Apprentissage par renforcement profond dans une architecture cognitive pour l'aide à la conduite de missions navales

Eva Artusi

PSL UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à MINES ParisTech

Apprentissage par renforcement profond dans une architecture cognitive pour l'aide à la conduite de missions navales

Soutenue par

Eva ARTUSI

Le 13 décembre 2021

Ecole doctorale n° 621

Ingénierie des Systèmes, Matériaux, Mécanique, Energétique

Spécialité

Sciences et génie des activités à risques

Composition du jury :

Myriam, MERAD Université Paris-Dauphine	*Présidente du jury*
Christophe, CLARAMUNT Ecole Navale	*Rapporteur*
Thomas, DEVOGELE Université de Tours	*Rapporteur*
Sébastien, LOUSTAU Université de Pau et des Pays de l'Adour	*Examinateur*
Fabien, CHAILLAN NAVAL GROUP	*Examinateur*
Aldo, NAPOLI Mines Paris	*Directeur de thèse*

Remerciements

Je tiens tout d'abord à remercier mes encadrants de thèse Aldo Napoli et Fabien Chaillan sans qui cette thèse n'aurait pas eu lieu. Tout au long de ces trois années, j'ai pu bénéficier d'un encadrement et d'une disponibilité de mes encadrants sans faille. Merci à vous d'avoir cru en mes capacités, plus particulièrement en moi et de m'avoir toujours poussée à aller plus loin. Mais surtout, je vous remercie de m'avoir laissée une grande liberté et autonomie dans mes travaux, d'avoir été à l'écoute de toutes mes propositions, idées et de m'avoir permis de m'approprier mon sujet comme je le souhaitais. Grâce à vous j'ai appris le métier de chercheur et pris en maturité. Outre sur le plan professionnel, nous avons échangé de très bons moments, discuté informellement sur différents sujets et travaillé dans la bonne humeur avec nos humours respectifs. Je tiens particulièrement à remercier Aldo pour m'avoir fait commencer l'exercice de rédaction dès la première année, cela m'a permis de progresser en rédaction, en synthèse et d'éviter un gros état de stress les six derniers mois.

Je tiens ensuite à remercier l'ensemble des membres du jury d'avoir accepté de prendre part à l'évaluation de cette thèse. La diversité de leurs horizons scientifiques permet de valider, d'améliorer ou de critiquer la réflexion proposée par des points de vue variés. Pour cela, j'adresse un profond merci à Christophe Claramunt et Thomas Devogele pour avoir été rapporteurs. Je remercie également Myriam Merad et Sébastien Loustau d'avoir été les examinateurs de cette thèse.

Merci à tous mes collègues de travail de NAVAL GROUP pour tous les bons moments partagés au travail, au café, en séminaire et autres. J'y ai rencontré des personnes de bonne humeur, avec une mention spéciale pour l'humour de certaines, ouvertes d'esprit qui m'ont soutenue, donné des conseils, aidée scientifiquement et apporté de la joie durant ses trois années. Certaines sont devenues des amis et se reconnaitront. Parmi mes collègues de travail, je remercie particulièrement Romain L, Laurent P, Michel C, Alexis B, Rozenn C, Chantal M, Thomas C, Estelle C, Alexandre Ga, Bénédicte L, Jean-Michel V et toutes les personnes qui se sont portées volontaire pour participer aux expérimentations de ma thèse. Romain L a implémenté entièrement l'IHM proposée dans ma thèse, il a été un fort contributeur et a toujours été disponible pour m'aider ; sans lui mes travaux de thèse seraient moins riches. Laurent P et Michel C m'ont apporté le côté opérationnel dont j'avais besoin, ils ont toujours été disponibles pour répondre à mes questions et m'ont été d'une grande aide pour mener à bien ma thèse. Alexandre Ga m'a transmis une partie de sa culture et de son expérience sur la lutte contre menace asymétrique me permettant d'avancer dans mes travaux. Thomas C m'a permis de trouver des solutions aux problèmes rencontrés lors de l'implémentation des algorithmes de Deep Reinforcement Learning et Estelle C m'a aidée dans mes recherches et ma compréhension sur l'optimisation de trajectoire. Un grand merci à Alexis B, qui m'a encouragée, m'a aidée dans mes travaux de thèse, a voyagé en conférence avec moi, m'a fait partager son humour, mais surtout m'a permis d'embarquer trois jours sur le Chevalier Paul pour interroger les opérationnels en situation. Merci aux deux personnes du service Facteurs Humains, Rozenn C et Chantal M, qui m'ont aidée dans l'élaboration de mon protocole expérimental et de validation. Je tenais aussi à remercier Bénédicte L qui sur son temps personnel a réalisé ma slide présentée à MT180 et une superbe présentation Prezi lors de ma participation au prix Pierre Laffitte 2020. Merci aussi à Jean-Michel V qui a relu et validé mon manuscrit de thèse en interne pour l'envoi aux rapporteurs. Pour terminer, je remercie Gregory B, Aymeric B et

l'ANRT pour avoir financé ma thèse et m'avoir permis de l'effectuer au sein de NAVAL GROUP.

Merci aussi aux personnes du CRC avec qui je n'ai malheureusement pas pu échanger autant que prévu à cause de la crise sanitaire. Je remercie Franck Guarnieri, directeur du centre, pour m'avoir permis de faire ma thèse au CRC. Merci au CRC de m'avoir permis de participer même succinctement à l'encadrement des élèves des Mines de Paris de première année lors du MIG Forensic Engineering. Je tiens à remercier plus particulièrement Sandrine Renaux pour sa gentillesse, son soutien et son aide lors des démarches administratives ainsi que Myriam Lavigne-Perrault. Merci aussi à Corinne Matarasso d'avoir réalisé des articles et posters sur moi.

Merci à « ma famille de la danse » qui m'a soutenue grâce à la danse et plus particulièrement à Nadia Bourguet-Delpy qui m'a permis d'évacuer une grande partie de mon stress durant ses trois années.

Et bien sûr merci à mes parents Chantal et Thierry, qui m'ont énormément soutenue dans les moments difficiles. Ils ont toujours su m'écouter, m'épauler et m'apporter leurs conseils. Ils sont et ont toujours été là pour moi, leur soutien est sans faille et je peux compter sur eux les yeux fermés. Merci d'avoir supporté mes états de stress, d'angoisse souvent très communicatifs, mes doléances, de m'aider à tenir en place, à prendre du recul sur les choses et d'avoir activement participé à ma thèse en me filmant, me faisant répéter, …. Sans leur soutien initial et l'expérience de mon frère, je ne me serais sûrement jamais lancée dans une thèse.

Et pour finir, un grand merci à B qui se reconnaitra qui a souvent lu ma prose, corrigé mon anglais parfois médiocre, écouté mes répétitions et évidemment supporté mon tempérament (pas très calme). Il a presque supporté tout au long de ces trois ans mes états d'âme et particulièrement la dernière année. Son soutien a été presque sans faille tout au long de cette thèse.

Ces trois années auront été riches en émotions avec des « hauts et des bas », mais je sors grandie de cette expérience que je ne regrette pas.

« *Le travail paie toujours* » Thierry et Chantal Artusi

« *Science sans conscience, n'est que ruine de l'âme* » François Rabelais

«Vi Cha Se !» Nadia Bourguet Delpy alias Sotcha

« C'est le destin ! » Claudine Artusi

Table des matières

Liste des acronymes :

ANRT	Association Nationale de la Recherche et de la Technologie
AHC-learning	Adaptative Heuristic Critic - learning
AIS	Automatic Identification System
API	Application Programming Interface
AR	Apprentissage par renforcement
BDI	Belief Desires Intention
BRDM	Bayesian Recognitional Decision Model
CA	Composite Agent
COLREG	Convention on the International Regulations for Preventing Collisions at Sea
CPA	Closest Point of Approach
DBP	Décisions à Base de Patrons
DQN	Deep-Q-Network
DRL	Deep Reinforcement Learning
ECMWF	European Centre for Medium-Range Weather Forecasts
EM	Etat-Major
ENL	Equipement Non Létal
FH	Facteurs Humains
HRL	Hierarchical Reinforcement Learning
IA	Intelligence Artificielle
IHM	Interface Homme Machine
LCMA	Lutte Contre Menace Asymétrique
LTM	Long Term Memory
MEDOT	Méthode d'Elaboration d'une Décision Opérationnelle Tactique
MMCAD	Méthodologie Multicritère d'Aide à la Décision
MOE	Measure Of Effectiveness
NDM	Natural Decision Making
OODA	Observe-Orient-Decide-Act
PA	Plan d'actions
PDM	Processus de Décision Markovien
PDMPO	Processus de Décision Markovien Partiellement Observable
POO	Programmation Orientée Objet
PPO	Proximal Policy Optimization
RA	Reactive Agent
RAV	Résistance à l'avancement
R-CAST	Collaborative Agents for Simulating Teamwork
RL	Reinforcement Learning
ROE	Rules Of Engagement

RPD	Recognition Primed Decision-Making
RPgD	Recognition Primed group Decision
RPM	Recognition Planning Model
SAD	Système d'Aide à la Décision
SCA	Symbolic Constructor Agent
SMA	Système Multi-Agents
TRANS	Tractable Role Agent prototype for concurrent Navigation Systems

Liste des figures :

Liste des tableaux :

Introduction

Ce travail de thèse s'articule autour de la notion **d'aide à la décision dédiée au niveau tactique du commandement militaire**. A ce niveau de décision, l'organe de commandement dirige un groupe d'individus afin d'accomplir une opération en fournissant un but, une direction et une motivation. Ce commandement est composé d'une ou plusieurs personnes ayant la capacité de reconfigurer rapidement une approche dans des situations souvent très complexes.

Dans une première partie, le contexte applicatif de la thèse est introduit avec les spécificités des missions navales militaires et la définition de leur réalisation. Dans une deuxième partie, les notions de prise de décision, d'aide à la décision et de systèmes d'aide à la décision sont définies. Dans une troisième partie, la problématique et les objectifs de la thèse sont explicités et pour terminer, la démarche méthodologique de la thèse et sa structure sont présentées.

Contexte applicatif : Spécificités et réalisation des missions navales militaires

Cette première partie introduit les spécificités du domaine et le déroulement d'une mission navale militaire afin de situer le contexte applicatif de la thèse et d'introduire les problématiques associées.

Spécificités des missions navales militaires

Le but premier de la Marine nationale est de contrôler et protéger son espace maritime dans ses trois dimensions, sous, sur et au-dessus de la mer. Pour cela, il existe cinq types de missions navales militaires [1] :

1. **Le Renseignement** : collecte, analyse et diffusion des renseignements sur la situation maritime mondiale.
2. **La Prévention** de crises qui peuvent menacer le territoire national.
3. **L'Intervention** en zone de conflit pour rétablir la paix, évacuer ou assister les populations.
4. **La Protection** : sauvetage, assistance aux navires, lutte contre la piraterie, la pollution et les trafics.
5. **La Dissuasion** : protection des intérêts vitaux de la France en maintenant en permanence en mer un sous-marin nucléaire lanceur d'engins.

Pour assurer ces différentes missions, la Marine nationale s'organise autour de quatre forces : la force d'action navale, la force aéronautique navale, la force océanique stratégique et la force des fusiliers marins et commandos marine. **Seule la force d'action navale concerne le sujet de thèse.**

La force d'action navale regroupe les bâtiments de surface, avec 10 000 marins et 72 bâtiments, elle fournit le cœur de la contribution de la Marine aux missions de prévention et de protection.

Dans le cadre de cette thèse nous rappelons que l'aide à la conduite de missions ne se fera qu'au niveau tactique.

Réalisation d'une mission navale militaire au niveau tactique

Les missions navales militaires sont décidées par l'Etat-Major de la Marine et porte sur trois niveaux, le niveau stratégique, opérationnel et tactique. Le niveau stratégique désigne le niveau le plus important où un dialogue s'opère au plus haut rang de l'Etat entre les responsables politiques, diplomatiques et militaires. Le niveau opérationnel, sous la responsabilité du Commandant du théâtre des opérations, planifie et conduit la campagne militaire interarmées qui répond aux objectifs fixés par le niveau stratégique. Le niveau tactique correspond à une opération ou une action, limitée dans le temps et/ou l'espace, planifiée et conduite par l'Etat-major (EM), qui regroupe les officiers chargés d'assister le Commandant.

L'accomplissement d'une mission navale militaire se déroule selon trois phases, (1) une phase de planification, (2) une phase de conduite et (3) une phase de restitution, introduites dans la suite pour des missions tactiques où seule la force d'action navale est concernée. La réalisation d'une mission tactique se décompose en deux tâches principales : le choix de la route optimale adaptée aux objectifs, contraintes et buts fixés de la mission et la gestion des événements indésirables notamment les menaces attendues.

Planification d'une mission navale militaire au niveau tactique

La planification d'une mission tactique d'un bâtiment de surface se fait à bord, par tous les officiers. Les officiers de la Marine Nationale utilise la Méthode d'Elaboration d'une Décision Opérationnelle Tactique (MEDOT) [2] qui est une méthode militaire française pour produire des ordres opérationnels au niveau tactique. Cette méthode se décompose en deux phases principales : la première est la phase préalable permettant d'établir le contexte de la mission et le cadre général de l'action et la seconde correspond à l'élaboration du plan d'actions en étudiant les modes d'actions ami et ennemi. Un mode d'actions désigne une séquence d'actes dans l'espace et dans le temps afin d'accomplir les différentes tâches. Les modes d'actions amis déterminent les plans d'actions qui vont permettre de réaliser la mission alors que ceux ennemis s'opposent à l'accomplissement de la mission. Le plan d'action est l'aboutissement de la MEDOT et il est directement impacté par la philosophie du Commandant. Selon la personnalité du Commandant et sa vision de la réalisation d'une mission, le plan d'action pourra être différent.

Conduite d'une mission navale militaire au niveau tactique

Une fois la phase de planification terminée, l'EM passe à la phase de conduite de la mission. Lors de la conduite, l'EM va s'interroger sans cesse sur la faisabilité du ou des plans d'actions définis lors de la planification, en fonction de l'évolution du théâtre des opérations en temps réel. Si le plan d'actions n'est pas réalisable, l'EM réapplique certaines étapes de la MEDOT pour faire de nouvelles propositions au Commandant qui prendra la décision finale. Pour déterminer la faisabilité, l'organe de commandement applique lors de la conduite, le processus d'évaluation [3].

Le processus d'évaluation est la détermination des progrès réalisés en vue d'accomplir une tâche, de créer un effet ou d'atteindre un objectif [3]. L'évaluation aide le Commandant à déterminer les progrès réalisés en vue d'atteindre l'état final souhaité, d'atteindre les objectifs et d'exécuter les tâches. Elle consiste également à surveiller et à évaluer continuellement l'environnement opérationnel afin de déterminer les changements qui pourraient avoir une incidence sur la conduite des opérations. La MEDOT et le processus d'évaluation sont donc complémentaires, le processus d'évaluation permet de déterminer à quel point le plan d'actions initial est faisable et en fonction de la faisabilité à partir de quelle étape l'organe de commandement va réappliquer la MEDOT.

Le processus d'évaluation se décompose en trois étapes. La première étape est l'étape de surveillance de la situation tactique en temps réel permettant à l'organe de commandement de recueillir des informations importantes et nécessaires. Ces informations peuvent concerner le terrain, la météo mais aussi les anomalies maritimes telles que les attaques pirates ou plus généralement les comportements dangereux.
La deuxième étape est l'étape d'évaluation qui a pour but d'analyser les informations recueillies lors de l'étape de surveillance pour évaluer le progrès des opérations. Elle aide les commandants à déterminer ce qui fonctionne ou pas et à tirer des enseignements sur la façon de mieux accomplir la mission. Des indicateurs sous forme de mesures d'efficacité (Measure Of Effectiveness : MOE) sont utilisés pour évaluer l'impact des opérations et mesurer les changements. Ces indicateurs sont choisis et construits par les Commandants.
La dernière étape est celle de redirection d'actions d'amélioration, où le Commandant peut décider de poursuivre l'opération comme prévu, d'ajuster le plan d'actions initialement prévu ou de le modifier significativement en demandant à l'EM de réappliquer certaines étapes de la MEDOT.

Un processus d'évaluation efficace repose à la fois sur des indicateurs quantitatifs fondés sur l'observation et qualitatifs fondés sur l'opinion. Un indicateur est un outil d'évaluation et d'aide à la décision et le jugement humain fait partie intégrante de l'évaluation, il a pour but d'identifier l'information sur laquelle se concentrer. **Des aides à la décision sont donc déjà utilisées mais elles sont subjectives, non informatisées et non automatisées.** Il est donc beaucoup plus difficile de déterminer quelles actions impliquent un succès en raison des nombreuses interactions dynamiques parmi les forces alliées, l'ennemi qui s'adapte, les populations et d'autres aspects de l'environnement opérationnel.

En conduite l'environnement opérationnel tactique est donc incertain avec une nécessité de prendre des décisions rapidement sans qu'aucune aide à la décision informatisée et automatisée n'existe. Ainsi dans la suite, les notions de prises de décision en environnement

incertain et contraint par le temps, d'aide à la décision et de systèmes d'aide à la décision sont introduites. L'aide à la décision proposée dans cette thèse se concentrera sur les étapes de surveillance et d'évaluation du processus d'évaluation afin de préconiser au Commandant des actions à entreprendre. L'étape de redirection d'actions d'amélioration sera exécutée par le Commandant qui aura le choix entre suivre la préconisation proposée ou faire d'autres choix.

Prise de décision humaine en environnement contraint, aide à la décision et systèmes d'aide à la décision

Prise de décision humaine en environnement incertain et contraint par le temps

La prise de décision humaine a été initialement étudiée et définie comme le résultat d'une évaluation minutieuse d'options alternatives en termes de probabilité et de valeur des résultats associés à ces options. Beaucoup de recherches se sont concentrées sur les violations systématiques de ces principes de rationalité afin de comprendre les processus cognitifs, c'est-à-dire l'ensemble des processus mentaux qui permettent à un individu d'acquérir, de traiter, de stocker et d'utiliser des informations ou des connaissances, qui sous-tendent le jugement et la prise de décision humaine [4].

La prise de décision en environnement incertain et contraint par le temps est aujourd'hui définie par les travaux de Klein [5]. Ses études sur les sapeurs-pompiers et le personnel d'intervention d'urgence ont révélé que, contrairement aux prédictions des modèles normatifs, ces individus ne conçoivent pas de solutions alternatives qu'ils comparent entre elles, mais utilisent leur diagnostic de la situation pour construire très rapidement une solution de cas optimal, dont ils simulent ensuite l'exécution dans leur esprit pour voir si elle correspond bien à ce qu'ils imaginent : c'est le concept de décision amorcée par la reconnaissance (Recognition Primed Decision making : RPD). Dans la phase suivante, lorsque la solution est sélectionnée, les décideurs utilisent la simulation mentale pour valider si la situation répond à leurs attentes et prendre d'autres décisions lorsqu'ils voient se développer un événement inattendu. Ainsi, le RPD suggère un processus de décision en deux étapes :

1. La reconnaissance de schémas déjà vus.
2. La simulation mentale.

Le RPD est le modèle reproduisant le mieux la prise de décision humaine dans des environnements incertains et contraints par le temps [6].

Aide à la décision

Comme évoqué précédemment, la prise de décision en environnement militaire se réalise la plupart du temps dans des situations avec de fortes contraintes de temps et dans des environnements incertains. Afin de faciliter la prise de décision des décideurs en situation, des aides à la décision ont été mises au point. L'aide à la décision vise à construire à partir de

méthodes, de concepts ou de moteurs de règles des réponses permettant d'assister l'humain dans l'élaboration de sa réflexion, qui précède sa décision. Roy a introduit la définition suivante de l'aide à la décision [7] :

« L'aide à la décision est l'activité de celui qui, prenant appui sur des modèles clairement explicités, mais non nécessairement complètement formalisés, aide à obtenir des éléments de réponse aux questions que se pose un intervenant dans un processus de décision, éléments concourant à éclairer la décision et normalement à prescrire, ou simplement à favoriser, un comportement de nature à accroître la cohérence entre l'évolution du processus d'une part, les objectifs et le système de valeurs au service desquels cet intervenant se trouve placé d'autre part ».

Processus d'aide à la décision

Quel que soit le domaine, le processus d'aide à la décision se décompose en huit étapes [8] illustrées sur la Figure 0-1 ci-dessous.

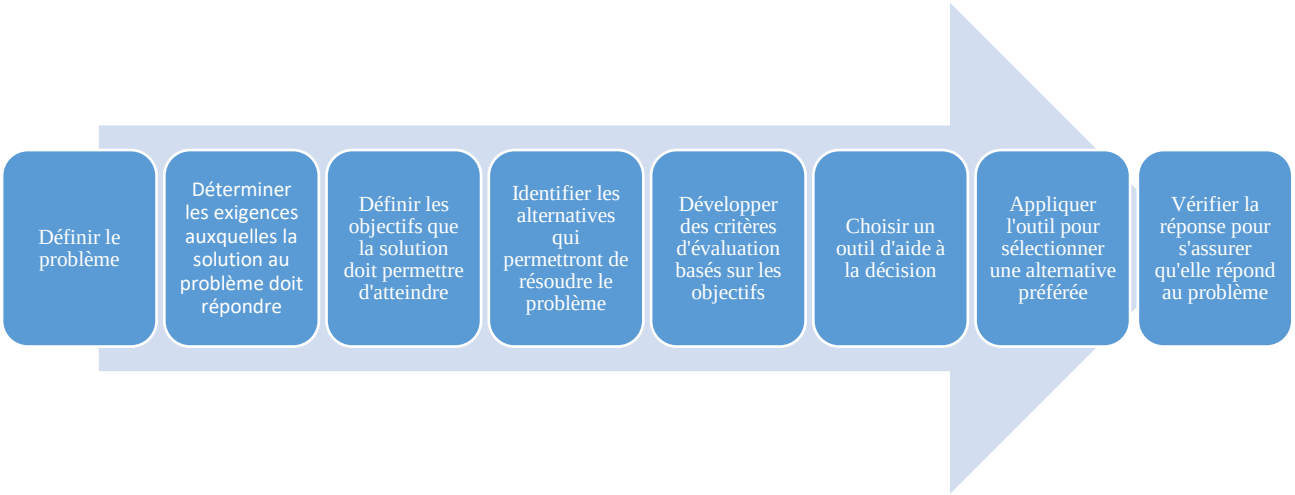

Figure 0-1 : Processus général d'aide à la décision [8]

Le processus d'aide à la décision commence avec la définition du problème. L'objectif est d'exprimer le problème dans un énoncé clair, qui doit être compris par tous les acteurs, décrivant à la fois les conditions initiales et les conditions souhaitées. Le(s) décideur(s) et le personnel d'appui s'accordent sur un énoncé de problème écrit afin de s'assurer qu'ils soient tous d'accord avant de passer aux étapes suivantes. Il s'en suit la détermination des exigences auxquelles la solution doit répondre et les objectifs à atteindre. Les décideur(s) évaluent ensuite les exigences et les objectifs pour proposer des alternatives répondant aux exigences et satisfaisant le plus grand nombre possible d'objectifs. En général les alternatives seront comparées entre elles pour sélectionner la plus adaptée au problème. Pour pouvoir les comparer, des critères d'évaluation sont définis en tant que mesures des objectifs afin d'évaluer le caractère atteignable de chaque alternative. Un outil d'aide à la décision adapté doit ensuite être choisi. Ces outils sont des processus rationnels permettant d'appliquer une réflexion critique aux informations, aux données et à l'expérience afin de prendre une décision équilibrée lorsque le choix entre les différentes options n'est pas clair. Parmi les outils d'aide à la décision, il peut

être cité par exemple, les matrices de décision issues de la Méthodologie Multicritère d'Aide à la Décision (MMCAD)[8], les arbres de décision ou encore l'analyse coût-bénéfice [9]. Une fois l'outil d'aide à la décision choisie, les alternatives peuvent être évaluées. Les critères d'évaluation peuvent être pondérés et utilisés pour évaluer les alternatives. Des analyses d'incertitude peuvent être utilisées pour améliorer la qualité du processus de sélection. Une fois la solution choisie, il est vérifié que celle-ci résout réellement le problème identifié. Une solution finale doit répondre à l'état souhaité, satisfaire les exigences et atteindre au mieux les objectifs.

Dès les années 1960, les processus d'aide à la décision ont pu être informatisés, permettant aux décideurs de résoudre des problèmes plus complexes et avec plus de données à traiter, marquant l'émergence des systèmes d'aide à la décision.

Systèmes d'aide à la décision

Un système d'aide à la décision (SAD) est un système d'information qui soutient les activités de prise de décision. Les SAD sont des systèmes et sous-systèmes informatiques interactifs destinés à aider les décideurs à utiliser les technologies de communication, les données, les documents, les connaissances et les modèles pour mener à bien les tâches liées au processus décisionnel [10]. Un SAD est un moyen de rassembler, analyser et synthétiser les données et de prendre des décisions de qualité sur la base de celles-ci, car la prise de la bonne décision est généralement basée sur la qualité des données et la capacité à les analyser.

Il existe trois modèles de SAD, les modèles passifs, actifs et coopératifs. Les modèles passifs regroupent les applications se contentant de collecter des données et de les organiser, aucune décision spécifique n'est suggérée. Un système actif, quant à lui, traite des données et présente explicitement des solutions basées sur celles-ci. Un système coopératif est un système dans lequel les données sont collectées, analysées et ensuite fournies à une composante humaine qui peut aider le système à les réviser ou les affiner.

Les progrès dans les techniques d'intelligence informatique ont permis l'essor des systèmes intelligents d'aide à la décision [6]. L'intelligence informatique est utilisée dans l'aide à la décision pour aider le décideur à sélectionner des actions en temps réel, à résoudre des problèmes de décision stressants dans des conditions incertaines, à réduire sa surcharge d'informations et lui fournir une réponse dynamique à l'aide d'agents intelligents. En intelligence artificielle (IA), un agent intelligent désigne une entité autonome qui agit en dirigeant son activité vers la réalisation d'objectifs dans un environnement en utilisant l'observation par des capteurs et des actionneurs conséquents [11]. Les agents intelligents peuvent également apprendre ou utiliser des connaissances pour atteindre leurs objectifs. Ces agents intelligents reposent par exemple sur la logique floue [12], les réseaux de neurones artificiels [13], l'apprentissage par renforcement [14] ou encore les algorithmes génétiques [15].

Maintenant que les notions de prises de décision en environnement incertain et contraint par le temps, d'aide à la décision et de systèmes d'aide à la décision ont été définies, la section suivante introduit la problématique et les objectifs de recherche de la thèse.

Problématique et objectifs de recherche

Comme évoqué précédemment, les bâtiments de surface de la Marine nationale sont amenés à accomplir différentes missions comme la surveillance de zones d'intérêt, la protection des infrastructures, la participation à des exercices ou encore l'intervention sur un théâtre d'opération. La conduite de ces missions nécessite d'intégrer simultanément un bon nombre d'informations relatives au navire et à l'environnement dans lequel il évolue. L'organe de commandement d'un navire doit analyser ces informations en très peu de temps afin de prendre les décisions adaptées à la situation en vue de mener à bien la mission. De manière générale, en fonction de sa position, de son état, de son environnement, de ses objectifs, le Commandant du navire évalue en temps réel le succès de la mission à l'aide d'indicateurs construits, choisis et calculés par l'humain muni de son discernement et de son expertise.

Cependant, **les capacités humaines, soumises à la surcharge cognitive et à leur propre « rationalité limitée », ne suffisent plus à traiter de manière fiable et rapide toutes les données relatives aux navires et à leurs environnements respectifs et recueillies par de nombreux capteurs, dans des conditions de plus en plus stressantes et contraintes par le temps.** En d'autres termes, la complexité des situations tactiques augmente alors que la contrainte opérationnelle de réalisation de la mission, rapidité et efficacité, demeure. La prise de décision dépend de plusieurs caractéristiques de la situation telles que :

- **Le temps disponible,**
- **Les informations disponibles** : trop d'informations peut ralentir le processus de planification et de conduite, tandis que pas assez peut augmenter le risque et l'incertitude.
- **L'incertitude** : l'élément « inconnu » est présent dans toutes les situations, rendant les décisions plus difficiles à prendre. Analyser l'incertitude, c'est-à-dire caractériser le manque d'information et son impact sur le résultat final peut aider les décideurs à décider mieux.
- **Le risque** : l'avenir est incertain et chaque décision peut mener à de multiples résultats. Le risque est lié à la possibilité que des résultats négatifs se produisent, et les processus de réflexion d'un décideur peuvent être influencés par le degré de risque qu'il est prêt à accepter [16].
- **Les coûts et conséquences** : les résultats positifs et négatifs d'une action [17].

Les décideurs sont donc confrontés aujourd'hui en mission à des situations incertaines dans lesquelles ils manquent de temps et de ressources. Pour faire face à cette situation, la conception et le développement d'un système d'aide à la décision permettrait grâce à la combinaison des sciences cognitives et de la puissance de calcul des moyens informatiques modernes de réduire les temps (1) d'analyse de la situation et (2) de sélection d'une action appropriée à la situation.

