, w') \Bigg[\sum_{a^H \in A^H} \mathbb{P}_{a^H, w'} R_t^H \\
& + \gamma \sum_{x_{t+1}^H \in X^H} \sum_{a^H \in A^H} \mathbb{P}_{a^H, w'} T_x^H(x_t^H, w', g, a^H, x_{t+1}^H) V^R(x_{t+1}^H, w') \Bigg]
\end{aligned}
$$

ANNEXE B COMPORTEMENTS DE L'HUMAIN SUIVANTS LES PARAMÈTRES DE RATIONALITÉ

Cette sous-section illustre le comportement de l'humain lorsqu'il suit le paramètre de rationalité défini pour les expérimentations présentées dans la section 4.3.1.

Indice de rationalité à 0.1

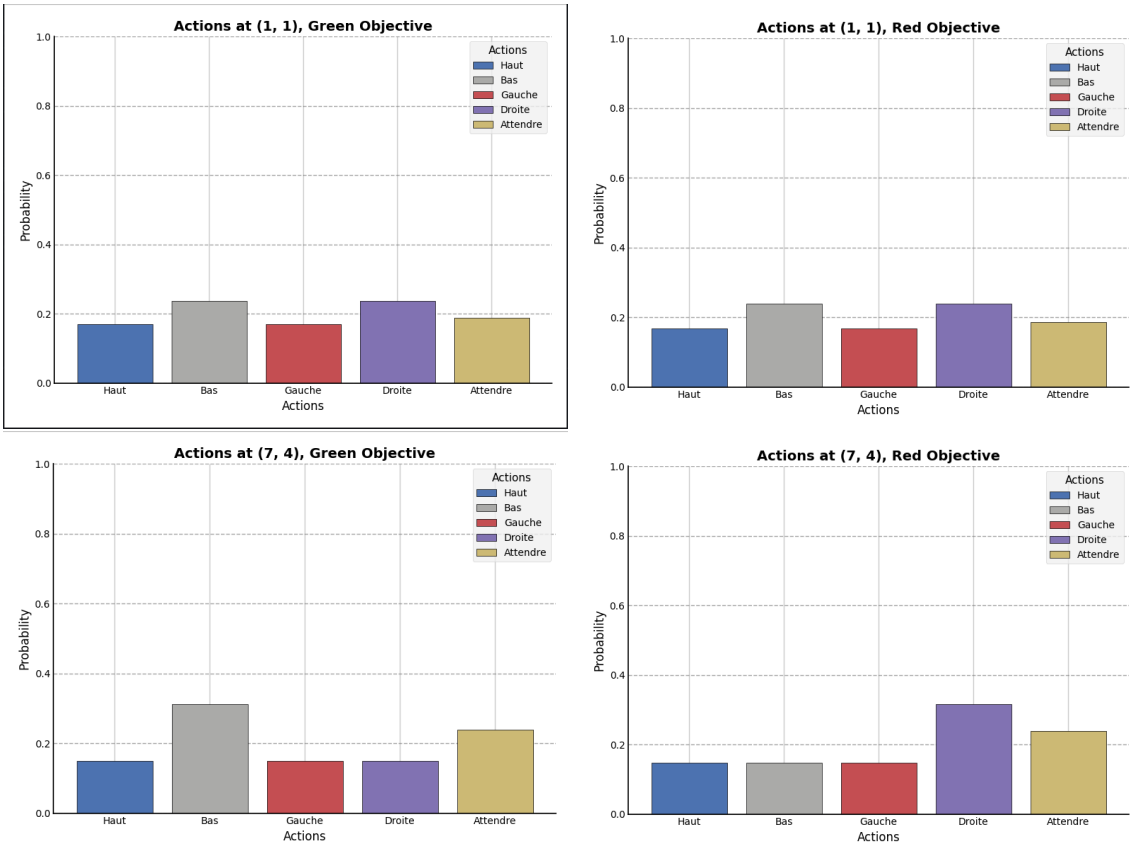


FIGURE B.1 Distribution de probabilité des différentes actions à différentes positions, $(1, 1)$ et $(4, 4)$ de la figure 4.2b, pour les deux différents objectifs du scénario 2. Les probabilités sont déterminées grâce à (3.16) avec un indice de rationalité $\beta = 0.1$ (agent quasi aléatoire).