L'objectif principal de cette thèse est de formaliser une aide à la décision dans le champ de la conduite de missions navales militaires tactiques.

L'aide permettra de préconiser des décisions à prendre en temps contraint pour réussir la mission (ex : changement de route imprévu, réponse à une menace). L'organe de

commandement aura alors la possibilité de suivre la décision proposée ou de la refuser, auquel cas l'aide s'adaptera en surveillant la nouvelle situation et proposera en temps réel de nouvelles décisions.

La problématique et les objectifs de recherche étant énoncés, la section suivante introduit la démarche méthodologique adoptée dans cette thèse.

Proposition de recherche et approche adoptée

La problématique générale est d'élaborer une méthodologie afin de concevoir et développer un système d'aide à la décision pour la conduite de missions navales militaires tactiques. Pour cela, la thèse se divisera en plusieurs étapes :

- **Conception du système d'aide à la conduite de missions navales**

Différentes méthodes existent pour aider les décideurs et plus particulièrement l'organe de commandement à prendre des décisions. Elles sont parfois composées d'indicateurs généralement sous la forme d'un ratio bénéfice/risque permettant aux décideurs d'analyser les décisions qui s'offrent à lui et de les départager. Après un état de l'art sur les différentes méthodes existantes, une méthodologie de prise de décision sera choisie pour conceptualiser le système d'aide à la conduite de missions navales. En sortie du système, la préconisation des décisions se fondera sur une série de scénarios possibles. Ils seront caractérisés sous la forme d'actions qui auront été recensées, accompagnées éventuellement d'indicateurs qui seront proposés aux décideurs. Les indicateurs correspondront à ceux utilisés par le Commandant afin qu'il puisse analyser la pertinence des scénarios proposés.

- **Sélection des méthodes et des technologies d'aide à la décision**

Les décisions préconisées par le système seront déterminées à l'aide d'agents intelligents ayant la capacité de s'adapter aux données du navire et de son environnement en temps réel. Ces agents ou algorithmes interagiront avec l'environnement dans lequel évolue le navire pour trouver et proposer la solution optimale à un ou des objectifs donnés. Les algorithmes devront s'adapter et respecter la méthodologie appliquée pour la conduite de missions. Les algorithmes répondant au mieux à ces exigences sont les algorithmes d'apprentissage par renforcement. L'apprentissage par renforcement ou Reinforcement Learning issu de l'IA permet à un agent informatique d'apprendre des conséquences de ses actions en interagissant avec un environnement dynamique. Astreint au principe de la sanction et de la récompense, le modèle apprend de chacune de ses actions. Après apprentissage, il sera capable de répondre de façon autonome à des situations inédites et de préconiser en tant qu'aide au commandement des actions et des décisions adaptées, de sorte que le navire atteigne l'objectif de sa mission [18]. De plus l'apprentissage par renforcement permet de reproduire la majeure partie du processus cognitif de prise de décision humaine [19].

- **Prototypage de l'architecture cognitive pour l'aide à la décision**

Le prototypage de l'aide à la décision reposera :

(1) Sur l'implémentation du ou des algorithmes d'apprentissage par renforcement qui viseront à apprendre à un agent, représentant un navire dans notre cas, à réagir dans un environnement dynamique, l'environnement du navire, en fonction d'un ou plusieurs objectifs de la mission, tout en respectant la méthodologie préalablement choisie. Ce ou ces algorithmes permettront grâce à la modélisation du navire et de son environnement de surveiller la situation en temps réel et de proposer en fonction de son évolution et des objectifs de la mission, des couples indicateurs, actions. Selon la complexité de la mission des fonctionnalités supplémentaires seront proposées.

(2) Sur une interface Homme-Machine qui proposera une représentation des scénarios à l'organe de commandement.

- **Evaluation de l'architecture cognitive**

L'évaluation se fera dans un premier temps sur des scénarios de missions réalistes en condition simulée. Dans une première phase, les préconisations de l'architecture cognitive seront comparées aux résultats d'un algorithme robuste et éprouvé. Dans une seconde phase, des expérimentations seront conduites avec des personnes ayant une expérience opérationnelle de niveaux différents afin d'évaluer (1) l'influence du système d'aide à la décision sur le type de participant et (2) la performance des équipes assistées de l'aide à la décision lors de la conduite de missions navales. L'évaluation du système se fera avec deux groupes (G1 et G2). G1 sera constitué de participants tandis que G2 sera constitué de participants différents du G1, couplés au système d'aide à la décision. Les participants de chaque groupe réaliseront un même nombre de tâches et seuls les participants prendront les décisions. Afin de comparer les performances entre les deux groupes et d'évaluer le système d'aide à la décision, des critères de comparaison et de validation, c'est-à-dire des grandeurs permettant de comparer les performances des équipes et des seuils devant être atteints pour valider le système, seront préalablement définis et établis avec les opérationnels et experts du domaine. Ces expérimentations nous permettront d'évaluer le système et d'identifier les pistes d'amélioration.

Structure de la thèse

En introduction dans une première partie, le contexte applicatif de la thèse est présenté avec les spécificités et la réalisation des missions navales militaires ; suite à cela la problématique générale et les objectifs de la thèse sont introduits. Dans une troisième partie, les notions de prise de décision humaine, d'aide à la décision et de systèmes d'aide à la décision sont définies et pour terminer la démarche méthodologique adoptée est décrite.

Le Chapitre 1 de cette thèse se consacre à un état de l'art sur la prise de décision au sein des missions navales militaires. La première partie présente la prise de décision humaine avec les deux familles principales de modèles existantes. Pour terminer cette première partie une comparaison entre ces deux familles est faite. La seconde partie porte sur les modèles de prises de décision utilisés pour les missions navales militaires au niveau tactique. Tout d'abord, la méthodologie utilisée par la Marine Nationale est décrite en détail. Une deuxième partie décrit les méthodologies issues du Natural Decision Making (NDM) ou prise de décision en situation naturelle et particulièrement le Recognition Primed Decision (RPD) ou prise de décision

amorcée par la reconnaissance utilisé pour l'aide à la conduite de missions. Une troisième partie énonce les limites du RPD.

Le Chapitre 2 présente dans une première partie un état de l'art relatif aux SAD fondés sur le RPD pour venir en aide aux décideurs. Suite aux limites des SAD existants, l'apprentissage par renforcement est introduit dans une seconde partie et proposé comme méthode de préconisation de décisions de notre système d'aide à la conduite de missions.

Dans le Chapitre 3, nous présentons notre proposition de système d'aide à la décision pour la conduite de missions navales fondé sur le RPD et l'apprentissage par renforcement profond. Ce chapitre explicitera particulièrement la modélisation de l'architecture cognitive de notre système d'aide à la décision.

Dans le Chapitre 4, le prototypage du système d'aide à la décision, et plus particulièrement de l'architecture cognitive mise en œuvre pour les actions préconisées à l'organe de commandement est décrit. Un premier exemple d'IHM développée pour l'évaluation de l'aide à la conduite de missions est également présenté.

Dans le Chapitre 5 nous exposerons les premiers éléments d'évaluation des préconisations de l'aide à la décision qui ont été comparées (1) aux résultats d'un algorithme de calcul de route optimale et (2) aux résultats obtenus d'une campagne d'expérimentations qui a mobilisé d'anciens opérationnels et des personnes ayant de l'expérience opérationnelle.

Dans Conclusion et perspectives nous reviendrons sur les apports de cette thèse au domaine de l'aide à la décision pour la conduite de missions navales militaires. Nous présenterons alors plusieurs perspectives d'amélioration de notre travail de recherche.

Chapitre 1 : Prise de décision au sein des missions navales militaires

Ce premier chapitre présente tout d'abord un état de l'art sur les modèles de prise de décision utilisés par la Marine Nationale pour accomplir une mission tactique. Cet état de l'art va nous permettre de faire un choix de méthodologie pour la conception de notre aide à la décision. Ce choix s'orientera vers des méthodes de prise de décision en situation naturelle (NDM - Natural Decision Making), une famille de modèles intuitifs. Puis une deuxième partie expliquera plus en détail les modèles analytiques et intuitifs afin d'introduire leur fonctionnement, de les comparer et de justifier notre choix de fonder notre système d'aide à la décision sur le modèle de prise de décision amorcée par la reconnaissance de la situation (RPD - Recognition Primed Decision), un modèle intuitif. Pour terminer une dernière partie introduira l'utilisation du Recognition Planning Model pour la planification et du RPD pour la conduite de mission et les limites du RPD seront énoncées.

1.1. Les modèles de prise de décision pour les missions navales militaires

Dans cette section, une première partie va décrire en détail la méthodologie utilisée par la Marine Nationale pour planifier et conduire une mission. Une seconde partie va justifier pourquoi l'aide proposée aux militaires va s'appuyer sur des méthodes NDM pour préconiser des décisions, alors que les officiers utilisent aujourd'hui des méthodes analytiques pour la planification et leur expérience, pour les situations à forts risques nécessitant d'agir dans l'urgence.

1.1.1. Planification et conduite d'une mission tactique par la Marine Nationale

Comme évoqué dans le chapitre introductif, la réalisation d'une mission se décompose en une phase de planification, de conduite et de restitution. La Marine Nationale utilise la MEDOT [2] pour la planification d'une mission au niveau tactique et le processus d'évaluation pour la conduite. Ces deux méthodes sont décrites de manière plus détaillée dans les sections suivantes.

1.1.1.1. Planification d'une mission au niveau tactique

La MEDOT se décompose en deux phases principales, la phase préalable et la phase d'élaboration du plan d'actions.

La première phase permet d'établir le contexte de la mission en fixant le but à atteindre ainsi que les tâches à réaliser avec leurs contraintes et impératifs. Une fois le contexte mis en place,

le cadre général de l'action est déterminé en établissant tout d'abord les règles d'engagement et les règles de comportement, puis l'analyse du théâtre de la mission est réalisée en tenant compte des critères suivants :

- Les éléments environnementaux : les conditions météorologiques, avec une influence plus ou moins importante selon la mission, où seulement le vent, la visibilité et les vagues sont pris en compte,
- Les rails commerciaux : le but est de les éviter en connaissant leur localisation et éventuellement les horaires de passage,
- La cartographie : tous les obstacles statiques tels que les côtes, mais aussi le profil bathymétrique pour naviguer en eaux suffisamment profondes sont énumérés,
- Les « No Go area » : prise en compte des contraintes spatiales, par exemple les zones de pêche ou offshore où le navire ne doit pas aller,
- Les contraintes temporelles : le schéma temporel, c'est-à-dire l'enchaînement des tâches de la mission, à définir sera différent selon que la mission se déroule de jour ou de nuit,
- Les forces en présence : cela permet de prendre en compte la trajectoire des forces navales ennemies, alliés et neutres, en planification et de déterminer les contraintes spatio-temporelles.

Cette première phase permet de tirer des conclusions partielles sur le cadre général de l'action avec le lieu et les moments clés de l'action, les contraintes et impératifs ainsi que la capacité de réussite de la mission. L'analyse de ces conclusions permet de savoir quels sont les avantages du bâtiment de surface par rapport aux ennemis et de déterminer un premier « centre de gravité » ami. Le terme « centre de gravité » fait référence aux faiblesses du bâtiment et il est déterminé pour savoir où l'ennemi va essayer de le toucher en premier et quels sont les équipements à préserver absolument pour la réussite de la mission. Un premier centre de gravité est également déterminé pour les ennemis.

La seconde phase correspond à l'élaboration du plan d'actions en étudiant les modes d'actions amis et ennemis. Les modes d'actions amis consistent à déterminer les actions que le bâtiment en mission doit entreprendre pour toucher le centre de gravité ennemi et à déterminer les incapacités, les points forts et faibles et le centre de gravité ami. Les modes d'actions ennemis permettent d'énumérer toutes les actions qui vont s'opposer à la mission ainsi que leur probabilité et sévérité pour le centre de gravité du bâtiment. Une fois les modes d'actions établis, les modes amis et ennemis sont confrontés à l'aide de critères établis, où pour chaque mode d'actions amies, les avantages, inconvénients et risques sont énumérés. Suite à cette confrontation, les meilleurs plans d'actions sont sélectionnés et présentés à l'autorité qui aménagera éventuellement un de ses plans qui deviendra le plan définitif avec la rédaction des ordres.

La MEDOT se fonde sur une méthodologie analytique de la prise de décision qui pousse à rechercher l'optimum décisionnel en choisissant le mode d'actions le plus approprié face au mode d'actions de l'ennemi le plus plausible**.** Cette méthodologie a l'avantage de conforter le décideur avec une méthode formalisée et éprouvée.

1.1.1.2. Conduite d’une mission au niveau tactique

Une fois la phase de planification terminée, l’organe de commandement passe à la phase de conduite. Pour déterminer la faisabilité, il applique lors de la conduite le processus d’évaluation [3] à l’aide des trois étapes suivantes, illustrées Figure 1-1:

1. **Surveiller** la situation en temps réel pour collecter des informations importantes.

La surveillance dans le cadre du processus d’évaluation permet à l’organe de commandement de recueillir des renseignements pertinents sur la situation tactique en cours pour être comparés à la situation prévue décrite par le Commandant lors de la phase de planification. La progression des opérations ne peut être jugée, ni les décisions d’exécution ou d’ajustement prises, sans une compréhension la plus précise possible de la situation en cours.

Les informations nécessaires à la surveillance concernent l’ennemi, le terrain, la météo, les anomalies maritimes ainsi que les considérations d’ordre civil classées par ordre de priorité. Les anomalies maritimes sont décrites de manière plus exhaustive en 1.1.1.2.1 ci-dessous. L’EM consigne les informations pertinentes et évalue continuellement les opérations en cours afin de déterminer si elles se déroulent conformément à l’intention, à la mission et au concept d’opérations du Commandant.

2. **Evaluer** les progrès accomplis pour atteindre les conditions de l’état final, les objectifs et réaliser les tâches.

L’organe de commandement analyse les informations pertinentes recueillies pour évaluer le progrès des opérations. L’évaluation est le cœur du « processus d’évaluation » où la majeure partie de l’analyse se produit et les critères sous la forme de MOE aident à déterminer les progrès réalisés en vue d’atteindre les conditions de l’état final, de réaliser les objectifs et les tâches. Les MOE [3] décrites de manière plus détaillée en 1.1.1.2.2 sont utilisées au niveau stratégique, opérationnel et tactique pour évaluer l’impact des opérations militaires et mesurer les changements dans les comportements du système. La pertinence des actions exécutées est mesurée à l’aide des indicateurs préalablement choisis et construits par le Commandant.

Cependant, prédire la progression des opérations dans des environnements incertains, contraints par le temps avec des informations incomplètes, même à l’aide de MOE, est souvent très complexe. Les conditions changent très rapidement, les menaces s'adaptent et de nouvelles peuvent survenir de manière inattendue. De plus, l’interprétation en ce qui concerne les changements de comportement, d'attitude et de perception de l'Homme, s'avère souvent difficile. Par conséquent, l’EM doit s’interroger très souvent sur la faisabilité du plan d’actions envisagé et le Commandant et l’EM doivent faire attention à l’introduction des biais cognitifs pour ne pas tirer de conclusions erronées.

3. **Diriger** des actions d’amélioration.

La surveillance et l’évaluation sont des activités essentielles ; toutefois, le processus d’évaluation est incomplet sans des recommandations ou des directives sur les mesures à prendre. L’évaluation peut permettre de diagnostiquer des problèmes, mais à moins qu’elle ne donne lieu à des ajustements recommandés, son utilité pour le Commandant est donc limitée. En se fondant sur l’évaluation des progrès, l’EM réfléchit à des améliorations possibles du plan

et porte des jugements préliminaires sur le mérite relatif de ces changements. Les membres de l'EM déterminent les changements qui ont suffisamment de mérite et les présentent au Commandant sous forme de recommandations ou apportent des modifications dans les limites des pouvoirs qui leur sont délégués.

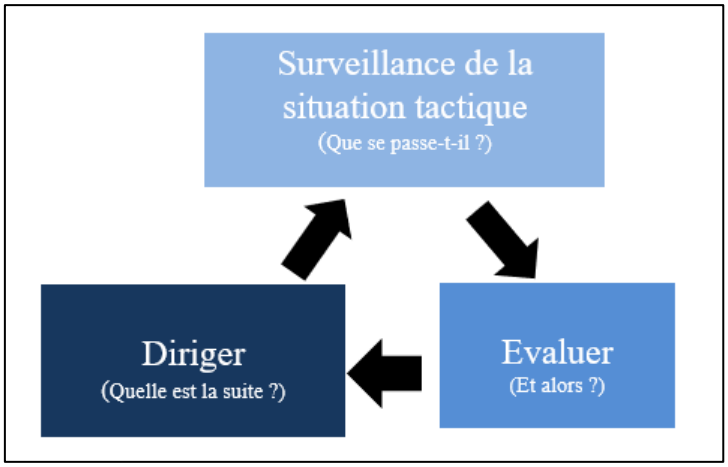

Figure 1-1 : Le processus d'évaluation [3]

1.1.1.2.1. Les événements indésirables ou les anomalies maritimes

Lors de l'étape de surveillance en conduite de mission, l'organe de commandement doit rester très vigilant quant à la survenue d'événements indésirables ou d'anomalies maritimes [20], [21], [22], [23]. Ces dernières peuvent être définies comme la détection de comportements inhabituels dans le domaine maritime et l'évaluation de leur potentiel de menace. Il en existe un certain nombre, telles que l'attaque de pirates, les menaces asymétriques, la pêche illégale ou encore le risque de collision. Roy et Davenport [24] ont alors créé des familles d'anomalies maritimes et des règles de détection. La Figure 1-2 extraite de [25] recense ces différentes familles :

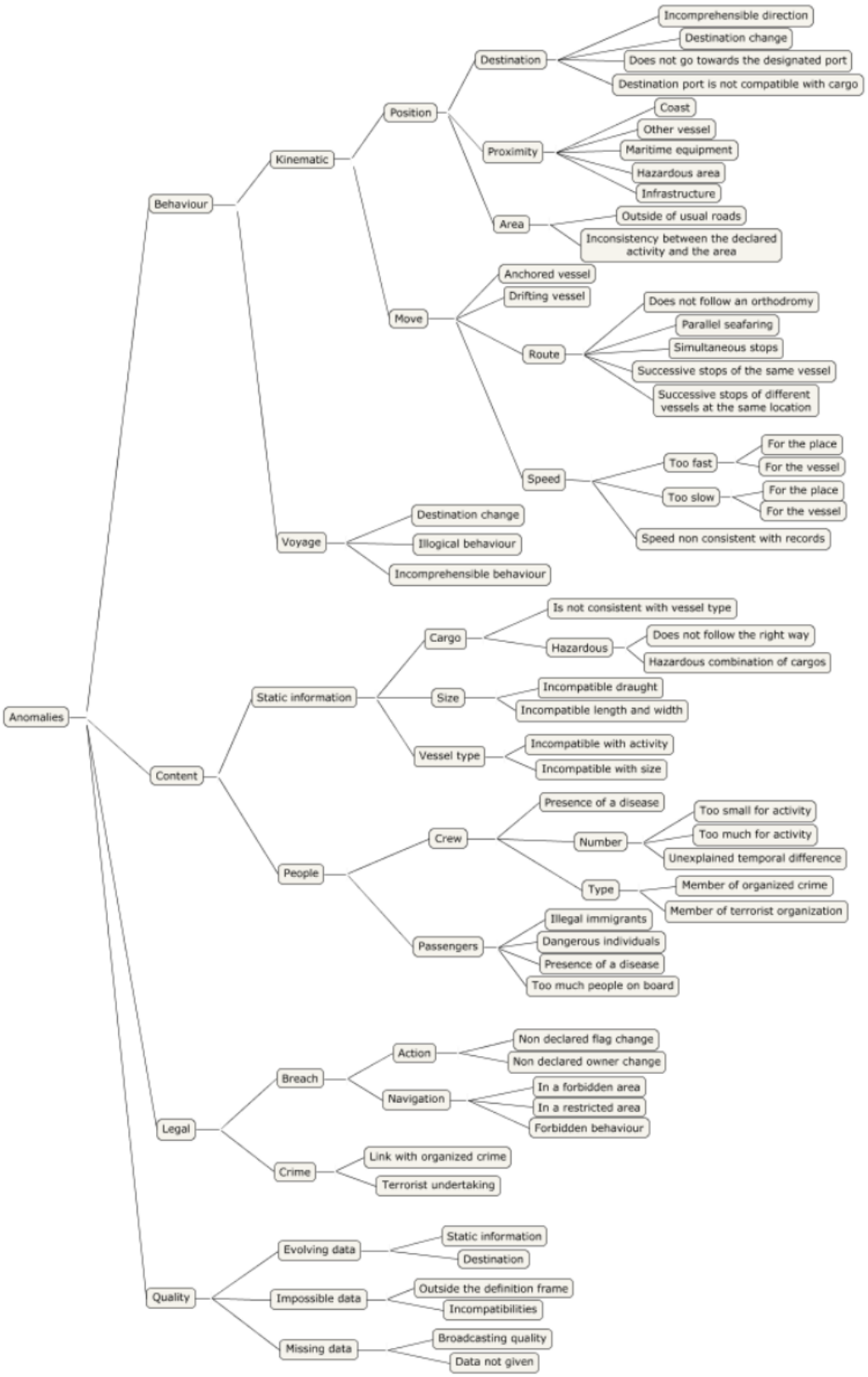

Figure 1-2 : Schéma des familles d'anomalies maritimes [25]

Dans le cadre de cette thèse, la famille la plus importante est celle regroupant les anomalies comportementales divisée elle-même en deux sous-familles. Les anomalies cinématiques constituent la principale sous-famille avec les anomalies basées sur la position (ex : un changement de direction soudain) et celles sur le mouvement (ex : une vitesse trop élevée). La seconde sous-famille désigne les anomalies basées sur l'itinéraire (ex : un comportement illogique).

Concernant les anomalies de contenu, elles se composent aussi de deux sous-familles. La première sous-famille est en rapport avec les anomalies relatives aux données statiques. Les données statiques sont des données propres à chaque navire, le caractérisant et qui ne changent généralement pas au cours du temps. Une anomalie dans ces données peut par exemple concernée les dimensions du navire (ex : la largeur déclarée est supérieure à la longueur déclarée) ou le type de navire (ex : incompatibilité avec les dimensions du navire). La seconde sous-famille concerne les anomalies relatives aux personnes à bord (ex : l'équipage appartient à une organisation criminelle ou terroriste, ou le nombre de personnes composant l'équipage est trop élevé par rapport à l'activité déclarée).

Il existe aussi les anomalies juridiques avec une sous-famille relative aux questions criminelles par exemple le terrorisme ou la criminalité organisée et une sous-famille relative aux infractions, telles que le changement de pavillon non déclaré ou un comportement maritime non autorisé.

La dernière famille d'anomalies concerne la qualité des données avec une sous-famille concernant les données qui changent de manière inattendue, une autre en rapport avec les données aberrantes et une dernière sur les données manquantes à cause d'une mauvaise réception du signal ou à une volonté de ne pas fournir les données.

L'ensemble de ces anomalies va permettre à l'organe de commandement lors de la phase de surveillance et d'évaluation de la situation de déceler les comportements dangereux des autres navires pour le navire en mission.

1.1.1.2.2. Les mesures d'efficacité - Measures Of Effectiveness (MOE)

Lors de l'étape d'évaluation appliquée en conduite de mission, le Commandant et l'EM doivent évaluer en permanence la situation actuelle pour la comparer au plan d'actions initial afin de s'assurer que l'opération est en accord avec l'état final et les objectifs souhaités. Les MOE sont des grandeurs scalaires utilisées pour évaluer la situation actuelle et permettre de déterminer la progression des opérations vers la réalisation des objectifs et l'état final. Les MOE aident l'organe de commandement à répondre à des questions telles que : « faisons-nous les bonnes choses ? Nos actions produisent-elles les effets souhaités ? Ou des actions alternatives sont-elles nécessaires ? » [3].

Des MOE, bien construites, indiquent avec précision le succès opérationnel et aident l'organe de commandement à prendre des décisions précises et opportunes. Au contraire, des MOE mal construites peuvent amener un Commandant ou un décideur à prendre des décisions inappropriées pouvant entraîner une multitude d'effets négatifs qui ne rapprocheront pas le navire de l'atteinte de ses objectifs. Les MOE utilisent des indicateurs d'évaluation qui doivent être pertinents, mesurables, adaptés et dotés de ressources afin qu'il n'y ait pas de fausse

impression de réalisation des tâches ou des objectifs. Une MOE peut être composée d'un ou plusieurs indicateurs telle qu'illustré Figure 1-3 et doivent être faciles à comprendre et à évaluer. Une MOE compliquée est plus difficile à évaluer et peut conduire à la confusion et à un manque de compréhension du véritable problème.

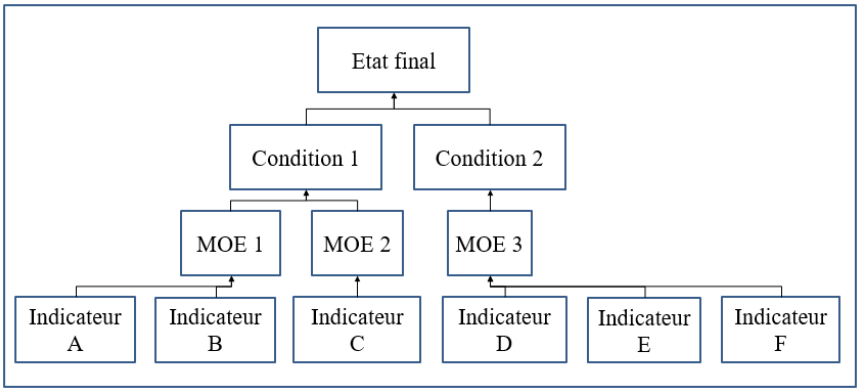

Figure 1-3: Exemple de la composition de MOE

L'objectif est donc de créer des indicateurs de MOE clairs et appropriés pour l'évaluation qui peuvent être décomposés en cinq éléments :

1. Titre abrégé : le nom de l'indicateur ;
2. Définition : une description claire de ce que l'indicateur mesure ;
3. Unité de mesure : peut-être quantifiable ou qualitative ;
4. Point de référence : une valeur qui définirait l'état souhaité par rapport à l'environnement opérationnel ;
5. Formule : une expression de la façon dont les changements de la valeur de l'indicateur affectent la MOE.

La mise en place de MOE pertinentes est donc une étape cruciale pour mener à bien l'étape d'évaluation et prendre des décisions adaptées à la situation.

1.1.2. Le processus de décision en situation naturelle pour la conduite de mission

Alors que le système français (MEDOT) s'appuie sur une prise de décision rationnelle, le système américain est fondé sur les méthodes NDM théoriques les plus récentes telles que le RPD afin de réduire le temps de prise de décision. Plusieurs modèles ont été proposés, le premier a été initié dans les années 1960 par John Boyd [26] avec le modèle OODA (Observe – Orient – Decide – Act) qui insiste sur l'importance de l'observation et de l'interprétation de la situation. Plus tard, ce sont les approches NDM qui sont devenues la référence dans la prise de décision en situation de crise.

Même si le système français (MEDOT) est toujours adapté au contexte dans lequel il est appliqué, en une vingtaine d'années le contexte des missions tactiques a évolué. D'un affrontement autrefois symétrique et planifiable mené par des Hommes avec une expérience

opérationnelle réelle limitée, l'affrontement est aujourd'hui asymétrique avec des Commandants ayant tous participé à plusieurs opérations extérieures [27]. La situation tactique à gérer est aujourd'hui devenue très complexe.

Ainsi, la prise de décision lors d'une mission navale tactique traitée dans le cadre de ces travaux de thèse est principalement caractérisée par (1) un temps restreint laissé au décideur pour prendre une décision et (2) un risque fort nécessitant une réaction rapide, c'est-à-dire une situation pour laquelle une perte conséquente pour le navire est inéluctable si une action n'est pas réalisée dans le temps imparti. C'est par exemple le cas d'une décision de changement de cap dû à un obstacle inattendu (conteneur en surface, changement de cap soudain d'un navire à proximité) ou d'une réponse adaptée face à une menace (navire de la flotte ennemie devenant menaçant, tentative d'abordage d'une embarcation légère). L'organe de commandement doit donc décider rapidement d'une action à entreprendre, c'est à dire trouver à l'aide de son expertise, le meilleur compromis entre (1) le bénéfice d'une action envisagée et (2) le temps nécessaire pour planifier et accomplir cette action.

Les nombreuses informations issues du navire et de son environnement ne sont pas utilisables lors de ces prises de décision. Même si ces informations caractérisent avec une certaine exhaustivité une situation, leur analyse, chronophage, ne permet pas à l'organe de commandement de trouver la meilleure solution dans un temps contraint. La prise de décision se fonde alors sur un nombre restreint d'informations, d'indicateurs et de processus de raisonnement afin d'aboutir rapidement à une solution efficace même si elle n'est pas optimale.