Indice de rationalité à 2

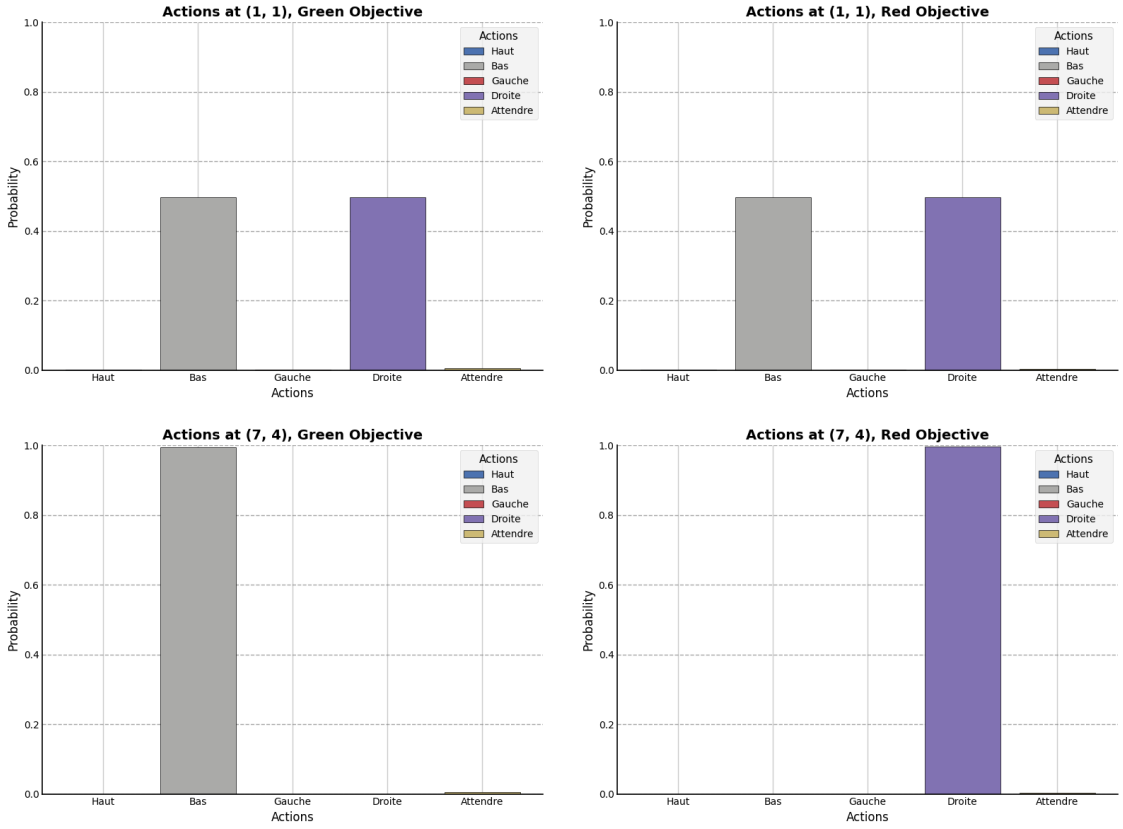
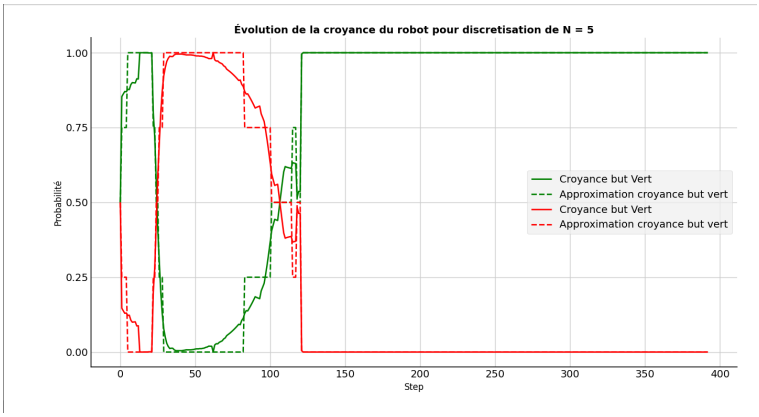


FIGURE B.2 Distribution de probabilité des différentes actions à différentes positions, $(1, 1)$ et $(4, 4)$ de la figure 4.2b, pour les deux différents objectifs du scénario 2. Les probabilités sont déterminées grâce à (3.16) avec un indice de rationalité $\beta = 2$ (agent quasi rationnel).

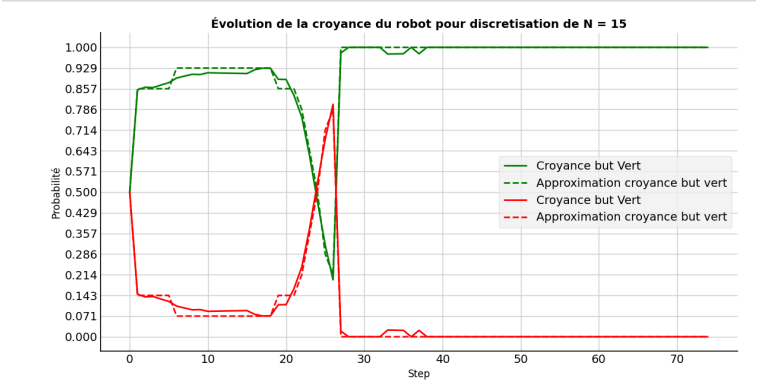
ANNEXE C ÉVOLUTION DE LA CROYANCE DU ROBOT