C'est pourquoi nous préconisons de fonder une aide à la décision sur les processus mobilisés par les décideurs lors d'une prise de décision dans l'urgence afin de (1) raisonner à l'aide d'informations présentant des degrés d'incertitude et proches de celles sélectionnées par les décideurs, et (2) préconiser des réponses que le décideur pourra rapidement comprendre et évaluer. La préconisation d'une action parmi l'ensemble des actions envisageables par le décideur augmentera la confiance du décideur dans le système qui n'aura pas besoin de mobiliser de ressources cognitives supplémentaires pour comprendre les décisions préconisées. Une aide à la décision fondée sur des méthodes analytiques préconisera certes une décision optimale à l'utilisateur dans un temps restreint. Cependant, en plus de se baser sur des informations complètes sans degré d'incertitude, condition invérifiable en réalité opérationnelle, il est fort probable qu'en temps restreint, le décideur n'ait pas pensé, ni envisagé la solution préconisée, ce qui lui demandera une réflexion supplémentaire pour analyser et comprendre cette décision introduisant un manque de confiance dans le système. Le décideur soumis à la pression préfèrera suivre son intuition que douter et perdre un temps « précieux » à comprendre ce qui lui a été proposé. Au contraire, une aide fondée sur des méthodes intuitives issues du NDM, telles que le RPD se base sur des informations utilisées par les décideurs et préconisera une décision satisfaisante et intuitive pour le décideur qui n'aura pas besoin de solliciter des ressources cognitives supplémentaires. L'aide à la décision diminuera ainsi le temps de réflexion de l'organe de commandement sans modifier les informations habituellement mobilisées et les mécanismes de raisonnement mis en œuvre pour trouver une solution satisfaisante rapidement.

Les méthodes analytiques comme la MEDOT, la méthode tactique interarmées, sont donc peu ou mal adaptées à des prises de décisions rapides à forts enjeux**. Dans ces situations, l'organe de commandement sera mieux assisté par une aide à la décision s'appuyant sur des méthodes intuitives issues du NDM.** Dans la suite, les modèles analytiques et intuitifs

vont être détaillés et comparés pour conclure sur les travaux de Klein concernant le RPD [5] sur lesquels va reposer notre aide à la décision.

1.2. Prise de décision humaine

Dans cette section les modèles analytiques et intuitifs sont d'abord explicités permettant de constater une opposition dans la manière de prendre les décisions. Ils sont ensuite comparés dans une dernière partie, justifiant le choix de fonder notre aide à la décision sur le RPD.

1.2.1. Les modèles analytiques et intuitifs

1.2.1.1. Les modèles analytiques

Un des premiers modèles de prise de décision humaine et analytique est le modèle de Savage [28] initié en 1954. Celui-ci fournit un cadre formel et cohérent pour réfléchir sur la prise de décision et insiste sur la différence cruciale entre les éléments que le décideur ne peut pas contrôler, l'ensemble E des événements et les éléments qu'il peut contrôler, l'ensemble A des actions. Il affirme que si le décideur suit un cheminement cohérent pour faire un choix, alors il existe un ensemble de probabilités et une fonction d'utilité U telle que le décideur peut chercher à maximiser l'utilité attendue pour les dites probabilités. Le théorème de Savage est souvent utilisé à l'inverse, c'est-à-dire pour suggérer quelle action maximise l'utilité attendue du décideur compte tenu d'un ensemble de probabilité connu pour des événements possibles. La théorie de l'utilité a servi de point de départ au développement des modèles analytiques ou prescriptifs basés sur le principe du choix rationnel.

Les modèles analytiques appliquent une procédure très détaillée afin d'analyser une décision et d'évaluer ses composantes. La plupart de ces modèles suivent la résolution de problèmes en cinq étapes, qui comprend l'analyse de la situation, la génération de solutions possibles et la comparaison des solutions par rapport à un ensemble de critères objectifs. Ils sont particulièrement utiles pour présenter des lignes directives permettant de prendre de meilleures décisions ou des décisions optimales. Une liste non exhaustive des avantages est :

- La nature générale des modèles permet leur application dans de multiples domaines.
- Les modèles assurent une compréhension commune partagée parmi les participants.
- Ils permettent d'incorporer des outils d'aide à la décision qui assistent et guident la prise de décision.
- Ils augmentent les chances de surmonter les biais cognitifs personnels [29] à l'origine des altérations de jugement.

Ces points forts permettent d'identifier et de sélectionner le plan d'action optimal, mais en raison du niveau de travail détaillé, les modèles analytiques requièrent beaucoup de temps.

Les modèles les plus utilisés sont les modèles classiques. Ils sont appliqués en cas de certitude, lorsque le décideur dispose de toutes les informations relatives au problème et connaît également toutes les solutions alternatives.

Le modèle classique repose sur quatre hypothèses principales :

1. Le problème clairement défini. Le modèle suppose que le décideur a des objectifs clairement définis et sait ce que l'on attend de lui.
2. Le modèle suggère que le décideur doit éliminer toute incertitude susceptible d'avoir un impact sur la décision. Par conséquent, il n'y a pas de risques à prendre en compte.
3. L'information est complète. Le décideur est capable d'identifier toutes les alternatives à sa disposition, de les évaluer et de les classer objectivement.
4. Les décisions sont rationnelles. Le décideur est supposé agir toujours dans le meilleur intérêt de l'organisation.

Le modèle classique propose trois étapes principales pour la prise de décision :

- Dresser la liste de toutes les alternatives disponibles : dans le modèle classique, le décideur n'est pas limité par le temps ou les ressources et peut continuer à chercher des alternatives jusqu'à ce qu'il identifie celle qui maximise l'utilité de la décision.
- Classer les alternatives listées : le décideur est censé posséder non seulement toutes les informations nécessaires, mais aussi la capacité cognitive de hiérarchiser les alternatives de manière précise et objective.
- Sélectionner l'alternative la mieux adaptée.

Cependant, les travaux de Savage et les modèles analytiques sont critiqués sur la manière dont la probabilité d'occurrence des événements est estimée [30], [31]. Des probabilités spécifiques peuvent être attribuées à chaque événement, mais il est assez difficile de faire la distinction entre les événements en fonction de leur probabilité d'occurrence. De plus, bien que les probabilités décrites dans le travail de Savage aient du sens théorique, elles ne signifient pour autant rien pour les décideurs en situation réelle. L'idée que le décideur puisse exprimer les probabilités relatives à tous les événements futurs et qu'il puisse ensuite maximiser leur utilité attendue n'est pas réaliste. Les probabilités attribuées par un décideur ne peuvent être qu'à priori et subjectives car elles ne reposent sur aucune connaissance spécifique des événements futurs.

Plus tard, l'observation des décideurs en situation réelle a permis d'illustrer une nouvelle voie de la prise de décision humaine : oublier la rationalité pure et prendre des décisions « satisfaisantes ». Ce cas correspond à la notion de « rationalité limitée » qui sera introduite par Herbert Simon [32] et qui donnera lieu au développement des modèles intuitifs (NDM) ou descriptifs.

1.2.1.2. La prise de décision naturelle (les modèles intuitifs)

1.2.1.2.1. Le principe de la « rationalité limitée »

Herbert Simon dans les années 1950-1960 [32], remarque que le comportement réel des décideurs est loin d'être représentable de manière purement rationnelle, car (1) les décideurs n'ont jamais une idée parfaitement claire de leur problème ; (2) les problèmes de décision se présentent souvent comme la recherche d'un compromis et (3) la solution d'un problème est soumise à des contraintes temporelles et des ressources disponibles.

Ainsi, Simon a initialement présenté la prise de décision comme comprenant trois étapes :

1. L'identification de toutes les actions ou alternatives possibles.
2. La détermination des conséquences de toutes les actions possibles.
3. L'évaluation des conséquences de chaque action possible.

Par la suite, Simon ajoute plusieurs autres aspects aux différentes phases de son processus décisionnel, notamment en ce qui concerne la représentation du problème, la manière de poser le problème et la recherche d'informations. Ceci l'amène à son travail fondateur sur les quatre phases de prise de décision suivantes [33] :

1. Renseignement.
2. Conception.
3. Choix.
4. Revue.

Le rôle de l'information est fondamental dans la phase de renseignement et de conception car le décideur ne choisit que parmi les actions identifiées et qu'il a pu documenter. Ainsi, comme l'a indiqué Simon : l'information contraint la décision.

Le cadre défini par Simon permet de relier la décision et l'information mais celui-ci n'est pas assez riche en termes de compréhension des choix et d'analyse du rôle des événements futurs. Il est précisément au cœur du débat sur les limites cognitives des décideurs humains et leur incapacité à prévoir des événements lointains. Les limites du cerveau et la nature des décisions ne permettent pas de gérer tous les scénarios possibles [34].

Le problème de l'évaluation des conséquences d'une action qui se retrouve dans les phases de choix et de revue est central dans tout processus décisionnel. Dans le travail de Savage, l'évaluation des conséquences suppose la connaissance de tous les événements futurs avec leurs probabilités. En théorie, il peut suffire de maximiser une fonction d'utilité pour un ensemble de choix, mais la difficulté est de déterminer quel est, en pratique, le rôle de la raison lorsqu'il n'y a ni choix clairs, ni fonction d'utilité complète et que les gestionnaires opèrent avec une connaissance minimale des événements futurs. Cela signifie que puisque les décideurs ne peuvent pas avoir une connaissance complète du monde, ils doivent viser à prendre des décisions sous-optimales ou satisfaisantes. En pratique, le processus de décision s'arrête dès que le décideur trouve une solution qui donne satisfaction en tenant compte du scénario le plus plausible, et qui ne risque pas non plus de se révéler catastrophique. La « rationalité limitée » a

été la première tentative pour fournir un cadre scientifique pour l'étude rigoureuse et significative de décisions en situation réelle.

1.2.1.2.2. Prise de décision amorcée par la reconnaissance de la situation : le Recognition Primed Decision making (RPD)

La prise de décision humaine ne peut pas être décrite sans tenir compte du rôle des événements futurs, même si les humains peuvent réagir directement à des stimuli de leur environnement sans anticiper les événements futurs. Cependant, dans le domaine du raisonnement, c'est-à-dire lorsque le décideur dispose de suffisamment de temps pour générer une projection d'événements futurs dans son esprit, deux phases clés sont à distinguer : le diagnostic et la projection.

La phase de diagnostic consiste à reconnaître l'état actuel de l'environnement, c'est-à-dire le passé et le présent. Cependant dans la réalité, la situation est souvent plus compliquée à reconnaître notamment lorsque le diagnostic ne permet pas d'identifier une situation déjà vécue et mémorisée. Dans la phase de projection, le décideur doit anticiper les conséquences des décisions potentielles, en fonction de sa perception de l'avenir.

En considérant les spécificités de la prise de décision humaine et l'importance de la phase de projection, il est utile d'examiner en détail les travaux de Gary Klein [5] et de ses associés sur la prise de décision amorcée par la reconnaissance (RPD) et la prise de décision en situation naturelle. La NDM est à la fois le fruit d'un ensemble de recherches sur la prise de décision humaine et une orientation méthodologique qui s'est concentrée sur l'étude de certaines fonctions cognitives qui émergent dans des contextes naturels, souvent dans des situations de prise de décision qui impliquent de fortes contraintes de temps et/ou des décisions très importantes. Initialement, Klein et ceux qui ont adopté ses idées ont étudié des cas concrets d'acteurs de la société civile confrontés dans leurs métiers à la prise de décision en environnement contraint tels que les pompiers, les personnels soignants des services d'urgence et le personnel paramédical afin d'obtenir des observations directes sur la façon dont cette prise de décision extrême peut se produire. Leurs observations révèlent que, contrairement aux prédictions des modèles analytiques, ces individus ne conçoivent pas de solutions alternatives qu'ils comparent entre elles, mais utilisent leur diagnostic de la situation pour construire très rapidement une solution de cas optimal, dont ils simulent ensuite l'exécution dans leur esprit pour voir si elle correspond bien à ce qu'ils imaginent : c'est ce qui est appelé le RPD. Dans la phase suivante, lorsque la solution est sélectionnée, les décideurs utilisent la simulation mentale pour valider si la situation répond à leurs attentes et prendre d'autres décisions lorsqu'ils voient se développer un événement inattendu Ainsi, le RPD suggère un processus de décision en deux étapes :

1. La reconnaissance de schémas déjà vus.
2. La simulation mentale.

De manière résumé, la NDM est « la façon dont les personnes utilisent leur expérience pour prendre leur décision » [35]. Plutôt que de se concentrer sur le développement de multiples solutions, les modèles NDM mettent d'avantage l'accent sur la capacité d'un individu à évaluer une situation afin d'élaborer une solution unique satisfaisante. L'une des caractéristiques du

modèle NDM est son application à des situations où les actions nécessitent d'être réévaluées en permanence. En raison de ce besoin continu de réévaluation, les méthodologies NDM mettent l'accent sur l'opportunité de suivre les rythmes changeants et potentiellement mal définis mais aussi sur la capacité des décideurs et des experts à prendre des décisions fondées sur leur expérience personnelle. Les méthodologies NDM ont besoin de décideurs experts dans leur domaine. L'action de reconnaître des situations caractéristiques dans un environnement complexe et de développer rapidement un plan d'actions nécessite un niveau d'expérience et d'expertise que les décideurs novices n'ont pas toujours [35]. En général les décideurs expérimentés n'évaluent qu'une seule option mais en examinent différents aspects mentalement. Ils prennent une décision lorsque l'option devient réalisable. Ainsi, ces méthodologies ne sont pas toujours appropriées dans des situations nouvelles, car il est peu probable que le décideur dispose d'une expérience fiable lui permettant d'identifier des pistes d'action potentielles [36].

1.2.2. Comparaison des modèles analytiques et intuitifs

Alors que les modèles analytiques mettent l'accent sur l'optimisation, les modèles intuitifs se concentrent sur le choix d'une solution viable le plus rapidement possible. Si la première solution fondée sur l'expérience semble inappropriée, le décideur continue d'évaluer mentalement les options jusqu'à ce qu'il trouve une solution qui fonctionne. C'est pourquoi le principal avantage des méthodes NDM est la rapidité. Le Tableau 1-1 [37] montre quelle stratégie est à appliquer en fonction de la situation :

Tableau 1-1 : Les différences entre les modèles analytiques et intuitifs [37]

Critères comparatifs	Modèle analytique	Modèle intuitif
Application	Paramètres du problème bien définis	Paramètres du problème mal définis
Variabilité	Applications générales	Applications dynamiques
Source du contrôle	Tous les processus du modèle	Les facteurs de situation
Processus de réflexion	Analytique et comparatif	Créatif et discriminant
Orientation	Orienté processus	Orienté but
Fondations	Objectifs clairement définis	Incertitude
Connaissances requises	Compréhension complète du problème	Compréhension incomplète
Informations requises	Complètes	Incomplètes
Buts	Prédéterminés	Basés sur la situation
Résultats désirés	La solution optimale	Une solution satisfaisante
Bases théoriques	Modèles et processus classiques	Pensée naturaliste

De plus, les modèles analytiques développent de multiples options concurrentes qui sont ensuite analysées pour déterminer la solution optimale. Les modèles intuitifs développent seulement une option, en plaçant en priorité une arrivée avec une solution satisfaisante aussi

vite que possible. Il y a des avantages à trouver une ligne de conduite optimale, tout comme il y a aussi des avantages à en trouver une rapidement satisfaisante. Ces compromis sont synthétisés dans le Tableau 1-2 [37] :

Tableau 1-2 : Avantages de chacune des deux méthodes [37]

	Modèle analytique	**Modèle intuitif**
Points positifs	Mieux adapté lorsqu'un problème implique une telle complexité de calcul que les processus de reconnaissance sont inadéquats	Optimal pour un personnel expérimenté sous temps restreint sur des tâches concrètes dépendantes du contexte
	Particulièrement utile lorsqu'il est nécessaire de justifier la décision auprès d'autres	Veille à ce que le décideur soit prêt à agir et à s'adapter à des situations changeantes
	Simplifie le processus de décomposition de tâches nouvelles et complexes basées sur des données abstraites que la reconnaissance ne peut pas traiter	Simplifie le processus de création d'un plan lorsque la situation est reconnaissable
	Si le temps le permet, il décrit un processus qui peut amener les décideurs à prendre la « décision optimale »	Réduit les exigences en matière de temps et de ressources
	Bénéfique pour les décideurs novices qui peuvent manquer d'expérience pour prendre une décision	
Points négatifs	Temps nécessaire pour prendre une décision très long	Elle ne peut pas garantir que la décision optimale soit prise
	Sollicitation de beaucoup de ressources humaines pour prendre des décisions	Inadapté pour les décideurs inexpérimentés qui ne connaissent pas la situation
	Le montant des ressources (humaines, temps) nécessaires pour prendre une décision est élevé	Difficile d'articuler la base de décision, ainsi que de concilier les conflits

Un décideur doit être capable de reconnaître quel est le processus le plus approprié pour pouvoir l'aider dans sa prise de décision. Lorsque le temps pour la planification ou la conduite est disponible ainsi que les informations, le processus de décision analytique est la meilleure solution. Au contraire, lorsque le temps et les informations à disposition sont limités, et que les décisions doivent être prises malgré un certain degré d'incertitude, le processus de décision en situation naturelle est plus approprié.

Cet état de l'art et pour terminer cette comparaison entre les modèles analytiques et intuitifs permet d'appuyer notre choix de fonder notre système d'aide à la décision pour la conduite de missions sur des méthodes issues du NDM telles que le RPD (Recognition Primed Decision-Making) [27]. **Nos travaux de thèse vont donc s'appuyer sur le RPD utilisé pour la conduite de missions**. Les sections suivantes détaillent ces modèles de prise de décisions.

1.3. Le Recognition Primed Decision-making (RPD) pour les décisions de petits groupes

Dans cette section, le Recognition Planning Model (RPM) pour la planification d'une mission est succinctement décrit. Le RPM, qui ne sera pas utilisé dans le cadre de cette thèse, est complémentaire au RPD et pourrait être utilisé pour la phase de planification pour obtenir rapidement un plan d'actions de référence. Pour finir, le RPD permettant de définir un nouveau plan d'actions lors de la conduite de missions et sur lequel va s'appuyer nos travaux de thèse est détaillé.

1.3.1. Accomplissement d'une mission au niveau tactique à l'aide du RPM et du RPD

La Figure 1-4 montre comment le RPM et le RPD sont utilisés pour mener à bien une mission.

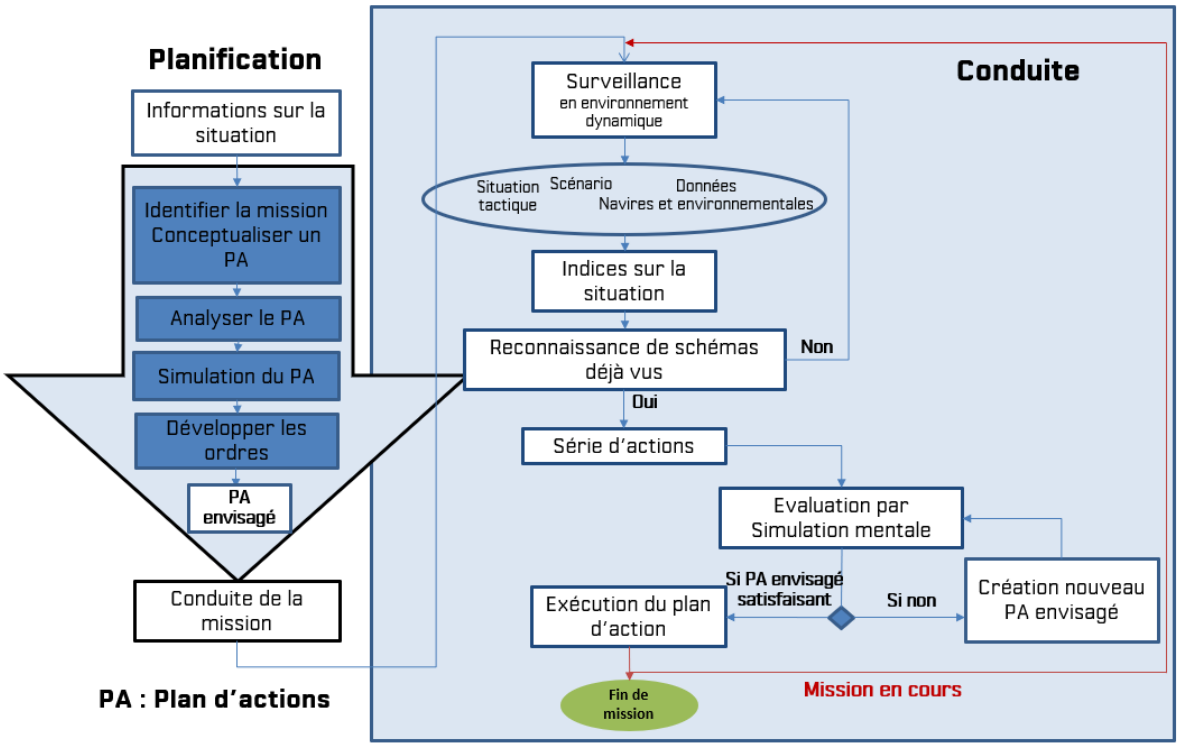

Figure 1-4 : Accomplissement d'une mission à l'aide du RPM et du RPD

Lors de la planification, le RPM est mis en œuvre. Lorsque celui-ci est terminé, un plan d'action est retenu et la conduite de la mission peut commencer. L'EM va surveiller l'évolution du théâtre des opérations et appliquer le raisonnement du RPD. Il va collecter des indices sur la situation jusqu'à reconnaître une situation déjà vue. Une fois que le décideur a confirmé ses attentes, il évalue si le plan d'action envisagé lors de la planification est encore viable, sinon il le modifie ou en propose un nouveau. Dans tous les cas, le plan d'action est évalué par

simulation mentale. Une fois satisfaisant, le plan d'action est mis en œuvre. Le raisonnement RPD est répété jusqu'à la fin de la mission.

1.3.2. Le RPM pour la planification

John F.Schmitt et Gary Klein ont développé le Recognition Planning Model (RPM) à partir des recherches sur le RPD et de plusieurs études d'exercices de planification militaire, afin de codifier les stratégies de planification informelles et intuitives utilisées par les équipes de planification de l'armée [38]. L'Armée britannique a mené des expériences sur le RPM, démontrant sa validité [39]. Ross et *al.* a montré que le processus permettait une augmentation du rythme de planification d'environ 20% sans perdre en efficacité [39]. Il a également observé que les plans issus du RPM étaient un peu plus audacieux et mieux adaptés aux exigences de la situation que les plans issus de méthodes analytiques, qui ont tendance à être plus contraints par une conformité excessive aux modèles doctrinaux actuels.

Le RPM [38] est une application du modèle RPD et consiste pour l'organe de commandement à identifier le plan d'action préféré afin qu'il puisse le détailler et l'améliorer, tel que l'illustre la Figure 1-5.

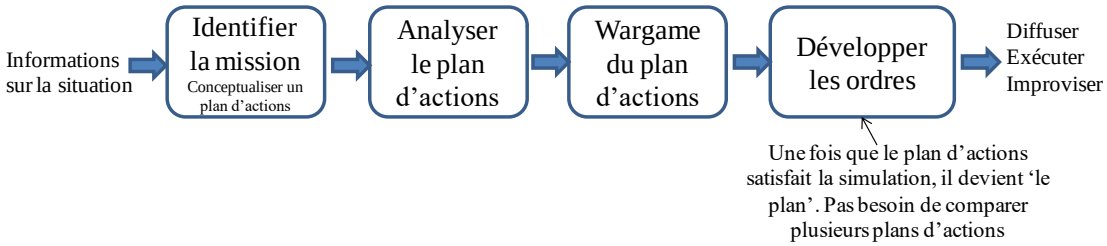

Figure 1-5 : Schéma du Recognition Planning Model (RPM) [38]

Lors de la première étape consistant à identifier la mission et conceptualiser un plan d'action, l'EM et le Commandant essaient de comprendre la mission qu'ils ont reçue tout en décidant de comment la réaliser. L'identification précoce d'un plan d'action de base peut orienter l'analyse de la mission. Si la situation est peu connue ou peu développée, une analyse approfondie de la mission peut précéder la conceptualisation du plan d'action. Si au contraire, le Commandant connaît bien la situation, l'analyse de la mission peut se faire rapidement. Le RPM ne fige pas la planification dans une stratégie unique ; il permet au Commandant et à l'EM de rechercher des options. Le résultat de cette première phase est une décision représentée par un plan d'opération provisoire élaboré très tôt dans le processus de planification. Cette décision précoce permet d'accélérer le rythme de la planification et facilite la planification parallèle. Une décision précoce peut permettre à d'autres échelons et organisations de commencer leur propre planification plus tôt. De plus, dans les modèles analytiques existants, les plans d'actions candidats restent statiques jusqu'à ce qu'un soit choisi. Dans le RPM, la sélection précoce d'un plan permet son amélioration continue à mesure que la situation est d'avantage connue.

L'étape suivante consiste à analyser et rendre opérationnel le plan d'actions. Une fois qu'un plan a été décidé, il doit être analysé et opérationnalisé. Ces tâches ont déjà commencé lors de la première étape. Pendant que les membres de l'EM analysent le plan, ils peuvent déjà

préparer des ordres d'opération ou trouver des failles qui disqualifient le plan et en choisir un autre. L'EM se livre ensuite à un « jeu de guerre » (Wargame) pour voir si le plan sera adapté face aux plans ennemis. Si le temps presse, le jeu de guerre peut également servir de répétition, permettant à l'EM de commencer à construire les rôles de chacun. Cette phase du RPM correspond à ce qui est plus communément appelé la planification de l'exécution.

Une fois que le plan d'actions a été validé, il ne reste plus qu'à produire les documents d'exécution nécessaires lors de l'étape « d'élaboration des ordres ». Cette étape est rapide car dès la première phrase, « analyser et rendre opérationnel le plan d'actions », un seul plan est pris en compte. Enfin, il est important de noter que le RPM comporte diverses boucles de rétroaction à chaque étape [38].

1.3.3. Le RPD pour la conduite de mission

Le RPD est l'un des modèles les plus connus et les plus étudiés du NDM, se concentrant sur la génération d'un seul plan d'actions possible plutôt que sur la comparaison de plusieurs plans d'actions concurrents. Il se concentre sur le choix d'une décision qui satisfait les besoins les plus importants, qui ne sont pas nécessairement les décisions optimales. Dans un contexte de commandement de mission, la réussite de l'application du RPD dépend de la capacité du Commandant à faire confiance à l'intuition et à l'expérience de ses subordonnés pour élaborer un plan d'actions. Cela signifie que les décideurs qui ont recours au RPD choisiront la première option qui répond à leurs exigences (intention, objectifs, sécurité, etc.) plutôt que de prendre le temps de trouver le plan optimal.

Le RPD est un modèle individuel de prise de décision basé sur une simple affirmation : les décideurs expérimentés sont capables de générer une option satisfaisante [25]. Il est supposé que les décideurs sont capables de comparer les situations actuelles avec des situations antérieures qu'ils ont vécues, de reconnaître les similitudes et d'utiliser ces similitudes pour combler les lacunes en matière de connaissances afin d'évaluer la situation et de déterminer rapidement une solution potentielle. Par rapport aux processus analytiques, le RPD est relativement rationalisé, son processus simplifié faisant appel à moins d'acteurs. Le processus impliqué dans le RPD, à base d'arbres de décision est publié à l'origine par Gary Klein et David Klinger.

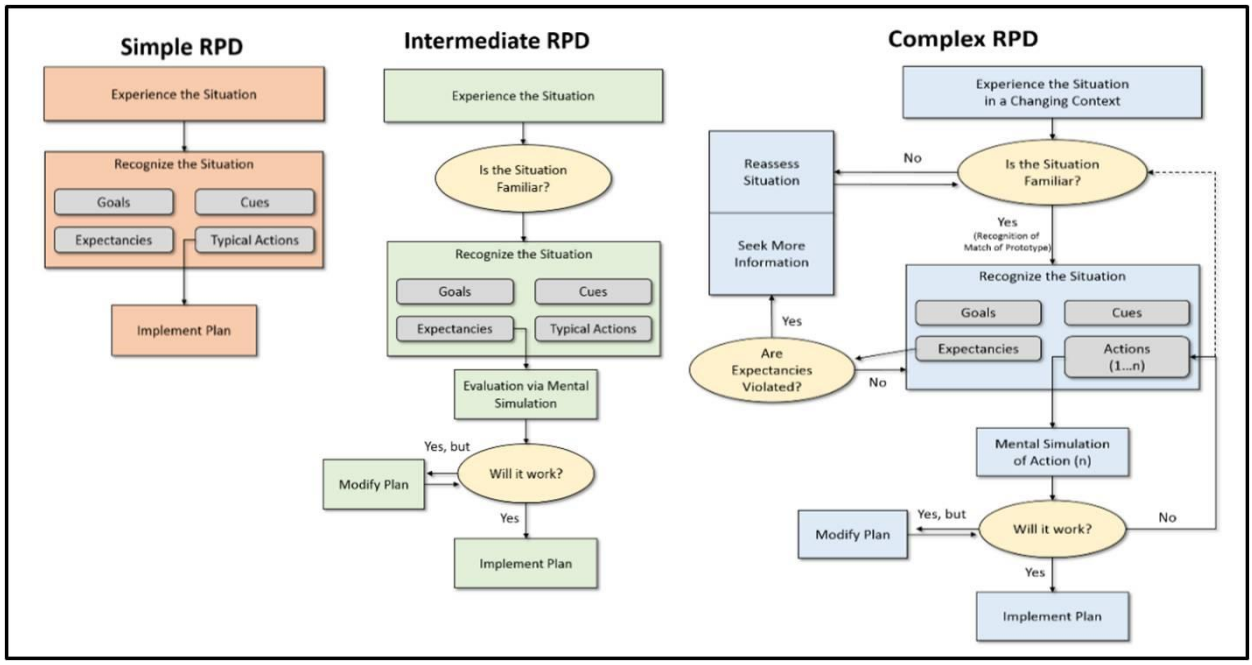

Figure 1-6 : Arbres de décision décrit par Gary Klein et David Klinger [25]

Sur la Figure 1-6, le RPD débute au moment où le décideur sonde la situation et s'appuie sur son expérience pour la reconnaître. Si le décideur ne reconnaît pas la situation, il peut demander des informations complémentaires jusqu'à ce que la reconnaissance de la situation ait lieu. Une fois la reconnaissance effectuée, il en résulte quatre sous-produits : **indices, objectifs, actions, attentes**. Ces quatre éléments sont issus de l'expérience et décrivent les concepts cognitifs sur lesquels un décideur fonctionne [40].

Les indices (**Cues**) représentent les éléments physiques et mentaux sur lesquels un décideur est en mesure de comprendre et de suivre une situation. Savoir quel petit ensemble d'indices il faut surveiller parmi toutes les informations disponibles pour décider est la marque d'un expert. Les indices sont souvent constitués d'éléments d'information agrégés qu'un décideur rassemble dans son esprit.

Les objectifs (**Goals**) sont des éléments clés du processus de reconnaissance. Ils représentent un état final que le décideur tente d'atteindre. Le décideur essaie de satisfaire plusieurs objectifs et doit choisir les actions qui peuvent le mieux les atteindre.

Les actions (**Actions**) représentent l'ensemble des décisions potentielles parmi lesquelles un décideur peut faire son choix. Un décideur expérimenté dans le domaine de la décision, sait intuitivement quelle action est susceptible d'être la plus favorable. La simulation mentale[1]est utilisée pour évaluer cette action afin de voir si elle est appropriée au contexte spécifique de la situation actuelle. Si c'est le cas, elle deviendra la décision. Dans le cas contraire, l'action suivante la plus favorable est évaluée jusqu'à ce que l'on trouve une action qui produira des résultats satisfaisants. La décision n'est pas nécessairement optimale.

Les attentes (**Expectancies**) agissent comme un mécanisme de contrôle dans le processus de décision. Elles représentent des critères permettant au décideur d'évaluer

[1] En psychologie cognitive, un modèle mental est une représentation permettant de simuler mentalement le déroulement d'un phénomène pour anticiper les résultats d'une action.

l'évolution de la situation et de déterminer si des ajustements sont nécessaires en raison d'un contexte changeant ou de résultats inefficaces.

La Figure 1-6 représente le RPD dans sa totalité avec les trois arbres de décisions différents. Le premier arbre souligne le RPD dans sa forme la plus simple, où le décideur reconnaît une situation familière qui requiert une réaction évidente à mettre en œuvre.

Le deuxième arbre décrit le processus RPD pour des situations où il y a un degré intermédiaire de complexité. Le décideur fait face à une situation non familière qui le force à se lancer dans un processus de diagnostic afin de développer l'ensemble des attentes nécessaires pour sélectionner une ligne d'action dans son répertoire.

Le troisième arbre montre comment le RPD peut être utilisé dans des situations complexes. Dans cette version, le décideur est familier avec la situation, bien qu'il soit peut-être dans un contexte nouveau, et a donc une série d'attentes qui peut être soit confirmée soit non atteinte. Si les attentes sont non atteintes, le décideur peut avoir besoin de plus d'informations avant de prendre une décision, ce qui peut l'amener à réévaluer l'ensemble de la situation. Une fois que le décideur a confirmé ses attentes, il élabore un plan d'action, procède à une simulation mentale, puis le modifie, si nécessaire et met en œuvre le plan ou, éventuellement, le raye et élabore un plan d'action entièrement nouveau. Une situation inconnue nécessitant une nouvelle ligne peut être représentée par une combinaison du deuxième et troisième arbre.

1.3.3.1. La simulation mentale

La simulation mentale est l'un des points les plus importants du RPD. Elle peut être considérée comme la construction cognitive de scénarios hypothétiques ou comme une reconstruction de scénarios réels. Il peut s'agir de répéter des événements futurs probables, d'imaginer des événements futurs moins probables, de revivre de manière réaliste des événements passés ou de reconstruire des événements passés en y incorporant des éléments hypothétiques. La simulation mentale est le processus utilisé pour évaluer en série les actions, si une évaluation d'un plan d'action est nécessaire [41]. Dans le RPD, la simulation mentale est utilisée pour :

(1) Observer la situation afin d'en prendre conscience : si les événements perçus sont compris, une situation de prise de décision pourra être établie. Les décideurs possèdent une image mentale dans laquelle les différents éléments s'imbriquent. Si à contrario les événements perçus ne sont pas compris, la situation ne peut pas être expliquée, la connaissance de la situation est insuffisante.
(2) Générer des attentes afin de vérifier la connaissance de la situation, et
(3) Evaluer un plan d'actions.

De manière simplifiée, les trois points évoqués correspondent à des phases de perception, compréhension et projection dans le futur. La phase de perception permet l'identification des éléments clés qui définissent une situation de prise de décision. La phase de compréhension permet la compréhension de la situation actuelle en termes de prise de décision. La projection dans le futur consiste à anticiper l'évolution prévue ou attendue de la situation actuelle.

Klein [42] a aussi identifié un certain nombre de points clés de la simulation mentale :

- Elle permet d'expliquer comment les événements passent du passé au présent et de projeter comment le présent permet de passer au futur.
- Construire une simulation mentale implique de former une séquence d'actions dans laquelle un état de fait est transformé en un autre.
- En raison des limites de la mémoire, les personnes construisent généralement des simulations mentales en utilisant environ trois variables autour de six transitions.
- Une grande expérience est nécessaire pour construire une simulation mentale utile.

La simulation mentale peut se heurter à des difficultés lorsque la situation devient trop compliquée ou lorsque la pression du temps ou d'autres facteurs interfèrent de façon prépondérante sur la situation.

1.3.4. Le Recognition Primed Group Decision-Making (RPgD) pour les décisions de grand groupe

Comme montré précédemment, le RPD est difficile à mettre en œuvre à l'échelle d'un groupe [29]. Le modèle RPD est plus adéquate pour des situations individuelles ou des petites unités que pour les décisions de grand groupe. Cela souligne un potentiel besoin pour le RPD d'être généralisé et adapté pour des groupes plus larges.

Tandis que le modèle RPD se concentre sur un seul décideur ou des petites unités, le RPgD (Recognition Primed Group Decision) modèle se concentre sur les décisions de groupe. Dans son modèle décrit dans la Figure 1-7, Slavkovik et Boella étendent le RPD et la décision en situation d'un seul agent à une équipe hiérarchique multi-agent [43]. Dans cette équipe hiérarchisée, le leader (initiator) est l'agent responsable de la décision tandis que les autres membres (executors) sont chargés d'informer pour la prise de décision et de l'exécuter [44].

Le RPgD a été développé pour être utilisé dans une équipe composée d'humains, de machines et potentiellement de robots utilisant de l'Intelligence Artificielle.

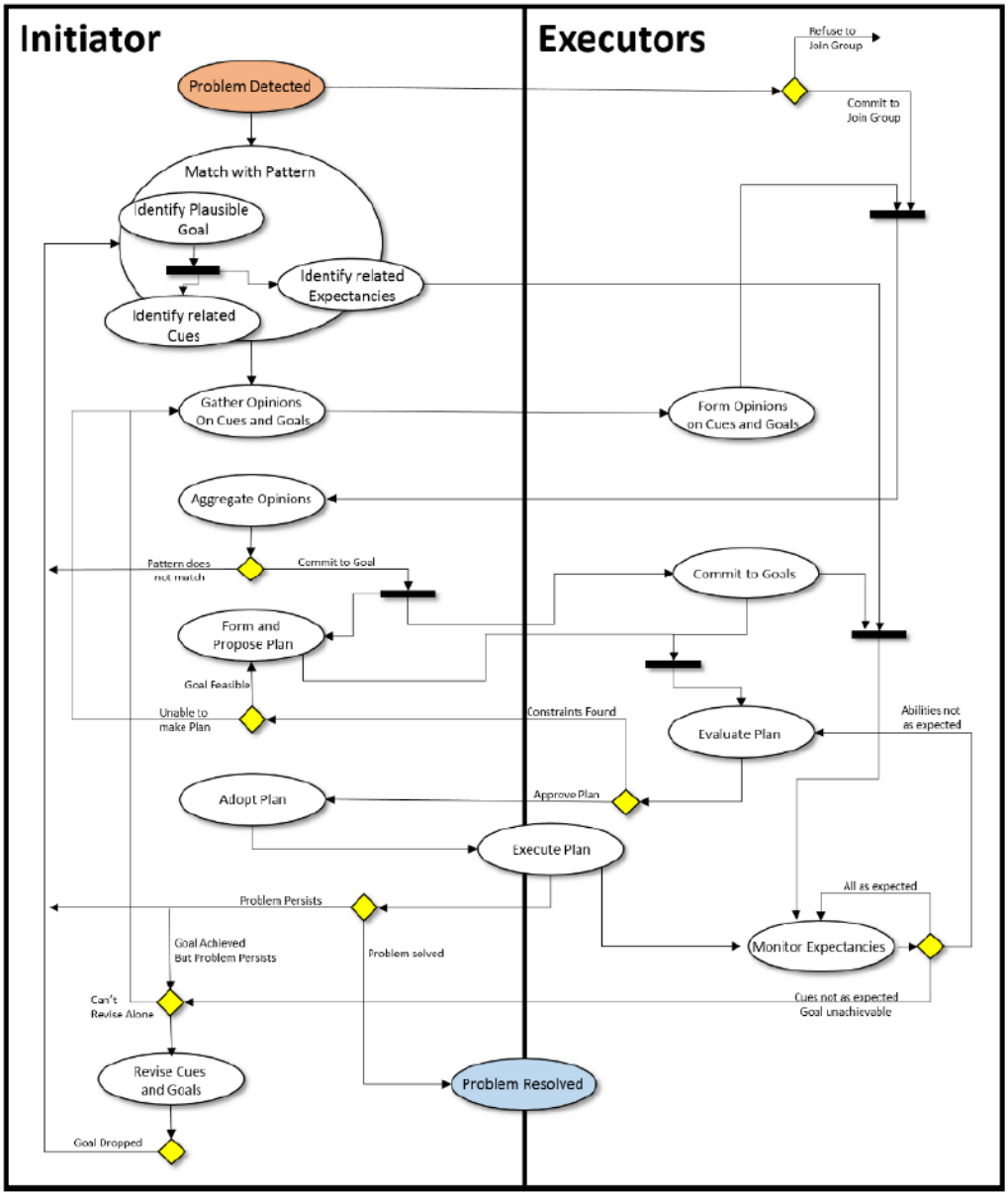

Figure 1-7 : Diagramme de décision détaillé developpé par Slavkovik et Boella pour décrire le processus RPgD [43]

1.3.5. Les limites du RPD

L'application réussie du RPD dépend néanmoins de deux variables clés : **les renseignements et l'expérience**. Ces variables appuient les deux principales actions que les décideurs exécutent, la reconnaissance de la situation et le choix des mesures à prendre. La première action, la reconnaissance de la situation, nécessite des renseignements pour fournir des indices au décideur. Ces renseignements peuvent prendre la forme de rapports de renseignements, de mises à jour de la situation ou d'actions que le décideur observe directement. La deuxième action, le choix de la ligne de conduite exige une expérience

pertinente. Le décideur doit disposer d'un cadre de référence ou d'une base de connaissances qui lui permette de reconnaître les similitudes entre ses précédentes expériences et la situation qu'il analyse [45].

Une analyse plus approfondie de cette dépendance à l'égard de l'information et de l'expérience révèle les failles potentielles des modèles intuitifs de prise de décision. Les décideurs qui utilisent le processus RPD pourraient fonder leurs décisions sur des renseignements qui ne sont pas parfaits ou ils pourraient ne pas avoir l'expérience requise pour prendre les décisions adaptées à une situation. Par exemple, un rapport inexact d'un subordonné ou un ennemi qui emploie une nouvelle tactique pourrait amener le Commandant à percevoir des indices trompeurs, l'amenant à ordonner un déploiement prématuré de sa réserve. Bien que le Commandant veuille prendre une bonne décision, ses propres perceptions sont soumises à son interprétation. Les psychologues décrivent la limitation des perceptions d'un décideur comme une **rationalité limitée**. Le concept de rationalité limitée suggère que même si les décideurs veulent prendre la meilleure décision possible, la qualité de l'information qu'il reçoive, leurs perceptions de cette information et de sa source, et leur capacité à traiter l'information les en empêchent souvent [46].

Dans le contexte du combat tactique, la rationalité limitée concerne **l'incertitude**, **l'information, les inhibiteurs, les attentes et l'expérience** qui limitent les décisions du Commandant.

L'incertitude limite le bien-fondé des décisions, en amenant le Commandant à mal juger certains des éléments d'une situation, comme le temps. L'incertitude fait référence à la perception qu'ont les Commandants de l'exactitude ou de l'exhaustivité de l'information, et se manifeste habituellement lorsque l'information n'est pas disponible ou que les situations ne sont pas familières. Par exemple, si une unité de reconnaissance ne parvient pas à détecter un ennemi, le Commandant peut croire qu'il a plus de temps pour se préparer ou réagir que ce qu'il a réellement. L'incertitude peut également survenir si le Commandant n'est pas certain de la crédibilité d'une source d'informations [47]. Même si l'information est disponible, les propres caractéristiques mentales d'un Commandant peuvent l'empêcher de la traiter correctement, ce qui l'empêche de reconnaître le besoin ou la possibilité de changement. Ces inhibiteurs d'information comprennent la personnalité du Commandant et le stress. Par conséquent, si un EM n'est pas au courant des besoins du Commandant et de son état de stress, il peut ne pas fournir les indices nécessaires qui permettent à un Commandant de reconnaître intuitivement qu'un changement est nécessaire.

Une autre catégorie de limitations concerne les attentes du Commandant, qui peuvent entraîner une mauvaise identification des indices ou des modèles. L'EM décrit ses attentes en termes de missions assignées et d'expression des résultats souhaités. Cette description, à son tour, encadre l'interprétation de l'information [48].

Une autre limitation, le niveau d'expertise du Commandant peut créer une forte réticence au changement. Les psychologues cognitifs proposent trois explications possibles : la première est la capacité de traitement limitée de l'esprit humain, un individu ne peut traiter un nombre fini de variables avant que la performance ne diminue. Deuxièmement, il est suggéré que l'expertise se décline en deux variétés routinières et adaptatives. Cette suggestion affirme que si une personne peut bien fonctionner dans des situations familières, elle n'a que des capacités modestes pour faire face à des types de problèmes nouveaux [49]. La dernière possibilité est que les décideurs expérimentés soient inconsciemment réticents au changement,

et adoptent des mesures cognitives « d'évitement défensif » faussant l'interprétation d'une situation [50].

Cet examen des aspects de la rationalité limitée donne un aperçu de ce dont le Commandant a besoin pour prendre des décisions. Plus précisément, ce qui précède suggère quatre tendances que l'organe de commandement doit tenter d'éviter : une évaluation inexacte du temps, le fait de ne pas reconnaitre le besoin ou la possibilité de changement, la mauvaise identification des indices et des modèles et la réticence à changer.

1.4. Conclusion

Les militaires de la Marine Nationale utilisent la MEDOT, méthode analytique de prise de décision, pour planifier et conduire une mission au niveau tactique. Néanmoins, ils sont aujourd'hui confrontés à des situations tactiques de plus en plus complexes où ils manquent de temps et de ressources pour appliquer correctement des méthodes analytiques. En conduite pour prendre des décisions rapides, les militaires n'utilisent que les informations les plus pertinentes recueillies par les capteurs du navire et mettent en œuvre la première solution « satisfaisante » qu'ils auront trouvée sans comparer toutes les alternatives entre elles. Cette manière de raisonner correspond au RPD, issu des méthodes NDM ou intuitives permettant de prendre des décisions rapides dans des environnements incertains et contraints par le temps. Le PRD permet au décideur de s'adapter et d'agir rapidement à des situations changeantes, de choisir rapidement un plan d'actions satisfaisant et d'ainsi diminuer le temps de prise de décision et les ressources humaines nécessaires. Malgré ces avantages, le RPD ne garantit pas que la décision optimale soit prise ; la conciliation entre différents objectifs est difficile à prendre en compte, et le RPD est réputé inadapté pour les décideurs inexpérimentés confrontés à une situation inédite. Pourtant, le fait qu'une situation soit nouvelle ne renforce pas nécessairement l'avantage relatif d'un modèle analytique par rapport à un modèle intuitif, ce qui démontre que le RPD n'est pas plus inadapté aux décideurs inexpérimentés que les modèles analytiques.

Cependant, même si les militaires en conduite n'ont plus le temps d'appliquer une méthode analytique et préfèrent s'appuyer sur une stratégie RPD pour répondre aux nouvelles exigences des situations tactiques, les capacités humaines soumises à la rationalité limitée et à la surcharge cognitive telle que la surcharge d'informations et les nombreuses prises de décisions ne permettent plus de répondre aux exigences de manière aussi efficace et rapide. L'émergence de ces situations tactiques de plus en plus complexes exige un soutien crucial du Commandant par des systèmes d'aides à la décision appropriés. Lors de la prise de décision, le Commandant recherche la supériorité de l'information, c'est-à-dire l'acquisition des données pertinentes, crée une prévision des opportunités ou des risques, puis prend sa décision. Pour ce faire, il a besoin d'un système de soutien pour améliorer sa connaissance de la situation en lui présentant une image explicative et en appuyant l'évaluation de la situation, y compris la prédiction et l'évaluation de la situation future, dans lequel il aura confiance et de ce fait réduira sa charge mentale. C'est pourquoi il a été proposé d'implémenter des systèmes d'aide à la décision fondés sur le RPD. En proposant un système qui s'appuie sur les mêmes informations traitées par les humains et qui préconise une solution « satisfaisante » adaptée à la compréhension rapide du décideur, il sera dans la capacité de gagner la confiance du décideur et donc de l'aider à prendre ses décisions tout en réduisant sa charge cognitive.

Le prochain chapitre établit un état de l'art des systèmes d'aides à la décision existants fondés sur le RPD, ainsi que leur utilisation au sein des missions militaires et leurs limites. Enfin, l'apprentissage par renforcement sera introduit pour pallier certaines limites des SAD existants.

Chapitre 2 : Systèmes d'aide à la décision fondés sur le RPD et apprentissage par renforcement

L'environnement militaire tactique est soumis à la contrainte du temps, et exige la collecte rapide d'informations pertinentes, changeantes et incertaines [51]. Dans ces conditions, l'organe de commandement de la mission n'a plus les ressources nécessaires ni le temps pour appliquer correctement des méthodes analytiques (ex : MEDOT). Plusieurs recherches portant essentiellement sur l'armée américaine ont montré que les Commandants militaires emploient le RPD dans la majorité de leurs décisions militaires en conduite de mission. Par exemple, dans leur étude, Kaempf et *al.* [52] ont examiné les décisions de commandement et de contrôle prises par l'équipe de lutte anti-aérienne sur un navire de la marine américain. Ils ont constaté qu'une stratégie de RPD était utilisée dans 95% des situations de décisions auxquelles l'équipe de lutte anti-aérienne avait été confrontées. Cependant, malgré les avantages de mobiliser le RPD pour la prise de décision rapide, les capacités humaines sont aujourd'hui menacées par la surcharge cognitive et la « rationalité limitée » qui introduisent certains biais cognitifs, entravant la qualité et la rapidité de la prise de décision [53].

Il a donc été important d'étudier des idées nouvelles et de développer des solutions efficaces pour réduire les risques de la surcharge d'informations, de la « rationalité limitée » et de la pression liée au temps à un niveau acceptable.

La création de systèmes d'aide à la décision reproduisant le RPD permet d'aider d'une part les décideurs à prendre en compte l'hétérogénéité et la complexité de l'environnement de décision, et d'autre part d'avoir un système mobilisant (1) les mêmes informations pertinentes recueillies par les opérationnels, (2) le même raisonnement mobilisé lors de la prise de décision dans l'urgence pour préconiser des décisions « satisfaisantes », intuitives, adaptées au décideur qui ne mobilisera pas de ressources cognitives supplémentaires pour les comprendre. Un tel système aura la capacité de gagner la confiance de ses utilisateurs, de réduire leur surcharge cognitive et leur temps de prise de décision. A ce titre, plusieurs SAD du RPD existent déjà ou sont en cours de développement.

Ce chapitre va dresser dans une première partie un recensement des SAD existants avec leur utilisation au sein de missions militaires ainsi que leurs limites. Cette première partie permettra de déterminer un choix de SAD sur lequel s'appuiera notre système, ainsi qu'un choix technologique pour modéliser le RPD. Dans une seconde partie, un état de l'art sur l'apprentissage par renforcement et plus spécifiquement sur l'apprentissage par renforcement profond sera introduit afin de justifier l'utilisation de ce dernier pour surpasser certaines limites existantes des SAD, et ainsi être en mesure de proposer une modélisation du RPD.

2.1. Systèmes d'aide à la décision basés sur le RPD

Comme évoqué dans le chapitre introductif, les SAD s'appuient souvent sur des agents intelligents issus de l'IA. Un agent intelligent est une entité autonome qui agit en orientant son activité pour réaliser des objectifs (rôle d'un agent), dans un environnement en utilisant l'observation par des capteurs et des actionneurs conséquents (rôle d'un agent intelligent) [11].

Les agents intelligents peuvent également apprendre ou utiliser des connaissances pour atteindre leurs objectifs. Un SAD peut être composé d'un ou plusieurs agents. Dans le cas où il en possède plusieurs, le système est qualifié de multi-agents (SMA). Les systèmes multi-agents peuvent résoudre des problèmes qui sont difficiles ou impossibles à résoudre pour un seul agent.

Les SAD peuvent aussi s'appuyer sur des architectures cognitives. Une architecture cognitive est un système informatique représentant les processus cognitifs, notamment l'attention, la mémoire et la correspondance de modèles, organisés selon une structure analogue à celle du cerveau humain [54]. Elle fournit une plateforme pour développer des modèles de comportements humains et prédictifs des performances humaines permettant de reproduire et d'atténuer les limitations du raisonnement humain telles que les biais cognitifs.

2.1.1. Systèmes basés sur un ou plusieurs agents

La partie suivante dresse l'inventaire des SAD fondés sur le RPD composés d'un ou plusieurs agents.

2.1.1.1. SAD à base de réseaux de neurones

Les réseaux de neurones [55] peuvent être utilisés avec succès pour reconnaître des schémas sous-jacents dans des données. Cela pourrait être un outil utile pour implémenter la partie reconnaissance du RPD. Cependant, dans le contexte de l'apprentissage supervisé [56], pour atteindre une telle performance une quantité significative de données d'entraînement est nécessaire pour préparer le réseau à reconnaître des situations dans différents scénarios militaires. Les réseaux de neurones devront être couplés avec d'autres techniques pour leur permettre d'apprendre de nouvelles situations. Après que le réseau de neurones ait reconnu la situation, d'autres traitements seront requis pour déterminer un plan d'actions satisfaisant. Cette étape pourrait être accomplie par un autre réseau de neurones ou d'autres techniques logiques comme le raisonnement basé sur des règles ou encore la logique floue [57].

Une étude [57] a été réalisée pour implémenter le RPD à l'aide de réseaux de neurones. Le réseau de neurones accomplit la surveillance de la situation et la sélection du plan d'actions. Pour entrainer le réseau, 12 scénarios différents ont été présentés à 12 experts militaires. Il leur a été demandé de concevoir des plans pour accomplir chaque scénario. Les données de départ du scénario et les plans générés par les experts ont été échantillonnés sous la forme de données d'entraînement. Une fois le réseau entraîné, les chercheurs lui ont donné des scénarios inédits, et ont demandé aux experts militaires d'analyser les solutions du réseau. Les résultats de ces tests ont montré que les réseaux de neurones étaient une technique viable pour implémenter la partie reconnaissance d'expérience du RPD.

Cependant, dans cette étude le RPD n'a pas été reproduit dans sa totalité. En effet, la simulation mentale, une étape clé du RPD, n'a pas été implémentée. De plus, les données d'entrée du scénario utilisées étaient parfaites. En réalité, un Commandant militaire aura

rarement des données parfaites pour l'aider à reconnaître la situation. Ces facteurs doivent être pris en compte pour avoir un modèle plus complet et plus précis de prise de décision.

Robichaud [58] a amélioré le réseau de neurones précédent en y ajoutant des variables floues et des règles d'inférence floues. L'interprétation floue de l'environnement extérieur a été ajoutée pour prendre en compte d'une part les fluctuations des valeurs mises en jeu, et d'autre part la manière dont les humains interprètent les données provenant de leur environnement extérieur. Cet ajout a permis d'améliorer la mise en œuvre du RPD avec des réseaux de neurones.

2.1.1.2. SAD à base de structure LTM

Warwick et *al.* [59] ont implémenté la partie reconnaissance du RPD en encodant l'expérience d'une personne dans une structure Long Term Memory (LTM). Ce type de couche de neurones permet de prendre en compte les informations passées. Lorsqu'un agent expérimente son environnement, il laisse une trace d'expérience. Ces traces composées de variables sur la situation et de décisions sont stockées dans la base LTM et représentent la totalité des expériences de l'agent. Lorsqu'une nouvelle situation de décision est présentée, une valeur de similarité [59] est calculée et utilisée pour « reconnaître » les expériences et le plan d'actions associé, stockés dans le LTM.

La structure LTM a été utilisée par des chercheurs de la NASA qui ont développé le prototype MOCOG1 [60]. Dans cette simulation, le RPD a été implémenté en utilisant des heuristiques. Cependant, étant basé sur des règles, le modèle est limité dans sa décision par l'ensemble des règles explicites programmées. Il a été utilisé dans un environnement statique et ne pouvait pas être utilisé pour simuler l'environnement dynamique complexe d'une guerre au niveau tactique.

2.1.1.3. SAD à base d'agents composites

Les systèmes multi-agents sont composés généralement de nombreux agents de haut niveau opérant dans un environnement commun et partagé. Les agents résidant dans cet « environnement extérieur » interagissent entre eux, ainsi qu'avec les objets de l'environnement. Ils interprètent les données sensorielles et prennent des décisions quant aux actions à entreprendre. Ces actions affectent à leur tour l'environnement. Afin de reproduire les fonctionnalités de ces agents, tout en simplifiant la conception d'un système complexe, Hiles et *al.* [61] ont développé le concept d'agent composite (Composite Agent, CA).