Modèle réel de l'humain pour indice à 0.1

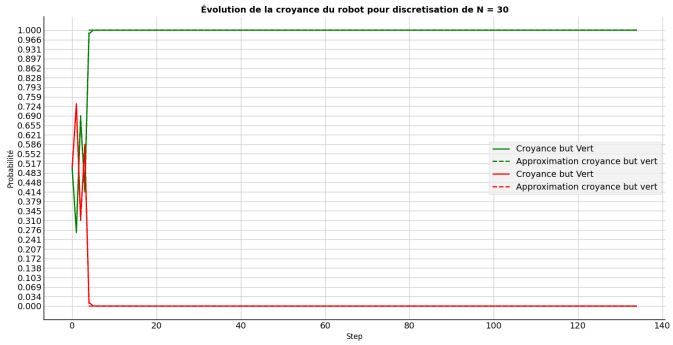
Le robot possède un modèle erroné du comportement de l'humain : il le considère comme semi-rationnel alors qu'il est en réalité irrationnel. Les figures C.1 illustrent cette divergence, montrant que pour N=5, l'approximation de la croyance ne correspond pas fidèlement à la croyance réelle calculée par le robot. Toutefois, cette approximation s'aligne progressivement avec la croyance réelle à mesure que N augmente. Par ailleurs, le robot étant sensible à chaque mouvement observé de l'humain, les courbes des graphes présentent des fluctuations marquées, reflétant ces ajustements successifs.



(a)



(b)

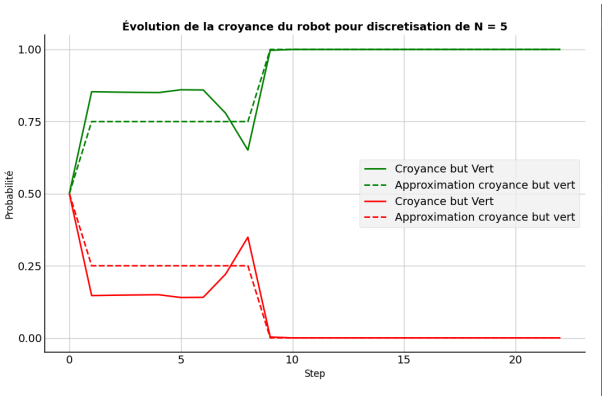


(c)

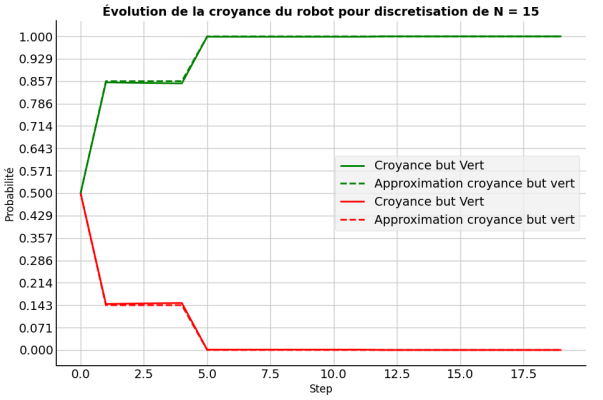
FIGURE C.1 Croyance avec et sans approximation du robot pour $\beta_R = 0.8$, alors que le paramètre réel de l'humain est $\beta_H = 0.1$. L'espace des croyances est discrétisé selon trois niveaux : (a) N=5, (b) N=15 et (c) N=30. L'objectif caché de l'humain est la case verte pour toutes les simulations.

Modèle réel de l'humain pour indice à 0.8

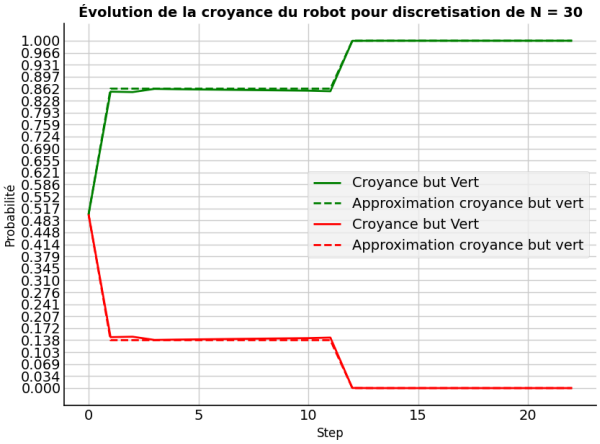
Pour cette expérimentation, le robot dispose d'un modèle exact de l'humain, ce qui se traduit par des transitions plus fluides de la croyance. Cependant, cela n'élimine pas totalement les erreurs d'approximation, comme l'illustre la figure C.2a. Ces erreurs diminuent néanmoins avec l'augmentation du paramètre de discrétisation de l'espace des croyances.



(a) Trajectoires de l'humain.



(b) Position sur la grille.



(c) Troisième figure ajoutée.

FIGURE C.2 Croyance avec et sans approximation du robot pour $\beta_R = 0.8$, alors que le paramètre réel de l'humain est $\beta_H = 0.8$. L'espace des croyances est discrétisé selon trois niveaux : (a) N=5, (b) N=15 et (c) N=30. L'objectif caché de l'humain est la case verte pour toutes les simulations.