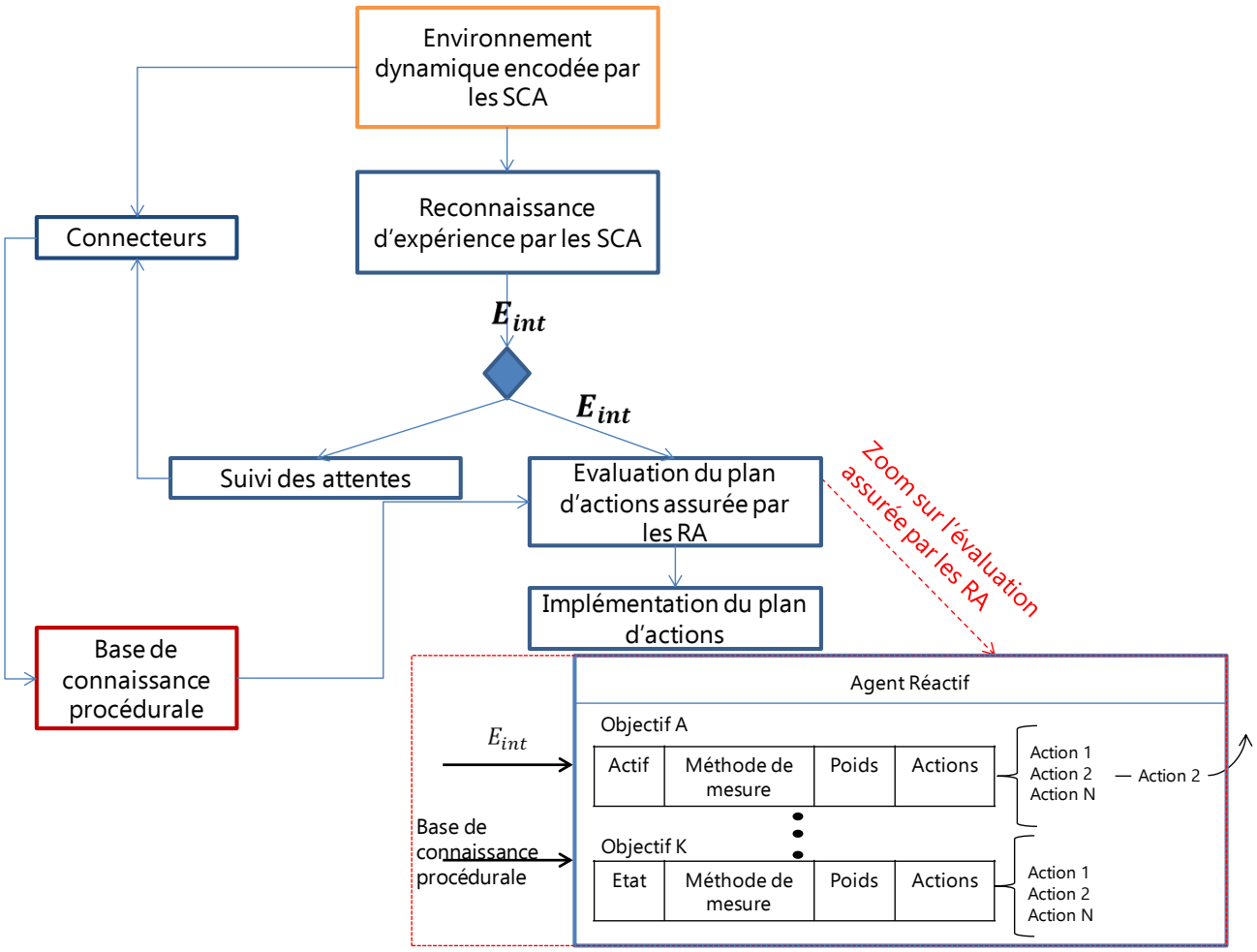

Figure 2-1 : Agent composite [61]

Un CA est un système multi-agent composé d'agents symboliques (Symbolic Constructor Agents, SCA) et d'agents réactifs (Reactive Agent, RA). Les agents composites sont construits à partir des forces de ces deux types d'agents pour produire un seul méta-agent capable de comportements complexes, comme le décrit la Figure 2-1. Les agents symboliques perçoivent l'environnement extérieur E_{ext}, et le convertissent en une représentation interne E_{int}, de la même façon que les humains se font leur propre représentation interne de leur environnement. Les SCA contrôlent et filtrent également les informations de E_{ext} afin que le CA ne soit pas submergé par des entrées sans importance. Les SCA sont associés à un ensemble d'agents réactifs.

2.1.1.3.1. Les agents réactifs et la gestion des objectifs

Les agents composites contiennent de nombreux agents réactifs (RA), où chaque agent réactif est responsable d'un comportement spécifique de l'agent composite. L'ensemble des RA pris en groupe, définit l'ensemble des comportements de haut niveau de l'agent composite. Les RA fonctionnent dans le monde de l'environnement intérieur. Ils prennent en entrée les informations sensorielles d'E_{int} et produisent en sortie les actions que l'agent doit effectuer.

Chaque RA a un ou plusieurs objectifs spécifiques pour faire progresser son comportement ou sa fonction. Ainsi, à tout moment, de nombreux objectifs sont en concurrence pour attirer l'attention de l'agent composite. Tout comme les humains ont des objectifs multiples, un agent a lui aussi plusieurs objectifs qu'il souhaite satisfaire. Dans la prise de décision humaine, les objectifs changent constamment de priorité, en fonction du contexte et de l'état de la personne. Les agents peuvent imiter la flexibilité et les capacités de substitution de la prise de décision humaine grâce à l'utilisation d'un appareil de gestion des objectifs variables au sein des RA. C'est de cet appareil de gestion des objectifs qu'émerge un comportement intelligent et adapté au contexte. Les RA interprètent l'environnement intérieur symbolique et, grâce à leur appareil de gestion des objectifs, traitent ces informations pour équilibrer leurs objectifs et renvoyer une action appropriée pour atteindre le ou les objectifs prioritaires tel qu'indiqué Figure 2-1.

Les objectifs ont quatre composantes : un état, une méthode de mesure, un poids et une action ou un ensemble d'actions pour atteindre l'objectif. L'état de l'objectif indique si un objectif est dans un état actif, inactif ou dans un autre domaine spécifique. La méthode de mesure traduit l'information sensorielle reçue par le RA en une mesure quantifiable de la force actuelle d'un objectif et de son degré de réalisation. Cela permet à un agent de hiérarchiser les objectifs et d'ajuster les états des objectifs en fonction du contexte.

Les agents peuvent de façon sélective rejeter les comportements qui ne favorisent pas la réalisation de leurs objectifs, et augmenter le recours aux comportements qui se sont avérés efficaces pour atteindre les objectifs. Ce comportement sert de système d'apprentissage réactif où l'agent apprend de l'environnement, en se basant sur ce qui fonctionne sans expertise ni intervention humaine. Le changement d'objectif basé sur un environnement en évolution dynamique produit un comportement innovant et adaptatif, cependant, un équilibre doit être fait avec des actions doctrinalement correctes et appropriées. Cet équilibre est atteint grâce à l'encodage des connaissances procédurales dans une structure de données appelée « tickets ».

2.1.1.3.2. Tickets

Les SCA et l'appareil de gestion des objectifs ont été développés pour contrôler la capacité sensorielle et la prise de décision de l'agent. Afin de fournir aux agents une riche base de connaissances procédurales tout en favorisant un comportement adaptatif, une structure de données appelée tickets a été développée. Les tickets permettent aux agents réactifs d'appliquer les connaissances procédurales dans leur contexte. Ils définissent l'ensemble des actions de l'agent, c'est-à-dire ses moyens pour atteindre ses objectifs. Ils sont utilisés pour organiser les connaissances procédurales.

A chacun des objectifs d'un agent sont liés un ou plusieurs tickets qui définissent comment atteindre les objectifs. Les tickets peuvent comporter des conditions préalables ou des conditions connexes qui doivent être remplies pour qu'un ticket soit actif.

Encoder les connaissances procédurales et les relier à divers objectifs pour créer un comportement intelligent n'est cependant pas suffisant. La méthode la plus appropriée des procédures pour une situation donnée doit être appliquée. Or, dans l'étude d'un système dynamique, la situation donnée par l'état courant non seulement change constamment, mais est

aussi si complexe à représenter que le concepteur du système ne peut pas prendre en compte toutes les possibilités. Par conséquent, le mécanisme de détermination des procédures les plus appropriées doit être flexible et capable de supporter le même niveau de complexité que les contextes changeants du système dynamique. La capacité à prendre les mesures correctes pour s'adapter à la situation est assurée par l'utilisation de « connecteurs ».

2.1.1.3.3. Connecteurs

Les connecteurs représentent des travaux basés sur des types symboliques. Ils permettent des substitutions et des enchaînements logiques, et facilitent l'explication des raisonnements. Les connecteurs sont un moyen d'associer des impressions, des idées et des actions à un contexte donné et d'obtenir une séquence logique de comportement. Les connecteurs sont des objets actifs qui ressentent et réagissent à l'environnement. Ils s'activent et se désactivent en fonction du contexte actuel. Lorsque l'état de l'agent et l'état de l'environnement changent, les connecteurs détectent les changements et s'étendent ou se rétractent en conséquence. En reliant des connecteurs à divers éléments du système, y compris les tickets, les connecteurs signalent l'état de préparation des éléments et leur niveau d'adéquation à la situation actuelle. Cette liaison est basée non seulement sur l'état de l'environnement, mais aussi sur les objectifs de l'agent et ses interactions sociales avec les autres agents. De cette manière, les connaissances procédurales correctes peuvent être mises à profit dans la situation correcte.

2.1.1.3.4. Agents composites et RPD

Les agents composites permettent de modéliser le RPD pour plusieurs raisons. Tout d'abord, le RPD commence par le décideur qui vit une situation dans un contexte changeant et essaie de déterminer si la situation correspond étroitement à une expérience passée. Dans les CA, les SCA remplissent cette fonction. Ils examinent l'environnement extérieur et produisent une représentation interne de la situation en échantillonnant périodiquement l'environnement pour rechercher un contexte changeant.

Lorsque la reconnaissance de la situation s'effectue, le RPD fournit quatre sous-produits : les objectifs, les indices, les attentes et les actions. Les décideurs s'appuient sur des indices pour se concentrer sur les variables importantes d'une situation. Les SCA accomplissent la même tâche. Ils contrôlent et filtrent les informations de E_{ext} afin que le CA ne soit pas submergé d'entrées sans importance.

Les objectifs jouent un rôle central dans le RPD. Ils aident le décideur à se concentrer sur la tâche à accomplir et à orienter le choix des mesures à prendre pour résoudre un problème. De même un CA est axé sur les objectifs. Le RPD indique qu'un décideur se concentrera sur un seul plan d'actions à la fois afin de déterminer s'il correspond ou peut s'adapter à la situation actuelle de décision. Le CA ne génère pas de multiples plans d'actions. Il permet plutôt à un ensemble d'actions dominantes d'émerger des différents objectifs concurrents. Sur la base de

son expérience programmée, le CA sélectionne des actions pour produire une solution satisfaisante et n'essaie pas de trouver la meilleure.

Un CA gère les attentes en surveillant périodiquement les changements dans son environnement causés par des effets externes ou résultant de ses actions. La surveillance permet au CA de déterminer les effets de ses actions et de décider si ces effets sont attendus. La surveillance permet également au CA d'ajuster ses décisions en fonction de l'environnement dynamique extérieur. Ce processus est similaire au processus de surveillance du RPD.

Le dernier point à aborder concerne l'utilisation de la simulation mentale au sein du RPD pour modifier un plan d'actions sélectionné afin de l'adapter au contexte spécifique d'une situation de décision. Un CA accomplit partiellement la simulation mentale lorsqu'il effectue son processus de gestion des objectifs pour sélectionner l'ensemble des actions qu'il mènera. Toutefois, il n'existe pas de mécanisme clair au sein du CA pour modifier ses expériences existantes afin de fournir une solution de décision mieux adaptée. Le processus de simulation mentale devra être amélioré pour mieux reproduire le rôle de la simulation mentale au sein du RPD.

Un CA a été testé sur un scénario militaire [62]. Le scénario consistait en une série de quatre points de décision liés entre eux par un scénario d'assaut amphibien. Il fournissait des décisions typiques de celles auxquelles sont confrontés les Commandants lors d'opérations navales. Les points de décision consistaient en la détermination du lieu et du moment du débarquement amphibie, une modification du moment du débarquement en fonction de nouvelles informations, et une décision sur la poursuite de l'assaut en fonction d'une opposition ennemie imprévue ou d'un taux de pertes amies élevé. Les algorithmes ont été validés [62] par rapport aux décisions produites par des officiers militaires jouant le rôle de Commandant d'une force opérationnelle interarmées dans un scénario de décision opérationnelle.

Afin d'acquérir l'expérience nécessaire pour encoder l'agent, une analyse cognitive des tâches d'attaques amphibie historiques a été réalisée. Cette analyse fournit une méthode permettant de solliciter les connaissances et l'expérience d'experts en la matière [63]. Une analyse des attaques amphibie depuis la Seconde Guerre mondiale jusqu'à à la guerre du Golfe persique a été réalisée. Cet examen a fourni les informations nécessaires pour identifier les éléments d'expérience associés à ce type d'action militaire. Ces éléments ont été introduits dans la base de données d'expérience de l'agent.

Le scénario a ensuite été fourni à trente officiers militaires. Vingt et un étaient des officiers militaires américains et neuf des officiers provenant de divers pays. Tous avaient une expérience militaire opérationnelle commune. Leurs réponses aux points de décision du scénario ont été enregistrées et sont devenues l'ensemble de décisions par rapport auxquelles le modèle a été évalué.

L'évaluation finale a montré que, sur la base des statistiques d'équivalence pour chaque point de décision et des résultats du test de Turing [64], le modèle produisait des décisions équivalentes à celles des acteurs. Par conséquent, le modèle a pu imiter le processus de décision cognitif utilisé par ces officiers militaires jouant le rôle de Commandant d'une force opérationnelle interarmées. Il est à noter que l'agent a été capable de reproduire la variabilité de la prise de décision humaine et peut donc fournir des décisions adaptées à celles envisagées par les officiers.

2.1.1.4. SAD à base d'agents BDI

Le modèle BDI est théoriquement défini selon les trois composantes, croyances (Beliefs), désirs (Desires) et intention (Intention) [65]. Le modèle tente de saisir la compréhension commune de la façon dont les humains raisonnent à travers : les croyances qui représentent les connaissances de l'individu sur l'environnement et sur son propre état interne ; les désirs ou plus spécifiquement les objectifs ; et les intentions qui sont l'ensemble des plans ou la séquence d'actions que l'individu entend suivre afin d'atteindre ses objectifs. Les principales fonctionnalités que les systèmes BDI mettent en œuvre sont :

- une représentation des croyances, des désirs et des intentions;
- un processus rationnel et logique de sélection des intentions;
- un engagement flexible et adaptable à l'ensemble des intentions.

Un nombre important de systèmes BDI existe, par exemple JACK [66], Jason [67], 3APL [68], qui permettent de créer des agents en utilisant ces trois concepts et plusieurs systèmes les élargissent pour fournir des capacités de raisonnement humain supplémentaires. Le point commun à toutes ces implémentations est le concept de bibliothèque de plans, qui est une collection de plans que l'agent est capable d'exécuter. Chacun de ces plans est constitué d'un corps, d'un objectif que le plan est capable d'atteindre et d'un contexte dans lequel le plan est applicable. Le corps du plan peut comprendre à la fois des actions qui peuvent directement affecter l'environnement, et des sous-objectifs qui seront étendus à d'autres plans si nécessaire. Cette imbrication des objectifs dans les plans permet de construire une structure hiérarchique plan-objectif, avec des branches à partir d'un nœud objectif menant à des plans capables d'atteindre cet objectif, et des branches à partir d'un nœud de plan menant à des objectifs qui doivent être atteints pour compléter ce plan.

Les agents BDI contiennent également un moteur d'exécution, qui guide le processus de raisonnement des agents :

1. Perception : trouve tout nouvel événement susceptible d'avoir été déclenché, soit au sein de l'agent, soit en dehors de l'environnement.
2. Mise à jour : mettre à jour les croyances avec les nouvelles informations fournies par cet événement ;
3. Révision de l'intention : si le changement de croyances signifie qu'une intention ou un objectif n'est plus valable soit parce qu'il a déjà été atteint, soit parce qu'il n'est plus possible, il faut alors le retirer de l'ensemble.
4. Filtrage de plans : Si l'intention a été révisée, déterminer l'ensemble des plans applicable à l'événement en cours et approprié au contexte ;
5. Sélection du plan : s'il n'y a pas de plan en cours, sélectionner un nouveau plan dans cet ensemble et ajouter les éléments de son corps à la liste des intentions ;
6. Action : exécuter l'étape suivante du plan en cours, ce qui peut impliquer l'exécution d'une action ou l'extension d'un sous-objectif en déclenchant un nouvel événement.

2.1.1.4.1. *La modélisation de Lui et al.*

Lui et *al.* [69] ont utilisé le système JACK [66] pour modéliser les commandants militaires dans des scénarios d'opérations terrestres. Ils ont étudié un scénario d'attaque de compagnie

dans lequel l'agent devait contrôler une compagnie d'infanterie motorisée. Sans agent, les objectifs sont donnés à un opérateur qui suit la doctrine militaire et les procédures standards pour produire des plans d'actions à prendre pendant la bataille. En utilisant des agents BDI, les objectifs de l'attaque seront des apports aux agents. Les agents utiliseront ensuite leur raisonnement et leurs règles, basés sur la doctrine militaire, pour produire des plans et des plans d'action pour réaliser les quatre phases de l'attaque, 1) la préparation, 2) l'assaut, 3) l'exploitation et 4) la réorganisation. Pour mettre au point cet agent BDI, il faut avoir à disposition :

- La phase préparatoire d'un scénario d'attaque modélisée en partant du principe qu'il y avait des croyances et des hypothèses préconçues sur les positions de l'ennemi. Cela a permis aux agents BDI de produire des plans et des plans d'action en utilisant la capacité de raisonnement basé sur des règles.
- Les données de terrain et l'algorithme de recherche d'itinéraire ont dû être mis en œuvre et intégrés à l'agent BDI pour déterminer les itinéraires vers les cibles et objectifs potentiels. À ce stade, l'agent BDI pouvait assigner des tâches comme déplacer des véhicules, attaquer une cible, occuper une zone aux unités d'infanterie motorisées.
- L'acquisition des connaissances s'est faite par la consultation du personnel militaire. Les règles de la phase préparatoire sont acquises et peuvent être utilisées dans l'agent.

Cependant, l'implémentation du BDI en tant que RPD est concentrée sur une modélisation précise de la connaissance de la situation et de la réalisation des objectifs [58]. Elle n'inclue pas d'autres aspects du RPD tels que la personnalité et la simulation mentale.

2.1.1.4.2. *La modélisation de Danial et al.*

Danial et *al* [70] propose une méthodologie pour implémenter un agent artificiel capable de prendre des décisions basées sur le RPD. La méthodologie proposée modélise les principes du RPD en utilisant la logique Belief-Desires-Intention. Le RPD considère la prise de décision comme une synthèse de trois capacités principales de l'esprit humain. La première est l'utilisation de l'expérience pour reconnaître une situation et suggérer des réponses appropriées. La principale préoccupation ici est la prise de conscience de la situation, car le décideur doit établir qu'une situation actuelle est identique ou similaire à une situation antérieure, et que la même solution est susceptible de fonctionner cette fois aussi. À cette fin, l'approche de modélisation proposée utilise un réseau logique de Markov [71] pour développer un module d'apprentissage par l'expérience et d'aide à la décision.

La deuxième composante du RPD traite des cas où l'expérience d'un décideur devient secondaire car la situation n'a pas été reconnue comme typique. Dans ce cas, le RPD suggère un mécanisme de diagnostic qui implique la correspondance des caractéristiques et, par conséquent, une approche de raisonnement basée sur l'ontologie (du domaine d'intérêt) est proposée ici pour traiter tous ces cas.

La troisième composante du RPD est la proposition selon laquelle les êtres humains utilisent l'intuition et l'imagination (stimulation mentale) pour s'assurer qu'un plan d'action fonctionne ou non dans une situation donnée. La simulation mentale est ici modélisée sous la forme d'un réseau bayésien qui calcule la probabilité d'occurrence d'un effet lorsqu'une cause est plus probable.

Smith [72] a réalisé une expérience pour évaluer comment la formation à l'intervention d'urgence dans un environnement virtuel affecte la compétence humaine dans différents scénarios d'évacuation d'urgence. L'expérience de Smith a impliqué 36 participants répartis en deux groupes : Le groupe 1 comprenant 17 participants et le groupe 2, 19 participants. Les participants du groupe 1 ont été formés au cours de plusieurs sessions, tandis que les participants du groupe 2 n'ont reçu qu'une seule formation de base. Danial et *al.* ont testé leur agent artificiel sur des scénarios d'évacuation d'urgence sur des plateformes pétrolières en se basant sur l'expérience de Smith. Les données des participants du groupe 1 sont utilisées pour valider les résultats de l'agent artificiel. L'agent est censé fonctionner avec les mêmes données que celles perçues par les participants dans l'expérience de Smith. Les participants ont été testés au cours de trois sessions distinctes : S1, S2 et S3, chacune d'elles comprenant diverses sessions de formation et de test impliquant une série d'activités.

Tous les scénarios de formation et de test ont été enregistrés, et un journal de bord a été tenu pour chaque participant contenant des informations spécifiques sur la façon dont le participant a procédé dans un scénario pour prendre la décision requise. Les facteurs analysés ont été le type de décision d'urgence, ici INCENDIE ou EVACUATION, la reconnaissance des alarmes, et l'intention de se rendre à un lieu de rassemblement particulier en utilisant un itinéraire d'évacuation. Les résultats ont montré que les sorties de l'agent sont similaires aux décisions prises par les participants humains pour les mêmes indices d'entrée. Cependant malgré ces résultats, des travaux supplémentaires sont encore nécessaires pour améliorer les résultats et le modèle proposé. Le RPD comporte de nombreuses dimensions, comme l'utilisation de la simulation mentale qui n'a été modélisée que partiellement.

2.1.1.5. SAD à base de réseaux Bayésiens

Müller a développé le modèle de décision Bayésien de reconnaissance (BRDM) [73], axé sur le processus de reconnaissance des événements passés. Il utilise le terme « décision de reconnaissance » pour souligner le fait que ce modèle n'intègre pas bon nombre des aspects plus complexes décrits par le modèle RPD. Le mécanisme par lequel le BRDM identifie une situation sur la base d'informations générales et d'indices sur l'environnement est décrit ci-dessous.

1. **Modélisation de l'environnement** : Quelle que soit la situation spécifique, l'état de l'environnement sera signalé au BRDM par un capteur simulé, soit un rapporteur humain, soit un dispositif en réseau. Les propriétés par lesquelles un capteur signale l'état de l'environnement sont saisies dans un modèle probabiliste, que le modèle de décision utilisera ensuite comme modèle génératif pour tenir compte des données observées.

2. **Modélisation de l'apprentissage du décideur** : Pour que le BRDM puisse fonctionner, il doit apprendre à connaître l'environnement dans lequel il opère. Dans des situations réelles, cette expérience provient de plusieurs sources telles que des manuels, des informations transmises par des officiers, des événements réels. Pour simuler ce processus d'apprentissage, le modèle est exposé à un grand nombre d'événements pertinents. Le BRDM apprend deux choses essentielles sur ces

événements : (a) des caractéristiques typiques à chaque classe d'événement et (b) un taux de base d'apparition de chaque événement dans l'environnement. Bien que les rapports individuels fournissent une vision incomplète et peu fiable du monde, le modèle apprend des schémas en accumulant une représentation moyenne des exemples observés. Le modèle apprend également le taux de base d'apparition des différentes menaces dans l'environnement. Dans des contextes et situations spécifiques, l'officier aura une idée assez précise des menaces les plus probables, obtenue grâce à la coordination avec ses ressources de renseignement et d'autres officiers. En pratique, une grande partie du travail de l'officier consiste à apprendre la probabilité de ces différentes menaces, c'est-à-dire estimer leur probabilité d'apparition plutôt que de simplement rechercher des preuves d'attaques. Dans le modèle, cette connaissance est saisie par l'accumulation d'un compteur d'événements empiriques qui fournit une simple distribution de fréquence des événements vécus.

3. **Modélisation de la décision de reconnaissance** : Une fois qu'une représentation raisonnable d'un ensemble d'événements a été apprise, le modèle est capable de classer de nouveaux événements en fonction de son expérience passée. Il reçoit un rapport et classifie l'événement en identifiant lequel de ses schémas en mémoire est le plus susceptible d'avoir généré l'événement vécu. Pour ce faire, le modèle pondère les preuves dans le rapport avec la probabilité d'apparition d'une telle menace et avec les informations concernant la fiabilité ou non du rapporteur. Les caractéristiques de ce message sont comparées en parallèle avec les schémas pertinents de la mémoire à long terme, et pour chacun d'eux, une probabilité est calculée, qui intègre la fiabilité du capteur, les informations actuelles et les connaissances de base sur les menaces possibles.

Une expérimentation du BRDM [73] s'est basée en partie sur des entretiens avec des officiers de l'armée de terre et de l'armée de l'air américaine responsables de la défense chimique/biologique. Leur rôle principal était d'évaluer les rapports des capteurs et des humains concernant la présence de menaces et d'attaques chimiques ou biologiques. Comme l'occurrence de ce type d'événements réels est très rare, peu d'officiers peuvent témoigner d'un retour d'expérience, il s'agit donc d'une tâche d'ultra vigilance. Pour déterminer la validité d'une menace, le BRDM utilisera : les informations de base, les indices contenus dans les rapports et la fiabilité de la source des rapports.

Pour que le BRDM fonctionne, des hypothèses sont faites sur la façon dont les événements se produisent dans l'environnement, sont signalés au décideur et sont représentés par ce dernier. Tout d'abord, il a été supposé qu'un certain nombre d'événements classifiables pourraient se produire. Dans le domaine de la protection des armes chimiques/biologiques, ces classes d'événements correspondent à des choses telles que des attaques de missiles conventionnels, des fausses alertes dues à la poussière de l'environnement. Ces classes d'événements sont associées à des ensembles de caractéristiques typiques.

Le BRDM doit ensuite apprendre à connaître l'environnement dans lequel il opère. Dans des situations réelles, cette expérience provient de plusieurs sources : des manuels, des anecdotes et de l'expérience transmise par d'autres officiers, et des événements réels.

Une fois une représentation de l'ensemble des événements apprise, le modèle a pu classifier de nouveaux événements. Il recevait un rapport en entrée et classifiait l'événement en identifiant lequel de ses schémas en mémoire était le plus susceptible d'avoir généré l'événement vécu.

Ainsi le BRDM propose la mise en œuvre de la partie reconnaissance du RPD et, en particulier, introduit les notions de (a) la compréhension par le décideur du taux de base d'apparition des événements ; (b) le modèle mental du décideur d'un capteur (et sa fiabilité) ; (c) la notion d'évaluation d'une option sur ses propres mérites, plutôt que d'exiger la comparaison entre les alternatives ; et (d) l'utilisation de prototypes de connaissances pour prendre des décisions plutôt que l'utilisation d'un simple raisonnement basé sur des exemples ou des cas. Toutefois, les descriptions originales du modèle RPD vont bien au-delà du modèle BRDM. Le RPD suggère que la simulation mentale est importante pour déterminer si un plan d'action possible est tenable. La simulation mentale nécessite de disposer d'un modèle mental de la situation, afin que les résultats potentiels des actions puissent être évalués et acceptés ou rejetés. Le modèle BRDM a fait de petits pas vers cet objectif en incorporant des modèles mentaux explicites.

Maintenant que les SAD composés d'un ou plusieurs agents ont été recensés une seconde partie va introduire les SAD fondés sur des architectures cognitives.

2.1.2. Les SAD basés sur les architectures cognitives

Dans les sections suivantes, les SAD basés sur des architectures cognitives modélisant le RPD sont décrites ainsi que des exemples d'applications sur des missions militaires et/ou navales.

2.1.2.1. L'architecture R-CAST (Collaborative Agents for Simulating Teamwork)

L'architecture R-CAST est une architecture composée d'agents cognitifs, construite sur le concept de modèles mentaux partagés [74], d'échange proactif d'informations entre agents [75] et du modèle RPD. L'objectif d'un agent dans l'architecture R-CAST [76] est d'aider à résoudre ou résoudre un problème du monde réel. Le scénario doit clairement identifier les objets dans l'environnement de l'agent, les règles auxquelles l'agent doit obéir, un plan d'action et l'objectif global.

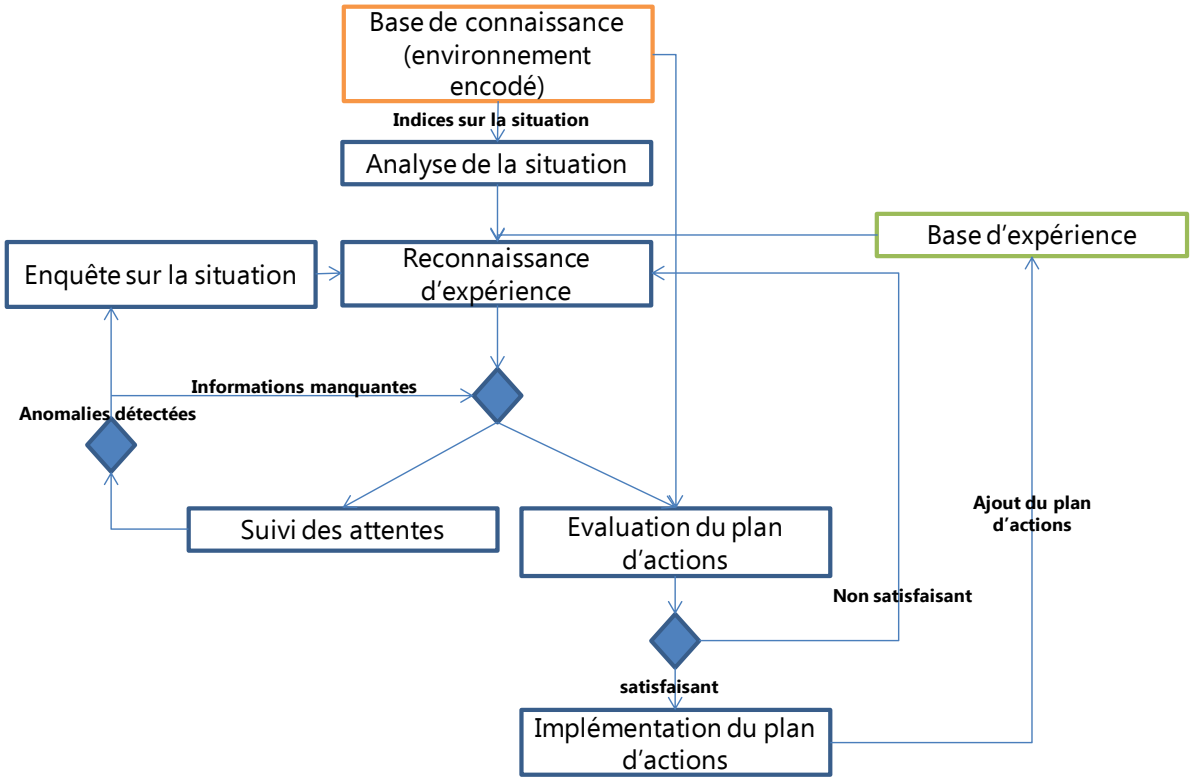

Figure 2-2 : Architecture R-CAST

En s'appuyant sur la Figure 2-2, l'agent débute sa phase de reconnaissance d'expérience dans l'analyse de la situation. À ce stade, l'agent décide s'il peut ou non reconnaître une expérience passée, ou s'il doit planifier une enquête sur la situation pour recueillir les informations manquantes. Si l'agent reconnaît une situation, ou si son enquête est terminée, l'étape de reconnaissance d'expérience est terminée et l'agent effectue une simulation mentale du plan d'actions prévu, un plan d'actions est écrit pour chaque expérience. Au cours de la simulation mentale, l'agent applique les règles qui le contraignent à son environnement à partir de la base de connaissances, ainsi que les attentes, les objectifs et les résultats de la dernière expérience et détermine si cette ligne de conduite particulière fonctionnera ou non. En supposant que le plan d'action est valide et accepté, l'agent le mettra en œuvre. Une fois que la ligne de conduite est terminée et que l'agent est satisfait, la nouvelle expérience sera ajoutée à la base d'expérience de l'agent. L'agent pourra désormais utiliser cette expérience pour une reconnaissance ou un soutien plus rapide pour la prochaine fois. Une interface utilisateur du RPD existe, mais n'offre pas la possibilité d'insérer manuellement des expériences ou des connaissances dans la base d'expériences, l'interface offre un moyen de visualiser l'état actuel du processus de reconnaissance de l'agent.

Une attente indique ce qui doit se passer, servant de condition de départ pour conserver la reconnaissance actuelle. Pour soutenir la prise de décision adaptative [77], un agent R-CAST surveille en permanence les attentes jusqu'à l'achèvement du plan d'actions sélectionné. L'invalidation de certaines attentes peut indiquer que la reconnaissance, autrefois valable, n'est plus applicable à la nouvelle situation. La partie déjà exécutée de l'action sélectionnée peut encore avoir un sens, mais le reste doit être ajusté. Dans ce cas, l'agent R-CAST peut entamer un nouveau cycle de reconnaissance.

La capacité de R-CAST en tant qu'aide à la décision a été évaluée dans une simulation de champ de bataille [78]. Les scénarios impliquaient une force bleue, amie, composée de trois zones fonctionnelles de combat décrites sur la Figure 2-3: la cellule de renseignement (S2), la cellule des opérations (S3) et la cellule logistique (S4). A la frontière du champ de bataille, une route d'approvisionnement reliait un aéroport A et une zone cible T. Les rôles simplifiés des cellules étaient les suivants : S2 contrôlait un drone pour collecter des informations et identifier si un objet en approche était une force neutre ou une unité ennemie ; S3 contrôlait un char pour détruire les ennemis et protéger la route de ravitaillement ; et S4 contrôlait un camion pour livrer du ravitaillement depuis A jusqu'à T, le camion sera détruit s'il se trouve dans le rayon d'attaque d'un ennemi en approche. L'objectif général de la force bleue était de protéger l'aéroport A et la zone cible T, et de s'assurer que le plus grand nombre possible de ravitaillements étaient livrés par S4 de A à T. Les cellules devaient collaborer entre elles afin d'avoir de bonnes performances.

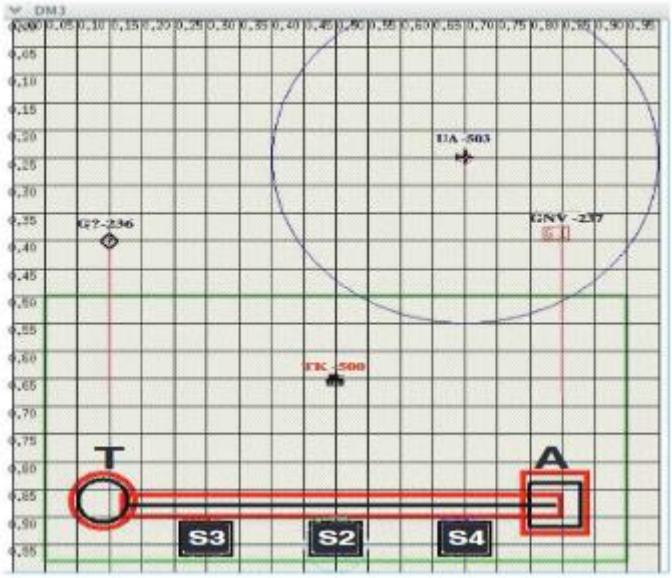

Figure 2-3 : Simulation du champ de bataille [57]

Une expérience [76] a été menée pour savoir si R-CAST pouvait améliorer les performances des équipes de commandement et de contrôle sous contrainte de temps. Deux types d'équipes ont été utilisés : des équipes humaines composées de trois sujets humains jouant les rôles de S2, S3 et S4 ; des équipes non-humaines composées de trois agents R-CAST jouant les rôles de S2, S3 et S4, l'agent S3 étant jumelé à un sujet humain. Après avoir comparé les équipes humaines et non-humaines sous différentes pressions de temps, les résultats ont confirmé que les agents R-CAST en tant qu'aide à la décision amélioraient les performances humaines.

2.1.2.2. L'architecture CogTRANS

Le modèle CogTRANS [79], [80], [81] cherche à imiter les comportements d'experts, ici des chefs de quart à bord de ferries et de cargos, via l'approche Agent-Groupe-Rôle, afin de simuler des évitements de collisions. L'expérience des individus est stockée à l'aide d'une base de patrons, associant à chaque situation, une décision. Le modèle de prise de décision en découlant et qui a été implémenté est le modèle de Décision à Base de Patrons (DBP). La Figure 2-4 présente le fonctionnement global de l'architecture CogTRANS et plus particulièrement du module cognitif de Décision à Base de Patrons. La partie bleue présente les modules externes de DBP : ceux-ci assurent l'interface entre l'environnement et le modèle cognitif. Les classes en orange représentent les modules internes de la prise de décision.

Le modèle DPB est décomposé en quatre phases. La première est la phase de perception de la situation qui est basée sur des sous-ensembles flous, employés pour transformer des données quantitatives en données qualitatives. S'ensuit la phase d'appariement de la situation à une ou plusieurs situations prototypiques. Le patron optimal, en fonction du profil de chaque agent, est retenu. La décision est ensuite traduite en une action. Le module DBP est validé par une extension du simulateur TRANS (Tractable Role Agent prototype for concurrent Navigation Systems) afin de reproduire le comportement d'experts maritimes.

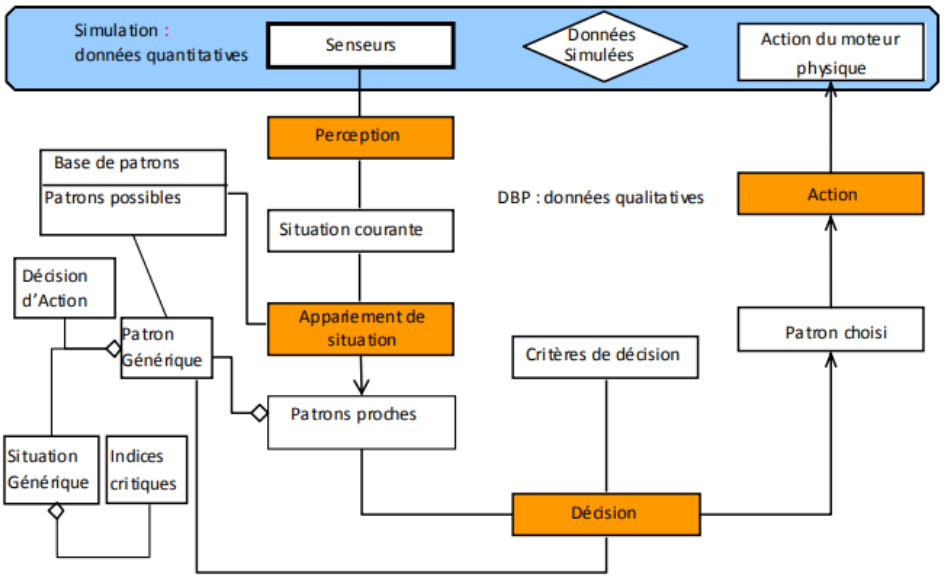

Figure 2-4 : Diagramme général du module DBP (Décision à Base de Patrons) [79]

Cependant certaines hypothèses ont été faites. Les indices critiques Figure 2-4 sont des éléments clés qui permettent à un individu donné ou à un groupe d'individus regroupés au sein d'un même groupe d'experts de prendre une décision en rapport avec un but donné, c'est-à-dire un rôle au sein du modèle. Dans le modèle proposé, chaque indice critique contribue à part égale à la reconnaissance d'une situation ce qui n'est pas le cas dans la réalité. Pour déterminer l'importance de l'un vis-à-vis de l'autre il aurait fallu passer par des méthodes d'apprentissage par renforcement ou par réseau neuronal, ce qui n'est pas été intégré dans ce modèle.

2.1.2.3. Les architectures cognitives de la Naval Posgraduate School (USA)

2.1.2.3.1. L'architecture cognitive de Kunde et Darken.

Kunde et Darken [82], [83], [84] ont développé une architecture cognitive adaptée au RPD en essayant de respecter au maximum les points clés de la simulation mentale (cf : partie 1.3.3.1). Pour simuler la simulation mentale, une approche probabiliste à l'aide des chaines de Markov [85] a été utilisée. Une chaîne de Markov peut fournir une probabilité de passage à l'état suivant et donc donner une estimation sur un événement futur. L'ensemble de leur architecture Figure 2-5 est composé de quatre éléments : l'environnement, qui couvre principalement le système de simulation, la composante de conscience de la situation, qui assure une image actualisée de la situation au moment de la décision, le simulateur mental, qui prédit et évalue, et la composante de décision, qui évalue les facteurs d'influence et prend la décision.

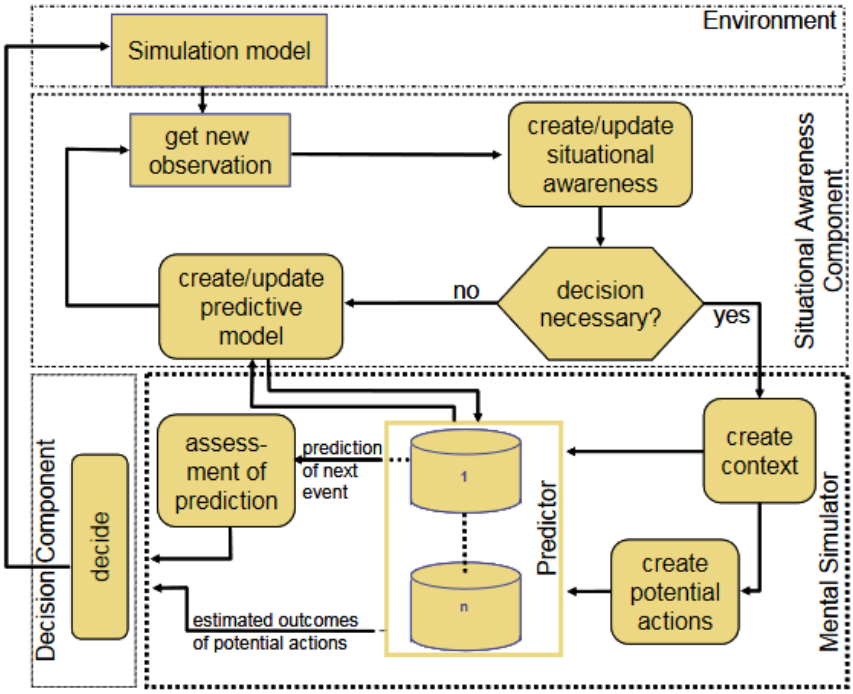

Figure 2-5 : Architecture générale [82]

Ils ont testé leur architecture sur un scénario militaire de combat terrestre entre deux forces. Les forces représentées étaient un peloton de chars bleus et un peloton de chars rouges. Les chars rouges suivent un chemin prédéterminé dans la zone de combat des bleus, et essaient d'encercler la position des bleus si ceux-ci se révèlent trop tôt. Seule la force bleue a utilisé l'architecture décisionnelle. Le comportement de la force rouge a été généré d'une manière relativement simple et conventionnelle.

Le simulateur mental, qui est la composante centrale de leur architecture, utilise les connaissances acquises dans le passé, prédit le prochain événement probable et place cet événement estimé dans le contexte de la situation prévue. Dans le contexte de leur scénario, trois indicateurs sont prédits, (1) l'estimation de la prochaine observation probable et le temps

moyen pendant lequel cet événement se produira, (2) la prévision du terrain dans un avenir proche et (3) la création des actions potentielles et l'estimation de leurs résultats.

Le simulateur mental est relié au module de décision qui a besoin de trois données d'entrée fournies par le simulateur mental :

- une prédiction du prochain événement susceptible de se produire. Ce sera le changement prévu de la richesse de la cible dans leur scénario.
- une évaluation de la prévision en tenant compte de l'influence prévue du terrain, et
- une évaluation des pertes probables des bleus et des rouges pour chaque action possible.

En d'autres termes, la composante décisionnelle prend le nombre prévu de chars visibles dans la prochaine observation, récupère un temps médian pour que cet événement se produise et estime l'emplacement prévu des chars visibles. La nouvelle estimation de l'emplacement est calculée géométriquement sur la base de la vitesse et de la direction estimées. Pour cette estimation de position, il existe un attribut de cellule de terrain qui indique la probabilité qu'une observation se produise à cet endroit. L'attribut de terrain sera utilisé pour évaluer la prédiction. Les pertes probables des bleus et des rouges pour chaque action possible sont extraites de la base de données.

Ces données d'entrée sont ensuite analysées par un arbre de décision. Celui-ci en fonction des données reçues devra choisir soit de « tirer » soit de « cesser le feu ». Dans leur scénario considéré, le char décide de tirer selon l'arbre de décision représenté Figure 2-6.

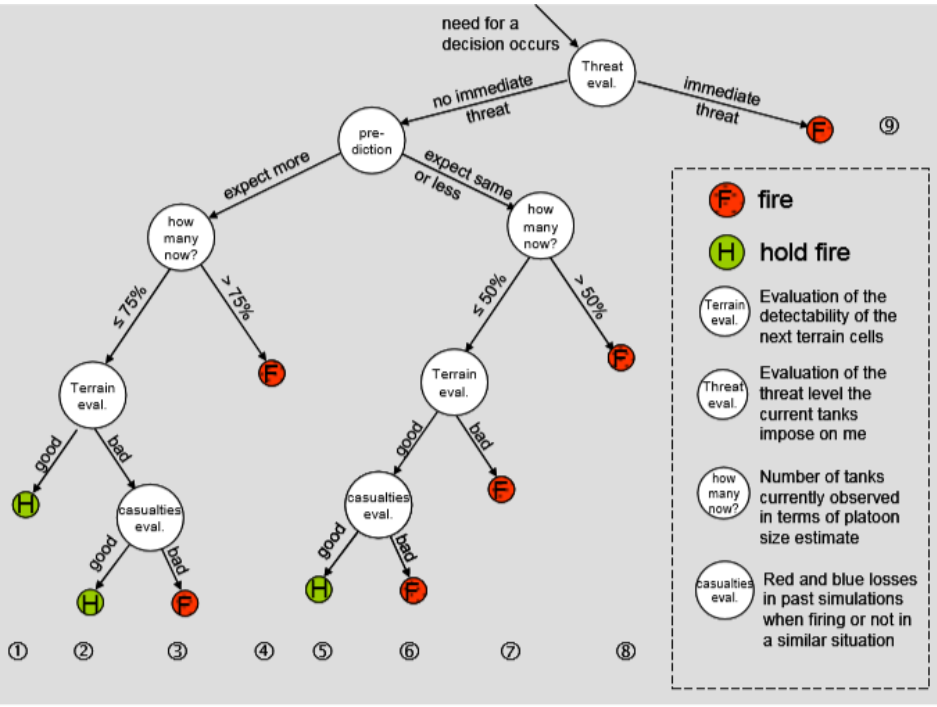

Figure 2-6 : Arbre de décision [82]

Une fois qu'un char adverse est en vue, cet arbre de décision est activé car une décision est nécessaire. La composante décisionnelle descend dans l'arbre jusqu'à atteindre une des

décisions possibles, « tirer » ou « cesser le feu ». A chaque nœud, une condition est vérifiée, permettant à la composante décisionnelle de choisir le chemin approprié. Le nœud supérieur évalue le niveau de menace des chars observés par rapport aux chars bleus amis, où le processus de décision du chef est modélisé. La détermination du niveau de la menace va inclure des facteurs tels que le cap du char ou si le canon ennemi pointe vers la position bleue. Le niveau de menace peut également être influencé par la mission, et pas seulement par le risque de se faire tirer dessus.

Plusieurs expériences ont été menées avec des officiers militaires de différents services. Le nombre de participants a varié entre 6 et 11. Le but des expériences était de comparer d'une part la prédiction du modèle par rapport à celle des humains et d'autre part de comparer le comportement de tir des humains et du modèle sur plusieurs scénarios. Les résultats de ces expériences ont montré des résultats encourageants et validé l'architecture. La mise en œuvre choisie montre que la simulation mentale peut être mise en œuvre avec succès dans un environnement de simulation de combat et permet de mieux imiter le comportement humain basé sur le RPD.

2.1.2.3.2. Le modèle de Géographie culturelle : une implémentation du RPD à l'aide de l'apprentissage par renforcement

Le modèle de géographie culturelle développé par le centre d'analyse du Commandement de la formation et de la doctrine de l'armée américaine (TRADOC) - Monterey (TRAC-MTRY) [86] implémente le RPD en utilisant l'apprentissage par renforcement (AR). Le but de ce modèle est de fournir un cadre pour étudier les effets des opérations dans la guerre asymétrique, c'est-à-dire dans un conflit qui oppose la force armée d'un État à des combattants matériellement insignifiants, en modélisant le comportement et les interactions des populations. L'apprentissage par renforcement [18] utilisé pour implémenter le RPD, se base sur un processus de décision markovien, un modèle stochastique, où un agent prend des décisions et les résultats de ses décisions ou actions sont aléatoires. En comparaison avec l'architecture cognitive de Kunde et Darken décrite en 2.1.2.3.1, les processus de décision markovien sont des extensions des chaines de Markov. La différence est l'addition des actions choisies par l'agent et ses récompenses gagnées. S'il n'y a qu'une seule action à tirer dans chaque état et que les récompenses sont égales, le processus de décision markovien est une chaîne de Markov. Le module de sélections des actions de leur architecture cognitive repose sur l'utilisation de deux concepts principaux décrits Figure 2-7 :

- La méthode d'apprentissage par exploration : Le modèle d'apprentissage par renforcement développé par Papadopoulos [87] s'appuie sur une méthode d'apprentissage par exploration, ici un algorithme de Q-learning [88] pour sélectionner les plans d'actions les plus appropriés.

- La mise en œuvre du RPD : L'application du modèle RPD est basée sur la technique d'apprentissage par renforcement décrite ci-dessus. Au début de chaque nouvelle situation jamais apprise, les agents utiliseront initialement la méthode d'apprentissage par exploration avec une stratégie ϵ-greedy [89]. La stratégie ϵ–greedy permet à un agent de réaliser un compromis exploration/exploitation de son environnement lors de

son apprentissage Au départ ϵ sera égale à 1, les agents auront 100% de chance d'explorer leur environnement, ils choisiront des actions de manière aléatoire. Au fur et à mesure de l'apprentissage la valeur de ϵ décroira et les agents exploiteront de plus en plus ce qu'ils auront appris Le nombre d'occurrences de chaque action est enregistré et comparé à un seuil minimum défini par l'utilisateur. Ce seuil atteint, l'agent est considéré expert de l'action réalisée, c'est-à-dire que l'agent a fini son entraînement et que le modèle RPD pour cet agent est finalisé. Pour la suite de la simulation il appliquera directement le modèle RPD lorsqu'il se retrouvera dans une situation déjà apprise.

Pour choisir la méthode appropriée, le modèle compte le nombre d'occurrences de chaque action possible effectuées dans le passé, le nombre le plus bas est alors considéré comme l'expérience de l'agent. La méthode RPD ou la méthode d'exploration est prévue, selon que l'expérience minimale a été atteinte ou non. Le paramètre de seuil d'expérience minimale, prédéfini par l'utilisateur, contrôle directement la quantité d'exploration que les entités sont autorisées à faire avant d'utiliser le RPD directement. Une fois la méthode de prise de décision déterminée, l'entité choisit l'action appropriée.

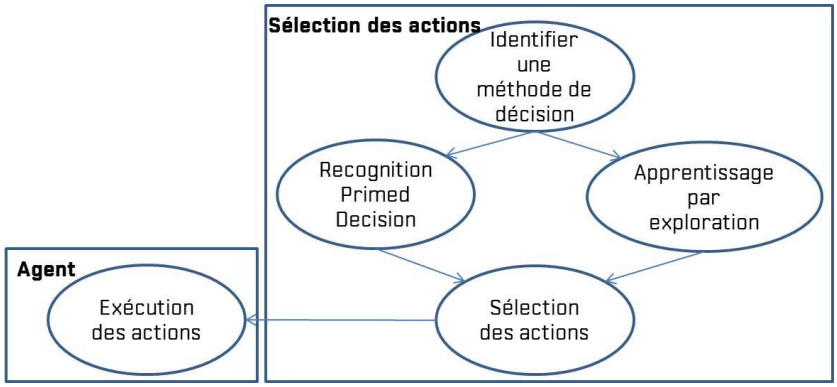

Figure 2-7 : Sélection des actions dans le modèle de géographie culturelle [86]

Cependant, une telle mise en œuvre présente des limites. Le modèle repose essentiellement sur une approche « gloutonne » (ϵ-greedy) [89] d'apprentissage par renforcement, où un agent a eu la possibilité d'explorer diverses options un grand nombre de fois dans l'environnement avant de prendre une décision. En revanche, pour une approche théorique du RPD, c'est-à-dire en appliquant l'arbre de décision proposé par Klein, les bénéfices de temps et de la connaissance de l'historique des actions peut ne pas être disponible pour un décideur humain. Un agent n'ayant pas fait de choix d'action préalable dans un scénario ou un environnement particulier devrait pouvoir, afin de jouer son rôle de manière satisfaisante et d'appliquer le RPD dans sa totalité, décider de sa ligne de conduite en fonction de l'ensemble limité des perceptions qu'il reçoit, en utilisant d'autres connaissances telles que son expérience antérieure, sa mémoire à long terme.

Plus généralement, un décideur possède la capacité de reconnaitre les changements d'une situation et d'écarter les mesures adoptées antérieurement si elles ne sont plus efficaces. La méthode mise en œuvre dans le modèle présenté plus haut ne permet pas aux agents d'avoir une telle polyvalence, limitant leur « expertise » à des situations relativement statiques.

2.1.3. Limites des systèmes informatiques

Les réseaux de neurones présents à la fois dans les systèmes composés de structures LTM et de réseaux Bayésiens sont entraînés pour produire des résultats spécifiques à des entrées spécifiques. Pour ces raisons ils sont utilisés pour la partie reconnaissance du RPD ne permettant pas une reproduction totale du RPD. Ils sont de type « boîte noire », sont difficilement explicables et ne sont pas capables d'agir lors de situations inédites jamais apprises en faisant preuve de parcimonie [90].

Les systèmes basés sur les agents BDI quant à eux utilisent des plans pour atteindre des objectifs en fonction de l'environnement actuel. Lorsqu'un agent BDI rencontre un problème qui l'empêche de réaliser le plan en cours, il arrête l'exécution de ce plan, réévalue la situation en fonction de l'environnement mis à jour et sélectionne un nouveau plan dans une bibliothèque de plans. Cela permet un certain niveau d'adaptation à un environnement dynamique. Cependant il ne dispose pas d'adaptation basée sur l'expérience passée. Une telle capacité peut être importante dans des environnements dynamiques changeant d'une manière non prévue [91]. De plus, ils ne font *a priori* pas de prévision ou de planification ; l'exécution est basée sur une bibliothèque de plans fournie par l'utilisateur pour atteindre les objectifs [92].

Ce système de plans utilisé par les agents BDI se retrouve dans l'architecture R-CAST sous forme de base de données et dans les CA avec l'expérience programmée. De même qu'énoncé plus haut, cela ne permet pas aux agents de s'adapter à des environnements dynamiques changeant de manière non prévue ou à des situations inédites.

De plus, les SAD cités ci-dessus n'intègrent pas ou que partiellement la simulation mentale du RPD, c'est-à-dire, la capacité d'un agent logiciel à simuler des événements futurs afin d'évaluer ses propres actions et d'émettre des hypothèses sur les événements qui pourraient se produire compte tenu de la situation actuelle et passée. La simulation mentale est une partie essentielle pour permettre à notre SAD de modéliser la prise de décision humaine rapide et de préconiser des décisions « adaptées » aux décideurs. Parmi les SAD présentés, seules ceux basés sur les deux architectures cognitives de la Naval Postgraduate School permettent de modéliser la simulation mentale à l'aide des chaînes de Markov et de l'apprentissage par renforcement. Les chaines de Markov et l'AR permettent de modéliser à différents niveaux la simulation mentale où le décideur évalue mentalement l'évolution des événements dans un futur proche avant de faire un choix d'actions. Cependant, contrairement à l'architecture de Kunde et Darken, le modèle de géographie culturelle n'implémente que très partiellement la « reconnaissance de schémas déjà vus » du RPD et n'explicite pas comment les autres étapes du RPD sont modélisées, exceptée la simulation mentale.

Certains avantages et inconvénients des SAD présentés sont résumés dans le Tableau 2-1 ci-dessous.

Tableau 2-1 : Résumé de certains avantages/inconvénients des SAD présentés

SAD	Avantages	Inconvénients
SAD à base de réseaux de neurones	Partie reconnaissance du RPD et sélection du plan d'actions modélisés	Nécessite beaucoup de données pour l'entraînement/ simulation mentale non modélisée
SAD à base de structures LTM	Partie reconnaissance du RPD modélisée	Implémentation partielle du RPD/ Ne fonctionne qu'en environnement statique
SAD à base d'agents composites	Modélisation quasi complète du RPD	Inadapté pour des situations inédites/ Modélisation partielle de la simulation mentale
SAD à base d'agents BDI	Modélisation précise de la connaissance de la situation et de la réalisation des objectifs	Pas d'adaptation basée sur l'expérience passée / simulation mentale non modélisée
SAD à base de réseaux Bayésiens de reconnaissance	Partie reconnaissance du RPD détaillée et aboutie	Inadapté pour des situations inédites/ Simulation mentale non modélisée
Architecture R-CAST	Modélisation quasi complète du RPD	Inadapté pour des situations inédites/ Modélisation partielle de la simulation mentale
Architecture CogTRANS	Modélisation de quatre phases, (1) perception, (2) appariement de la situation, (3) décision et (4) action.	Modélisation très partielle de la simulation mentale/ les indices pour la partie reconnaissance ont tous le même poids
Architecture cognitive de Kunde et Darken	Modélisation de la simulation mentale	Modélisation partielle de la partie reconnaissance du RPD
Architecture cognitive du modèle de géographie culturelle	Modélisation de la simulation mentale	Modélisation très partielle de la partie reconnaissance du RPD / Inadapté pour des situations dynamiques et inédites

2.1.4. Synthèse

Des SAD fondés sur le RPD ont été implémentés pour venir en aide aux décideurs et faire face à la surcharge cognitive telle que la surcharge d'informations, aux nombreuses prises de décision. Lors de missions navales militaires tactiques, l'EM est souvent confronté à des situations nouvelles et doit pouvoir avoir à sa disposition un SAD capable de fonctionner lors de situations changeantes et simulant le plus complètement le RPD, notamment la reconnaissance de schémas déjà vus et la simulation mentale, les deux phases essentielles du RPD, afin de s'adapter le plus possible aux décideurs. Le SAD répondant au mieux à ces exigences est actuellement celui conçu par Kunde et Darken à la Naval Postgraduate School [82], [83], [84] reposant sur une architecture cognitive adaptée à des opérations militaires tactiques.

Pourtant, les chaînes de Markov utilisées dans cette architecture ne permettent pas de modéliser la simulation mentale dans son intégralité supposant que l'estimation d'un événement futur ne dépend que de l'état précédent sans tenir compte de l'influence de l'action envisagée. Or lors de la réalisation de missions navales militaires, les opérationnels se « projettent dans le futur » en fonction de l'état dans lequel ils sont et de la ou les actions envisagées.

L'apprentissage par renforcement permet cette anticipation du futur en fonction de l'état et des actions. Cependant, comme le démontre le modèle de géographie culturelle, les algorithmes d'apprentissage par renforcement ne peuvent pas être utilisés pour des prises de décisions en environnement dynamique et inédit, les forçant à être utilisés pour des problèmes relativement simples et statiques. L'apprentissage par renforcement profond qui est la combinaison de

l'apprentissage par renforcement et des réseaux de neurones [93] permet à un agent de pouvoir disposer d'une expérience antérieure sur différentes missions et de s'adapter à un environnement dynamique. Le système est alors « dynamique » avec « effet mémoire ».

Afin de pallier les limites des SAD existants et de simuler au mieux le RPD, notamment la simulation mentale, la formalisation du système d'aide à la décision proposé va s'inspirer de l'architecture cognitive de la Naval Postgraduate School et de l'apprentissage par renforcement profond.

L'objectif de la partie suivante est de présenter plus en détails l'utilisation de l'apprentissage par renforcement (AR) pour modéliser la prise de décision humaine d'experts et donc le RPD et d'établir un bref état de l'art sur l'apprentissage par renforcement (RL - Reinforcement Learning en anglais) et l'apprentissage par renforcement profond (DRL – Deep Reinforcement Learning en anglais).

2.2. Apprentissage par renforcement pour le RPD

Pour rappel, le RPD est un modèle individuel de prise de décision rapide basé sur le fait que les décideurs expérimentés sont capables de générer rapidement des plans d'actions satisfaisants [25]. Les décideurs sont supposés capables de comparer les situations actuelles avec des situations antérieures qu'ils ont vécues, de reconnaître les similitudes et d'utiliser ces similitudes afin d'évaluer la situation par simulation mentale et de déterminer rapidement une solution admissible.

L'apprentissage par renforcement est une méthode issue de l'IA permettant à un agent informatique d'apprendre des conséquences de ses actions en interagissant avec un environnement dynamique. Astreint au principe de la sanction et de la récompense, le modèle apprend de chacune de ses actions. Après apprentissage il pourra être considéré dans une certaine mesure comme décideur expert et capable de répondre de façon autonome à des situations en s'appuyant sur son expérience passée. Il pourra préconiser en tant qu'aide au commandement des actions et des décisions adaptées, de sorte que l'objectif de la mission soit réalisé.

Les sections suivantes définissent le lien entre la prise de décision humaine et l'AR pour justifier son utilisation dans notre système d'aide à la décision pour la conduite de missions navales militaires. Dans les deux dernières parties, le principe théorique de l'AR et l'AR profond est introduit.

2.2.1. Prise de décision humaine et apprentissage par renforcement

L'apprentissage est l'une des principales applications de la simulation mentale. En situation, les humains apprennent en interagissant avec l'environnement et en observant les résultats. Après avoir acquis suffisamment d'expérience, le cerveau est capable de simuler cet environnement et de l'utiliser pour imaginer de nouveaux résultats en appliquant un comportement différent [94]. L'apprentissage par renforcement s'inspire de la façon dont

l'Homme apprend et permet de modéliser sur beaucoup de points la prise de décision humaine justifiant son utilisation pour modéliser le RPD et notamment la simulation mentale dans le système d'aide à la décision [95].

L'apprentissage par renforcement provient à l'origine des théories empiriques sur l'apprentissage animal [96] issu du béhaviorisme. Le béhaviorisme est une approche de la psychologie étudiant l'interaction de l'individu avec son milieu. En psychologie animale plusieurs modèles d'apprentissage existent, dont le conditionnement opérant de Buhrrus F. Skinner [97] qui a introduit pour la première fois la notion de renforcement. Ces modèles explicitent comment l'association entre des stimuli et des réflexes ou conséquences peut être renforcée.

Au début du XX^{ème} siècle, le physiologiste Ivan Pavlov [98] a mis en place le conditionnement pavlovien, où le renforcement est décrit comme « le renforcement » ou l'affaiblissement d'un comportement lorsque celui-ci est précédé d'un stimulus spécifique. Les travaux de Pavlov ont amené Edward L. Thorndike [99] à formuler la loi de l'effet établissant qu'une réponse est plus susceptible d'être reproduite si elle entraine une satisfaction pour l'organisme et d'être abandonnée s'il en résulte une insatisfaction. Cette méthode a inspiré les algorithmes d'apprentissage par essais et erreurs.

Le psychologue Buhrrus F. Skinner [100] approfondit les travaux de Thorndike et élabora au milieu du XX^{ème} siècle le concept de conditionnement opérant se distinguant du conditionnement pavlovien par le fait que la réponse n'est pas nécessairement un réflexe de l'organisme. Skinner observa que la pression exercée par un animal sur un levier s'accroît de manière considérable lorsqu'elle entraîne la distribution de nourriture. Les conséquences d'une action selon que ce soit des récompenses ou des punitions rendent plus ou moins probable la reproduction du comportement.

De ces études et de l'essor de l'IA, notamment menées par Alan Turing en 1950 [64], émane l'idée de programmer un ordinateur pour apprendre une tâche par essais et erreurs. Les systèmes conçus pour optimiser des actions dans des environnements complexes sont confrontés aux mêmes défis que les animaux, à l'exception que l'équivalent des récompenses et des punitions est déterminé par des fonctions récompenses lors de la conception. En 1968, R. Chambers et Donald Michie [101] écrivent un programme capable d'apprendre à maintenir un pendule inversé en équilibre. Puis Andrew G. Barto et *al.* [102] proposent une méthode AHC-learning (Adaptative Heuristic Critic), considérée comme la première méthode d'apprentissage par renforcement. Dans ces architectures, la structure de mémoire est séparée pour représenter la fonction qui sélectionne les actions (l'acteur) et celle qui les évalue (le critique).

La formalisation mathématique des algorithmes d'apprentissage par renforcement date des années 80 lorsque Richard Sutton [103] et Christopher Watkins [88] publient les algorithmes du TD(λ) et du Q-learning. Ces algorithmes ont été inspirés par des données comportementales sur la façon dont les animaux apprennent et modélisent la prise de décision humaine [19].

De nombreux aspects de l'apprentissage par renforcement ont également été étudiés directement pour expliquer certains phénomènes dans le cerveau. Par exemple, les modèles informatiques ont été une source d'inspiration pour expliquer des phénomènes cognitifs tels que l'exploration [104] et l'actualisation temporelle des récompenses [105]. En science cognitive,

Kahneman [106] a également décrit une dichotomie entre deux modes de pensée : un « Système 1 » qui satisfait et qui ne satisfait pas et un « Système 2 » qui est plus lent et plus logique. En apprentissage par renforcement, une dichotomie similaire peut être observée.

Pour conclure la modélisation du comportement humain dans des tâches dynamiques est encore un défi pour les sciences cognitives. Cependant, l'AR a été largement utilisé dans les domaines de la modélisation cognitive, du jugement et de la prise de décision humaine [107]. Dans la suite, il est montré que l'AR profond a une structure décisionnelle générique qui est bien adaptée à l'explication du comportement humain dans un environnement dynamique. Pour ces raisons, il a été choisi de modéliser le RPD à l'aide de l'AR profond. Dans la suite le fonctionnement mathématique de l'AR et de l'AR profond est explicité.

2.2.2. Apprentissage par renforcement

L'apprentissage par renforcement décrit schématiquement Figure 2-8 est une technique d'apprentissage automatique qui détermine comment un agent doit prendre des décisions dans un environnement afin d'optimiser une notion de récompenses cumulatives. Dans un environnement donné, un agent reçoit des informations qui permettent de mettre à jour de façon dynamique son état en sélectionnant l'action la plus vraisemblable parmi l'ensemble fini des actions possibles [108]. La transition qui permet de définir un nouvel état est sous la forme d'une récompense immédiate ou d'une pénalité accordée à l'agent.

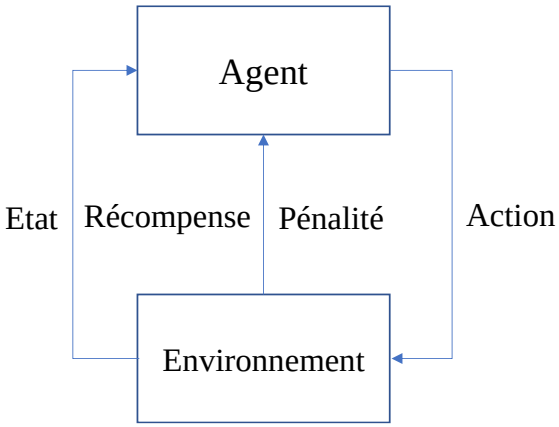

Figure 2-8 : Schéma de l'apprentissage par renforcement

2.2.2.1. Processus de décision markovien

La notion de processus de décision markovien (PDM) [18] explique une grande partie des fondements de l'apprentissage par renforcement. Un PDM est défini comme un 4-uplet défini par $\{S, A, R, T\}$, où

- S est un ensemble fini d'états,
- A un ensemble fini d'actions,
- $R : S \times A \rightarrow \mathbb{R}$ une fonction récompense,

- $T: S \times A \times S \rightarrow [0,1]$ une fonction transition.

La fonction récompense renvoie un scalaire, une récompense, pour une action sélectionnée dans un état donné. La fonction de transition spécifie la probabilité $Pr(s'|s,a)$ de passage de l'état s à l'état s' lors de la prise d'une action a notée $T(s,a,s')$. Un PDM fini est un processus dans lequel un ensemble d'états S, d'actions A, et de récompenses R ont un nombre fini d'éléments.

Une entité qui apprend et prend des décisions est appelée « agent » et tout ce qui est extérieur à l'agent est appelé « environnement ». L'agent apprend de l'interaction continue avec un environnement pour atteindre un objectif, réaliser une tâche. L'agent sélectionne des actions, puis l'environnement répond à ces actions en donnant une récompense permettant de mettre à jour l'état courant. L'objectif de l'agent est de maximiser le montant total de la récompense qu'il reçoit à long terme et non ses récompenses immédiates.

L'interaction entre l'agent et l'environnement Figure 2-9 est généralement décrite par une séquence à temps discrets. Au temps t, l'agent reçoit un état s_t et sélectionne une action a_t sur la base de l'état actuel. Au temps $t+1$, l'agent reçoit une récompense r_{t+1} et un nouvel état s_{t+1} à la suite d'une action a_t.

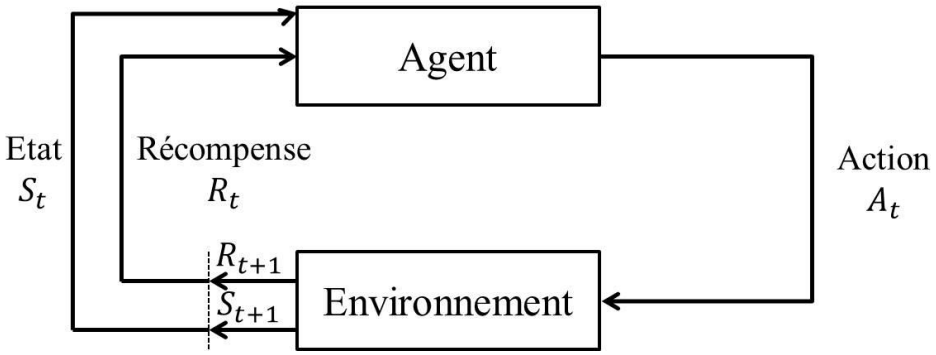

Figure 2-9: Interaction entre agent et environnement

Dans le cadre du PDM, l'état est supposé posséder la propriété de Markov : **la réponse de l'environnement au temps $t+1$ dépend uniquement des représentations de l'état et de l'action au temps t**. Cette propriété de Markov permet de prédire l'état suivant à partir de la transition $T(s'_t, a_t, s_t)$ et de récompenser l'état et l'action actuelle.

2.2.2.2. But et récompense

A chaque pas de temps [18], le but de l'agent est de maximiser le montant total des récompenses qu'il reçoit. Cela signifie ne pas maximiser la récompense immédiate mais la récompense cumulative sur le long terme. Selon ce principe, l'agent essaie de sélectionner l'action a_t pour maximiser la récompense cumulative suivante :

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \quad (1)$$

Où γ est un paramètre, $0 \leq \gamma \leq 1$, appelé facteur d'actualisation. Le facteur d'actualisation γ détermine la valeur actuelle des récompenses futures. Si $\gamma = 0$, l'agent n'est concerné que par la récompense immédiate. L'action de l'agent influence uniquement la récompense actuelle. Si γ se rapproche de 1, l'agent considère plus fortement les récompenses futures de ses actions.

L'agent doit ainsi suivre les états ayant les valeurs les plus élevées et non les récompenses immédiates les plus élevées, car ces états apporteront une récompense plus élevée à long terme.

2.2.2.3. Episodes et apprentissage

La tâche que l'agent tente de résoudre peut avoir ou non une fin naturelle. Les tâches qui ont une fin naturelle, comme un jeu, sont appelées tâches épisodiques. Les tâches qui n'en ont pas, comme l'apprentissage de la marche avant, sont appelées tâches continues. La séquence de pas de temps du début à la fin d'une tâche épisodique est appelée un épisode. Les agents peuvent avoir besoin de plusieurs étapes temporelles et épisodes pour apprendre à résoudre une tâche. La somme des récompenses collectées dans un seul épisode s'appelle la récompense cumulative totale et définie en 2.2.2.2. Les agents sont conçus pour maximiser cette récompense cumulative totale et une limite de temps est souvent ajoutée aux tâches continues, afin qu'elles deviennent des tâches épisodiques.

Chaque quadruplet d'expérience (s'_t, a_t, s_t, R_t) offre une opportunité d'apprentissage et d'amélioration des performances de l'agent. Celui-ci peut être conçu pour apprendre des correspondances entre les observations et les actions, appelées politiques. L'agent peut être conçu pour apprendre des correspondances à partir d'observations (et éventuellement d'actions) vers des estimations de récompense appelées fonctions de valeur ou fonctions de paires état-action.

2.2.2.4. Politique et fonctions valeur

La plupart des algorithmes d'apprentissage par renforcement impliquent l'estimation de fonctions valeur ou de fonctions paires état-action qui estiment la qualité de l'agent dans un état donné ou la qualité de l'exécution d'une action donnée dans un état donné. La notion de « qualité » est ici définie en termes de récompenses futures qui peuvent être attendues. Les récompenses que l'agent peut s'attendre à recevoir à l'avenir dépendent des actions qu'il prendra. En conséquence, les fonctions de valeur sont définies par rapport à des manières d'agir particulières, appelées politiques.

Une politique $\pi : S \times A \rightarrow [0,1]$ met en correspondance les états avec les probabilités de sélection d'une action. La politique π représente la sélection d'actions de l'agent dans un certain état s. Dans tout PDM, il existe une politique qui est meilleure ou égale à toutes les autres politiques pour tous les états [18]. L'objectif de l'agent est de trouver cette politique dite optimale et notée π^*, qui maximise la récompense totale à long terme à l'aide de l'estimation de fonctions valeurs ou de paires état-action.

La valeur d'un état s sous une politique π, notée $v_\pi(s)$ est définie comme la récompense future moyenne attendues en partant de l'état s et en suivant la politique π [109]:

$$v_\pi(s) = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | \, s_t = s\right] \tag{2}$$

De manière similaire la fonction paire état-action Q, notée $Q_\pi(s,a)$, est une mesure de la récompense globale attendue en supposant que l'agent est dans un état s et effectue l'action a, en suivant une certaine politique π. Elle peut être calculée de façon récursive [109]:

$$\begin{aligned} Q_\pi(s,a) &= \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | \, s_t = s\,, a_t = a\right] \\ &= \sum_{s' \in S} \Pr(s'|s,a)\left(R(s,a) + \gamma \max_{a' \in A} Q(s',a')\right) \end{aligned} \tag{3}$$

Ainsi la politique optimale suit les conditions suivantes :

$$\begin{aligned} \pi^* &= \operatorname*{argmax}_{\pi \in \mathcal{P}} v_\pi(s) \\ &= \operatorname*{argmax}_{\pi \in \mathcal{P}} Q_\pi(s,a) \end{aligned} \tag{4}$$

Où $\mathcal{P}$ est l'ensemble fini des politiques admissibles.

2.2.2.5. Exploration

Pour maximiser la récompense totale, l'agent doit sélectionner l'action ayant la plus grande valeur (exploitation), mais pour découvrir une telle action, il doit essayer des actions non sélectionnées auparavant (exploration). Le compromis entre l'exploitation et l'exploration est l'un des défis de l'apprentissage par renforcement. Deux méthodes d'exploration sont présentées.

(1) **ϵ-greedy**. La stratégie ϵ-greedy [89] est l'une des méthodes les plus utilisées. Afin d'équilibrer l'exploration et l'exploitation, un taux d'exploration ϵ est défini et initialisé à 1. Ce taux d'exploration est la probabilité que l'agent explore l'environnement plutôt qu'il l'exploite, au départ l'agent a donc 100% de chance d'explorer son environnement. Au fur et à mesure de son apprentissage la valeur de ϵ va décroître afin que la probabilité d'exploration devienne de moins en moins probable à mesure que l'agent apprend à mieux connaître son environnement.

(2) **Softmax**. Une alternative est la méthode softmax, ou encore de Boltzmann [110]. Elle permet de pondérer les actions en fonction de leur valeur estimée. Les bonnes actions ont une probabilité exponentielle d'être sélectionnées. Une action a est choisie avec une probabilité :

$$p(s,a) = \frac{e^{\frac{Q(s,a)}{\tau}}}{\sum_{a' \in A} e^{\frac{Q(s,a')}{\tau}}} \tag{5}$$

Où $\tau > 0$ est un hyperparamètre appelé « température », utilisée pour contrôler le degré d'exploration. Lorsque $\tau \to \infty$, les actions sont choisies au hasard. Lorsque τ s'approche de 0, les actions sont sélectionnées de manière gourmande. Une telle approche à base de schémas de décroissance de température est également utilisée dans la mise en œuvre de l'algorithme du recuit simulé [111].

Le choix entre les deux méthodes dépend de ce que doit accomplir l'agent et des fonctions de récompenses. En pratique, le paramètre ϵ est facile à définir, alors que pour définir τ les valeurs probables des actions doivent être connues.

2.2.2.6. Apprentissage par renforcement sans modèle et basé sur un modèle

En apprentissage par renforcement, deux approches principales existent pour trouver la politique optimale en estimant les fonctions valeur ou état-action : celles basées sur un modèle d'environnement et celles qui ne tiennent pas compte de ce modèle. Un modèle simule la dynamique de l'environnement et permet de déduire comment l'environnement se comportera. En pratique, le modèle est utilisé par la fonction de transition et par la fonction de récompense dans un PDM.

2.2.2.6.1. Méthodes fondées sur un modèle d'environnement

L'apprentissage par renforcement basé sur un modèle [112] fait référence à l'apprentissage indirect d'une politique optimale par l'apprentissage d'un modèle de l'environnement en prenant des actions et en observant les résultats incluant l'état suivant et la récompense immédiate.

Les méthodes basées sur des modèles sont plus efficaces que les méthodes sans modèle, mais une exploration exhaustive est souvent nécessaire pour apprendre un modèle satisfaisant de l'environnement [53]. Si l'agent effectue une action a dans un état s, la valeur de l'action $Q(s,a)$ est mise à jour par l'équation de Bellman [113]:

$$Q(s,a) = R(s,a) + \gamma \sum_{s' \in S} \Pr(s'|s,a) \max_{a' \in A} Q(s',a') \tag{6}$$

Où $0 \leq \gamma < 1$, la fonction optimale Q^* peut être obtenue par itération sur l'équation de Bellman jusqu'à ce qu'elle converge [94].

Les modèles de transition et de récompense sont généralement appris par un modèle de maximum de vraisemblance. Supposons que $C(s,a)$ soit le nombre de fois où l'action a est prise dans l'état s et que $C(s,a,s')$ soit le nombre de fois que l'état s passe à l'état s' par l'action a. La probabilité de transition de la paire (s,a) vers l'état s' est obtenue par :

$$\Pr(s'|s,a) = T(s,a,s') = \frac{C(s,a,s')}{C(s,a)} \tag{7}$$

Le modèle de récompense pour une paire (s, a) est obtenu de manière similaire :

$$R(s,a) = \frac{RSUM(s,a)}{C(s,a)} \quad (8)$$

Où $RSUM(s,a)$ est la somme des récompenses que l'agent reçoit lorsqu'il prend une action a dans un état s.

2.2.2.6.2. Méthode sans modèle

Par définition, les méthodes sans modèle ne reposent pas sur des modèles de transition et de récompense. La plupart des approches sans modèle tentent soit d'apprendre une fonction valeur et d'en déduire une politique optimale (méthodes basées sur la fonction valeur), soit de rechercher directement une politique optimale (méthodes de recherche de politique). Les méthodes sans modèles peuvent être également classées comme étant soit « on-policy » soit « off-policy ». Dans les méthodes « on-policy » la politique actuelle génère des actions et est directement mise à jour, alors que dans les méthodes « off-policy », une politique exploratoire génère des actions et est différente de la politique qui est en cours d'actualisation. Le Tableau 2-2 ci-dessous cite les méthodes sans modèle les plus connues.

Tableau 2-2 : Recensement des méthodes sans modèle

Méthodes basées sur la fonction valeur	Méthode de Monte Carlo [18]	« off-policy »
	SARSA [18]	« on-policy »
	Q-learning [114]	« off-policy »
Méthodes de recherche de politique	REINFORCE [115]	« on-policy »
	Acteur-critique [18]	« on-policy »
	Off-policy policy gradient [18]	« off-policy »

Cette première partie a expliqué le fonctionnement mathématique de l'AR et évoqué succinctement les méthodes de résolution existantes. Dans la suite quelques limites de l'AR sont évoquées pour justifier l'emploi de l'AR profond.

2.2.3. Quelques limites de l'apprentissage par renforcement

Même si en théorie l'apprentissage par renforcement modélise le fonctionnement du processus de décision humain, les algorithmes existants restent à ce jour encore limités et perfectibles.

Les approches d'apprentissage par renforcement sont confrontées à la taille et à la complexité de l'espace d'états et de l'environnement, appelé la « malédiction de la

dimensionnalité » [18], [113]. En AR, ce travers s'exprime par le fait que plus le nombre de variables paramétrant un environnement augmente, plus le nombre d'états possibles relatifs à cet environnement croît de manière exponentielle. Cela signifie que même de simples problèmes d'apprentissage par renforcement deviennent très rapidement insolubles sur le plan du calcul, et les problèmes qui possèdent une complexité du monde réel sont insolubles en termes pratiques.

Une autre limite concerne la stationnarité d'un environnement dans lequel les agents doivent agir. Une hypothèse de base de l'apprentissage par renforcement est que l'état réel d'un environnement est stable par rapport à un agent, c'est-à-dire que la dynamique de l'environnement ne change pas au cours du temps. Cependant, lorsque plusieurs agents travaillent dans le même environnement, la structure de cet environnement peut changer de manière inattendue d'un état à l'autre. Prenons le cas d'un agent A qui se dirige vers un objectif et qui est soudainement bloqué par l'agent B qui atteint cet objectif. Du point de vue de l'agent A, l'objectif n'existe plus car l'environnement est dynamique, changeant en fonction de l'activité qui s'y déroule. La conséquence de la non-stationnarité est que les agents doivent continuellement apprendre de nouvelles politiques pour faire face au nouveau contexte environnemental, ceci pour chaque phase transitoire rencontrée, de sorte que le comportement ne peut jamais s'installer ou être optimisé.

Ces limites, listées de façon non-exhaustive, soulignent la nécessité de modifier les processus d'apprentissage afin de permettre aux agents de réagir dans des environnements dynamiques, complexes et incertains. L'émergence de l'apprentissage par renforcement profond va permettre de répondre à ces exigences.

2.2.4. Apprentissage par renforcement profond

L'apprentissage par renforcement profond combine les réseaux de neurones [93] avec une architecture d'AR explicité sur la Figure 2-10. Les réseaux de neurones sont utilisés pour approximer directement la fonction $v_\pi(s)$ ou $q_\pi(s,a)$ et la fonction politique π. Ils peuvent apprendre à faire correspondre des états à des valeurs, ou des paires état-action à des valeurs Q, sans utiliser une table-Q pour stocker, indexer et mettre à jour tous les états possibles et leurs valeurs. Un réseau neuronal est construit à partir des échantillons de l'état ou de l'espace d'action pour apprendre à prédire leur valeur par rapport au but dans une architecture d'apprentissage par renforcement. Comme tous les réseaux neuronaux, ils utilisent des coefficients pour approximer la fonction reliant les entrées aux sorties, et leur apprentissage consiste à trouver les bons coefficients, ou poids, en ajustant itérativement ces poids, grâce à la minimisation d'une fonction de perte à l'aide d'un algorithme itératif de descente de gradient [116].

L'utilisation des réseaux de neurones en apprentissage par renforcement profond permet donc d'augmenter la complexité des espaces d'états en s'affranchissant des tableaux de valeurs ou de fonctions et de tenir compte de la dynamique de la situation. Après apprentissage, l'agent est capable de généraliser la valeur d'états qu'il n'a jamais vue auparavant en utilisant les valeurs d'états similaires auxquelles il aura déjà été confrontées. L'agent pourra donc agir dans des environnements dynamiques et inédits.

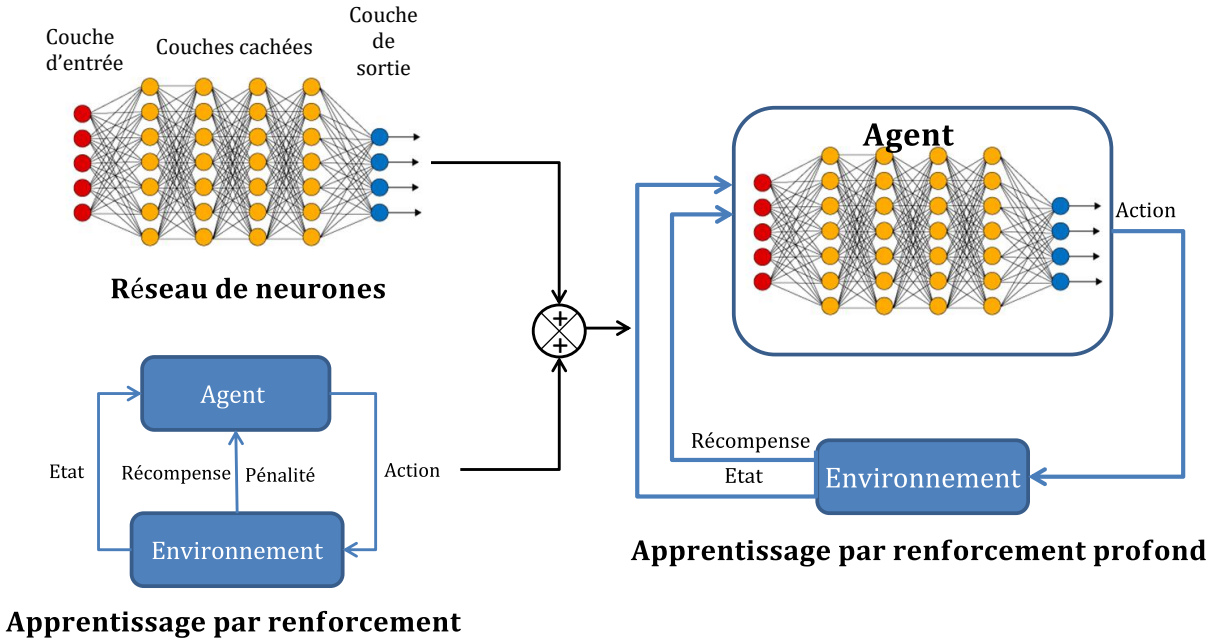

Figure 2-10 : Apprentissage par renforcement profond

La mise à jour de la valeur Q pour un état et une action donnés se fait pour rappel avec l'équation de Bellman [113] en utilisant une table Q :

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) \; + \alpha \left[R(s,a) + \gamma \max_{a' \in A} Q(s',a') \right] \tag{9}$$

En utilisant un réseau de neurones composé de poids θ à la place d'une table Q, l'équation de Bellman devient :

$$Q(s,a,\theta) \leftarrow (1-\alpha)Q(s,a,\theta) \; + \alpha \left[R(s,a) + \gamma \max_{a' \in A} Q(s',a',\theta) \right] \tag{10}$$

Au début de l'apprentissage, les poids du réseau de neurones sont initialisés de manière pseudo-aléatoire. Grâce à la réponse de l'environnement suite aux paires (s,a) le réseau neuronal doit minimiser la différence entre $Q(s,a,\theta)$ et $Q(s',a',\theta)$ à l'aide de l'erreur quadratique moyenne suivante :

$$L(\theta) = \frac{1}{N} \sum_{i \in N} \left(Q_\theta(s_i,a_i) - \; Q'_\theta(s_i,a_i) \right)^2 \tag{11}$$

Avec $Q'_\theta(s_i,a_i) = R(s_t,a_t) + \gamma \max_{a_i' \in A} Q_\theta(s'_i,a'_i)$

L'erreur quadratique moyenne $L(\theta)$ est minimisée par descente de gradient pour ajuster les poids du réseau de neurones et améliorer son interprétation des paires (s,a) :

$$\theta \; \leftarrow \theta - \alpha \frac{\partial L}{\partial \theta} \tag{12}$$

De manière identique à l'apprentissage par renforcement, des méthodes basées sur des modèles et des méthodes sans modèle existent en apprentissage par renforcement profond pour entraîner un agent. Les méthodes sans modèle sont actuellement les plus utilisées pour résoudre

des problèmes du monde réel. Le Tableau 2-3 montre un recensement non exhaustif de certaines méthodes les plus connues et utilisées :

Tableau 2-3 : Recensement de quelques méthodes sans modèle d'apprentissage par renforcement profond

Méthodes basées sur la fonction valeur	DQN[117]	« Off-policy »
	Double DQN[118]	« Off-policy »
	HER[119]	« Off-policy »
Méthodes de recherche de politique	A2C/A3C[120]	« On-policy »
	PPO[121]	« On-policy »
	ACKTR[122]	« On-policy »

Toutes ces méthodes possèdent des avantages et des inconvénients. Le choix d'une de ces méthodes pour un problème donné va dépendre du domaine d'application, de la complexité de celui-ci, de l'espace des états et actions de l'agent, et d'autres critères exogènes.

2.3. Conclusion

Pour venir en aide aux opérationnels lors de la conduite de missions navales tactiques des systèmes d'aide à la décision fondés sur le RPD ont été implémentés. Les premières parties de ce chapitre ont recensé les différents systèmes existants, ceux basés sur un ou plusieurs agents et ceux basés sur les architectures cognitives, ainsi que leurs limites. Parmi les limites évoquées, beaucoup des systèmes ne s'adaptent pas à des environnements dynamiques changeant de manière imprévue et/ou n'implémentent que partiellement la simulation mentale. Or, lors de la conduite de missions, l'organe de commandement est très souvent confronté à des situations nouvelles, changeantes de manière inattendue, dans lesquelles il a besoin d'un soutien. De plus, l'implémentation de la simulation mentale est très importante pour que le système puisse préconiser des décisions qui soient adaptées aux raisonnements des décideurs et rapidement compréhensibles afin de ne pas solliciter de ressources cognitives supplémentaires. Notre aide à la décision doit donc surmonter ces deux limites pour assister l'organe de commandement le mieux possible. Afin de proposer un système capable de s'adapter à un environnement dynamique, à des situations inédites et d'implémenter de manière plus complète la simulation mentale, la formalisation d'un système d'aide à la décision fondé sur le RPD et implémenté à l'aide de l'apprentissage par renforcement profond est proposée. Ce choix a été fait d'une part car l'apprentissage par renforcement profond permet de modéliser en grande partie le processus de décision humain et notamment la simulation mentale du RPD ; et d'autre part, cette méthode issue de l'IA permet à un agent informatique d'apprendre des conséquences de ses actions en interagissant avec un environnement dynamique. Après apprentissage, il sera considéré comme un décideur « expert », capable de répondre de façon autonome à des situations inédites et de préconiser en tant qu'aide au commandement des actions et des décisions adaptées en environnement dynamique, de sorte que le navire atteigne l'objectif de sa mission.

La formalisation du système proposé d'aide à la décision va s'inspirer de la structure de l'architecture cognitive de Kunde et Darken [82], [83], [84] décrite précédemment et s'appuiera sur l'apprentissage par renforcement profond, afin d'éviter une aide à la décision « statique » et « sans mémoire »,

Le prochain chapitre introduit la modélisation de l'architecture cognitive proposée pour notre système d'aide à la décision.

Chapitre 3 : Modélisation de l'architecture cognitive

Ce chapitre présente la modélisation de l'architecture cognitive de notre système d'aide à la décision fondée sur la méthodologie RPD. Comme évoqué dans le chapitre précédent, l'architecture cognitive proposée va s'inspirer de celle de Kunde et Darken [82], [83], [84] et s'appuiera sur l'apprentissage par renforcement profond afin de proposer une aide à la décision « dynamique » et avec « effet mémoire ».

3.1. Introduction

Cette section présente le système d'aide à la décision contenant notre architecture cognitive, puis décrit sa composition afin de modéliser le RPD et pour finir le profil des utilisateurs amenés à se servir du SAD proposé.

3.1.1. Introduction du système d'aide à la décision proposé

Le système d'aide à la décision (SAD) proposé est constitué de trois composantes principales, illustrées Figure 3-1. Une base de données permettant de recueillir et de stocker les données utiles au fonctionnement du système, un modèle regroupant l'architecture cognitive permettant au système de préconiser une décision en fonction des données reçues et une IHM permettant de restituer les décisions préconisées par le système et à l'utilisateur d'interagir avec le système.

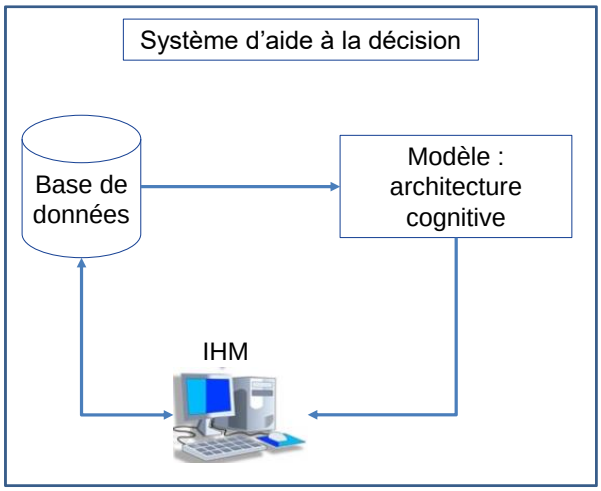

Figure 3-1 : Composition du SAD proposé

Le SAD proposé doit reproduire le schéma RPD utilisé par la grande majorité des décideurs lorsqu'ils sont confrontés à des prises de décision rapides en situation incertaine. Comme expliqué dans les chapitres 1 et 2, le processus RPD se base sur deux étapes clés : la reconnaissance de schémas déjà vus et la simulation mentale.

Afin de modéliser les étapes du RPD, le SAD sera composé d'une architecture cognitive « modulaire », représentée Figure 3-2. Chaque module aura pour fonction d'accomplir une ou plusieurs étapes du RPD et fonctionnera de manière similaire : les données d'entrée seront traitées par une ou plusieurs méthodes, et une ou plusieurs sorties sera fournies.

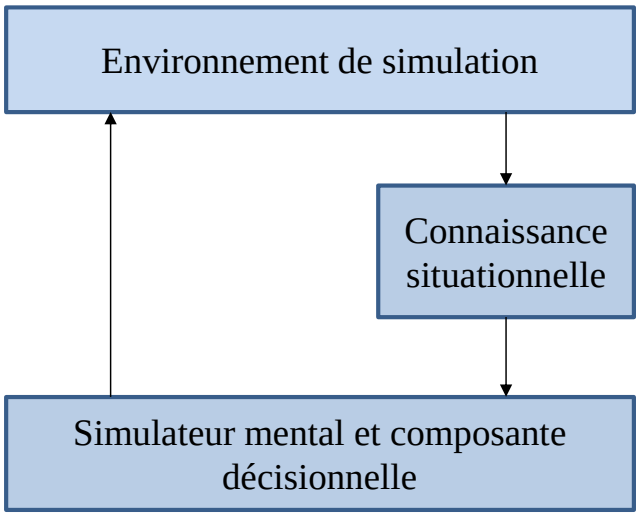

Figure 3-2 : Modules composant l'architecture cognitive

L'architecture cognitive proposée s'inspirant de celle de Kunde et Darken [82], [83], [84] est composée de trois modules (Figure 3-2). Le module **environnement de simulation** va permettre de reproduire l'étape de « surveillance en environnement dynamique » du RPD en simulant l'environnement maritime dans lequel va se dérouler la mission. Le module **connaissance situationnelle** va reproduire les étapes « indices sur la situation » et « reconnaissance de schémas déjà vus » et le module **simulateur mental et composante décisionnelle** accomplira toutes les étapes suivantes du processus, c'est-à-dire :

- La création de plans d'actions.
- L'évaluation par simulation mentale.
- La préconisation du plan d'action.

De la même manière que le RPD, tant que la mission n'est pas terminée, c'est-à-dire le but non atteint, le système continuera de préconiser des décisions de manière autonome. Si les objectifs et le but sont modifiés, la base de données du SAD devra être réadaptée et les modules composant l'architecture devront être reconfigurés afin de traiter les nouvelles données et de s'adapter aux nouveaux objectifs et but de la mission.

La Figure 3-3 illustre notre architecture où les fonctionnalités de chacun des modules sont détaillées. L'architecture proposée diffère de celle de Kunde et Darken sur les points suivants : (1) les modules de simulation mentale et de décision sont fusionnés, (2) l'apprentissage par renforcement profond est utilisé pour la simulation mentale et (3) le module simulateur mental ne calcule pas de prédicteurs mais s'appuie sur les indices sur la situation enregistrés par le module **connaissance situationnelle** et également d'éventuels indicateurs qui sont fournis aux agents entraînés par apprentissage par renforcement profond.

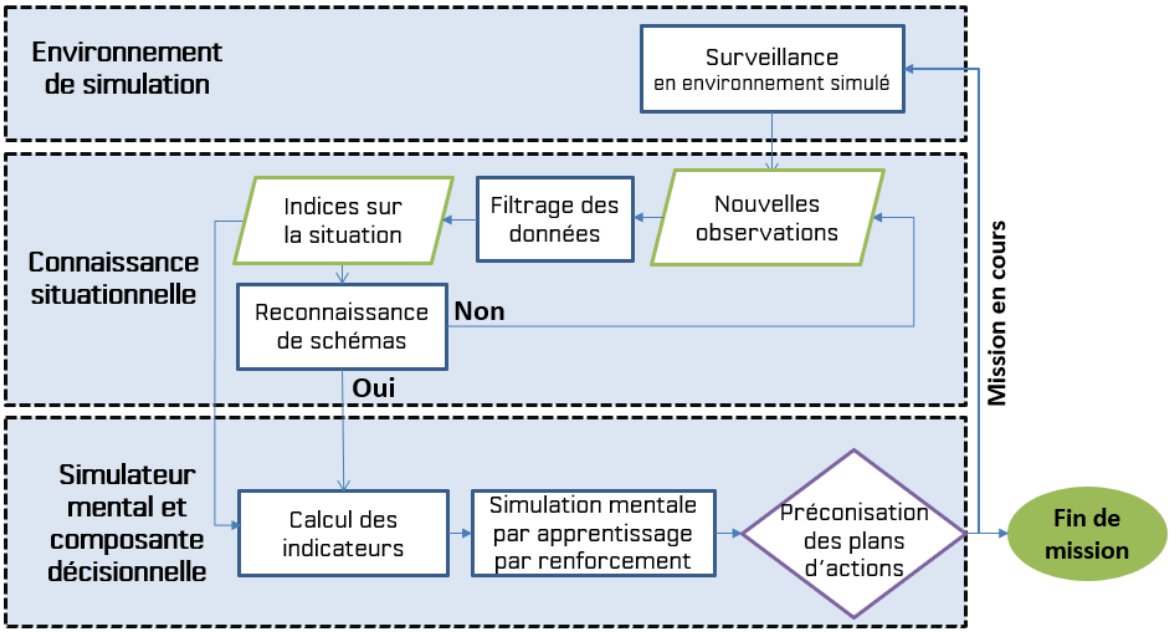

Figure 3-3 : Architecture cognitive proposée

3.1.2. Description des profils utilisateurs du système d'aide à la décision

Le but de cette thèse est de formaliser un système d'aide à la décision appliqué à la conduite de missions navales militaires au niveau tactique. Les utilisateurs de ce SAD seront donc les personnes à bord des navires militaires en charge de conduire la mission, à savoir les officiers (EM) et le Commandant se regroupant au sein de l'organe de commandement.

Ces différents utilisateurs peuvent avoir des niveaux d'expérience différents en conduite de missions, cependant aucun niveau d'expérience particulier n'est requis pour utiliser le SAD. De plus, comme évoqué dans le Chapitre 2, plus un décideur sera expérimenté, plus il sera rapide dans l'utilisation du système et en capacité d'analyser et de quantifier le bien-fondé de la préconisation des décisions faites par celui-ci.

Les utilisateurs doivent savoir aussi utiliser l'IHM qui fait office de lien avec le système, de sorte à pouvoir le faire fonctionner en interagissant avec lui. Dans le cadre de cette thèse, l'IHM réalisée sera très simple d'utilisation, une formation d'une demi-heure maximum est suffisante à un néophyte pour la prendre en main.

Dans la suite, l'architecture cognitive proposée appliquée à la conduite de missions navales est décrite en détail.

3.2. Description de l'architecture cognitive proposée

L'architecture cognitive proposée est composée de trois modules :

- Environnement de simulation
- Connaissance situationnelle
- Simulateur mental et composante décisionnelle

Ce découpage fonctionnel permet de modéliser toutes les étapes du RPD.

Dans la section suivante, les trois modules sont décrits de manière détaillée afin de montrer comment les étapes du RPD sont modélisées.

3.2.1. Le module environnement de simulation

Dans l'architecture proposée la surveillance en environnement dynamique est assurée par le premier module **environnement de simulation**. Ce module va permettre de recréer l'environnement dans lequel se déroule la mission et de le mettre à jour au cours du temps. Il s'agit ici d'un environnement maritime avec un niveau de réalisme, dépendant des hypothèses faites. Avant de décrire la façon dont le module va assurer la surveillance en environnement dynamique, l'environnement maritime simplifié est décrit ci-dessous.

3.2.1.1. Modélisation de la cinématique des navires et de la zone maritime

Le comportement des navires présents dans la zone maritime sera modélisé par une simple expression décrivant leur cinématique à trois degrés de liberté dans un plan horizontal 2D. Ce plan est muni d'un repère local $(O; x; y)$ dans le référentiel terrestre, et d'un repère lié au navire dans le référentiel terrestre noté $(G; x_G; y_G)$, tel qu'illustré Figure 3-4.

La cinématique d'un navire est définie par le système dynamique :

$$\dot{\eta} = R(\psi)v(\eta) \tag{1}$$

Où $\eta = [x, y, \psi]^T$ représente les positions dans le repère terrestre $(x; y)$ ainsi que l'angle de cap ψ, $v(\eta) = \left[u_x(\eta), u_y(\eta), r\right]^T$ représente les vitesses dans le repère du navire. $R(\psi)$ est la matrice de rotation du repère terrestre vers le repère du navire définie par :

$$R(\psi) = \begin{bmatrix} \cos(\psi) & -\sin(\psi) & 0 \\ \sin(\psi) & \cos(\psi) & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2}$$

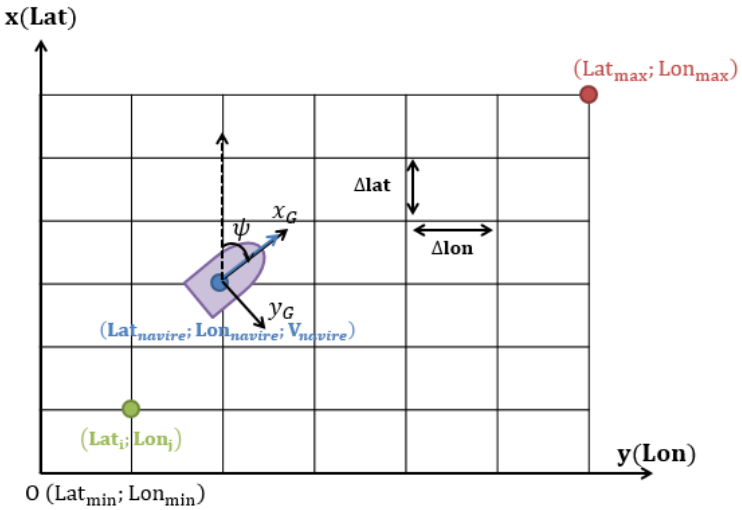

Figure 3-4 : grille de la zone maritime et paramètres cinématiques d'un navire

La zone maritime de la mission sera définie par $[\mathrm{Lat}_{min}; \mathrm{Lat}_{max}] \times [\mathrm{Lon}_{min}; \mathrm{Lon}_{max}]$ et discrétisée par une grille maillée régulièrement en latitude ($\Delta\mathrm{lat}$) et longitude ($\Delta\mathrm{lon}$), telle qu'illustrée Figure 3-4. Les données cartographiques seront exprimées en latitude et longitude selon un système géodésique connu comme par exemple le WGS84.

Les sections suivantes décrivent la modélisation de la surveillance en environnement dynamique par le module, en introduisant les données d'entrée à recueillir, les méthodes de traitement et les sorties attendues.

3.2.1.2. Les données à recueillir

Ce premier module va devoir acquérir les données utilisées par les décideurs, essentiellement issues des capteurs du navire. Ces données dépendront du type de mission, du but, des objectifs et des menaces attendues. Les données recueillies étant des données brutes provenant de plusieurs sources, elles peuvent être incomplètes, bruitées, erronées ou incohérentes.

3.2.1.3. Méthode(s) de prétraitement des données

La qualité des données est essentielle pour que les agents entraînés par apprentissage par renforcement profond, sollicités dans le dernier module **simulateur mental et composante décisionnelle,** soient performants. Voilà pourquoi des opérations de prétraitement appropriées doivent être implémentées, telles que :

- **La réduction des données** : les enregistrements de données peuvent être sous-échantillonnés pour faciliter leur manipulation tout en conservant l'information utile.
- **La transformation des données** : les données peuvent être normalisées afin d'augmenter la stabilité numérique.

- **Le nettoyage des données** : les valeurs manquantes peuvent être complétées (par exemple par interpolation), les données trop bruitées et aberrantes sont supprimées.

3.2.1.4. Sortie(s) attendue(s)

Pour **l'environnement de simulation** qui assure la surveillance en environnement dynamique, les sorties attendues sont les données brutes prétraitées ou les données simulées afin qu'elles soient directement exploitables par les modules suivants. Les données recueillies par le module **environnement de simulation** vont être introduites dans le module **connaissance situationnelle** qui va filtrer ces données afin d'obtenir les indices sur la situation.

3.2.2. Le module de connaissance situationnelle ou de reconnaissance de schémas déjà vus

Pour rappel, d'un point de vue cognition humaine le processus de reconnaissance de schémas consiste à faire correspondre les informations reçues d'un environnement extérieur aux informations déjà mémorisées. Le but du module **connaissance situationnelle** est de reproduire la reconnaissance de schémas déjà vus.

Le module **connaissance situationnelle** recueillera en entrée les indices sur la situation et devra déterminer en sortie si la reconnaissance de schémas déjà vus a eu lieu. Le module analysera les indices sur la situation jusqu'à ce qu'un schéma ou une situation, caractérisé par un contexte et des décisions ou actions possibles, soit reconnu. La section suivante décrit le choix des données d'entrée, des méthodes de traitement et des sorties possibles.

3.2.2.1. Les données à recueillir

Etant donné que l'architecture proposée doit traiter les données et les interpréter en suivant un processus de reconnaissance de schémas déjà vus conforme au raisonnement humain, le contexte et les données à analyser doivent être définis à l'aide d'experts lors d'entretiens ou lors de l'observation de leurs activités en situation réelle.

Dans le cas de l'application de l'architecture à la conduite de missions navales tactiques les indices ont été déterminés à l'aide d'opérationnels, c'est-à-dire d'acteurs ayant la connaissance du théâtre des opérations. En effet, lors de la conduite de mission navale l'organe de commandement analyse la situation tactique en recueillant des informations provenant de plusieurs sources telles que les capteurs à bord du navire ou les services de renseignement. Ces informations concernent à la fois l'ennemi, le terrain maritime sur lequel évolue le navire, les conditions météorologiques ainsi que les considérations d'ordre civil classées par ordre de priorité, et dépendent du type de mission. A partir de toutes ces informations, l'EM consigne les plus pertinentes appelées indices sur la situation en fonction du but et des objectifs fixés de la mission. Pour une mission navale générale à un niveau tactique, ces indices peuvent être :

- Des données issues du radar de navigation, du radar longue portée, des données AIS, des sonars, de l'infrarouge et des équipements optroniques pour surveiller les forces en présence, c'est-à-dire les forces ennemies, alliées et neutre.
- Les caractéristiques et les données propres au navire recueillies par les capteurs à bord afin d'avoir des informations sur « l'état de santé » du navire.
- Les conditions météorologiques notamment les vagues, la visibilité et le vent. Des prévisions météorologiques ou des données recueillies directement à bord peuvent être utilisées.
- Des données géoréférencées indiquant les côtes, les obstacles statiques ainsi que le profil bathymétrique pour éviter les eaux peu profondes.
- Les « no go area », où le navire ne doit pas aller comme les zones de pêche ou offshore par exemple.

Cette liste d'indices, non exhaustive, doit être modifiée selon les objectifs, le but de la mission et le type d'ennemi attendu. Par exemple, pour l'objectif d'une mission qui est de rallier une zone le plus rapidement possible, les conditions météorologiques, les « no go area » et les données issues du radar de navigation pourront être suffisantes, mais si l'objectif est de minimiser la consommation énergétique, des données concernant la consommation du moteur et la chaîne propulsive du bâtiment doivent être rajoutées.

Le module **connaissance situationnelle** devra être en capacité de recueillir cette liste d'indices parmi toutes les données fournies par le module **environnement de simulation**. Pour les obtenir, un filtrage des données recueillies par l'environnement de simulation doit être réalisé. Le filtrage des données est un processus consistant à affiner les ensembles de données pour obtenir simplement ce qui est utile. Différentes méthodes de filtrages de données peuvent être utilisés tels que des requêtes dans des bases de données.

3.2.2.2. Méthode(s) pour la reconnaissance de schémas déjà vus

Une fois les indices sur la situation déterminés, ils seront traités par une ou plusieurs méthodes pour la reconnaissance de schémas. Certaines méthodes ont déjà été citées dans le Chapitre 2 telles que les réseaux de neurones ou les structures LTM. Cette liste n'est pas exhaustive et d'autres méthodes peuvent être utilisées telles que la méthode de comparaison de modèles [123], ou la logique floue [124].

3.2.2.3. Identification des situations possibles

En sortie du module **connaissance situationnelle**, une identification de la situation est réalisée (contexte et actions associées). Dans le cadre de ce premier prototype, cette identification sera sous la forme d'un booléen permettant de valider si les indices sur la situation sont suffisants pour passer à l'étape de création de plans d'actions. Si le booléen renvoyé est TRUE la reconnaissance de schémas déjà vus a eu lieu, sinon le module récupèrera de nouveaux indices jusqu'à ce que le booléen renvoyé soit TRUE. Si des indices sont manquants ou indisponibles, le module **connaissance situationnelle** cessera de fonctionner, et aucune

décision ne sera préconisée. Ce cas est très peu probable dans le cadre de ce premier prototype car les indices sont pratiquement tous simulés.

3.2.3. Le module simulateur mental et composante décisionnelle

Pour rappel, dans le RPD la simulation mentale remplit diverses fonctions essentielles telles que l'évaluation de la situation, la prise en compte de la dynamique de la situation et l'évaluation des plans d'action.

Le but de ce module est d'anticiper l'évolution de la situation actuelle à partir des actions possibles et de choisir la ou les meilleures actions à accomplir en fonction du but et des objectifs fixés. Une fois les indices introduits, le module **simulateur mental et composante décisionnelle** va pouvoir préconiser une décision sous forme d'actions ou de plans d'actions à l'utilisateur. Il assurera les étapes de création de plan d'actions (i.e. : les actions possibles), d'évaluation par simulation mentale et de préconisation du plan d'actions (i.e. : choix et mise en œuvre des meilleures actions) du RPD. Ce module est sollicité lorsqu'une décision est requise, c'est-à-dire lorsque l'étape de reconnaissance de schémas déjà vus aura été validée par le module précédent. Dans la suite le choix des données d'entrée, des méthodes de traitement et des sorties possibles est décrit.

3.2.3.1. Identification des données à recueillir

Afin de pouvoir anticiper l'évolution de la situation et de choisir le meilleur plan d'actions le module aura besoin en entrée des actions possibles et des indices sur la situation. Les actions possibles sont sous la forme d'une base adaptée au contexte de la mission qui reste inchangée quel que soit la situation reconnue. Ces données sont en plus accompagnées d'indicateurs. Les indicateurs sont obtenus en corrélant certains indices sur la situation et peuvent rajouter de l'information pour le choix du meilleur plan d'actions. Ils sont propres au cas d'application et doivent être définis à l'aide d'experts. Lors de la conduite d'une mission navale, l'EM utilise des indicateurs pour l'aider à déterminer les progrès réalisés en vue d'atteindre les conditions de l'état final et de réaliser les objectifs. Ces indicateurs sont sous la forme de mesures d'efficacité (MOE), telles que définies dans le Chapitre 1. Dans le cadre de cette première modélisation, trois indicateurs communs à un grand nombre de missions sont définis et détaillés dans les parties suivantes : un premier pour calculer la vitesse maximale admissible du navire, un deuxième pour calculer le risque de collision entre le navire et les obstacles, utiles à la gestion de la trajectoire d'un navire, et un dernier pour calculer la probabilité de dissuasion d'une arme en fonction d'une menace, utiles à la gestion de la menace asymétrique d'un navire.

3.2.3.1.1. Calcul de la vitesse maximale admissible d'un navire.

La vitesse maximale admissible du navire est initialement calculée en fonction de sa position, des données météorologiques associées et de ses caractéristiques propulsives. La méthode de calcul réalisée en interne chez NAVAL GROUP, s'agence de la manière suivante :

(1) Le calcul de la vitesse maximale est issu de la formule suivante décrit dans [125] et [126] :

$$V = \frac{P(V)\eta_p}{\mathrm{RAV}(V)} \tag{3}$$

Avec V la vitesse maximale recherchée en nœud ; η_p le rendement propulsif global ; $P(V)$ la puissance propulsive requise en Watt et $\mathrm{RAV}(V)$ la résistance à l'avancement exercée par les forces extérieures sur le navire en Newton.

(2) La formule ci-dessus fait intervenir la résistance à l'avancement (RAV) [127]. Un modèle simplifié de RAV a été mis en place ne faisant intervenir que les quatre paramètres environnementaux les plus influents et les effets des vents contraires :

$$\mathrm{RAV}(V) = R_0(V) + \nu(V, \alpha, h)R_w(V, h) + F_x(V_v, \beta) \tag{4}$$

Avec R_0 la RAV sans vent ni houle ; ν le facteur correctif de l'angle de la houle, qui dépend de la hauteur de la houle et de l'angle d'incidence de la houle par rapport à l'axe du navire ; R_w la RAV pour une houle donnée de face et F_x la force du vent selon l'axe longitudinal du navire qui dépend de la vitesse du vent et de l'angle du vent par rapport à l'axe du navire.

(3) Une fois les formules de vitesse et de RAV du navire établies, la procédure de calcul de la vitesse maximale admissible a été mise en place. Comme la vitesse dépend de la RAV et de la puissance et ces dernières dépendent de la vitesse, plusieurs boucles de convergence ont été mises en place, chronophages, pour obtenir la vitesse maximale. Les étapes principales de l'algorithme sont les suivantes :

Etant donnée une vitesse initiale postulée V :

- Calcul du point de fonctionnement de l'hélice,
- Détermination de la RAV,
- Détermination du point de fonctionnement moteur,
- Vérification de la compatibilité avec le champ moteur,
- Calcul de la puissance disponible,
- Vérification de la possibilité effective d'atteindre V.

Si la vitesse V est atteinte, on l'augmente, sinon on décroit V et on itère jusqu'à atteindre une tolérance suffisante. La valeur Vest trouvée à l'aide de la méthode itérative du point fixe. Le calcul de la vitesse maximale est sollicité dans l'architecture cognitive à chaque nouvelle

préconisation de décision. Cependant, cette méthode de calcul est numériquement très gourmande en temps de calcul et demande souvent plusieurs dizaines de minutes pour l'obtention d'un résultat.

Afin que l'architecture cognitive puisse proposer une décision dans un temps acceptable, un modèle permettant de prédire rapidement la vitesse maximale admissible en fonction de 5 variables météorologiques impliquées dans le calcul de la vitesse maximale admissible, à savoir la hauteur et la direction de la houle, la vitesse et la direction du vent et la température de l'air et des caractéristiques propulsives du navire a été réalisé. Ce modèle prédictif se fonde sur la méthode des forêts aléatoires (forêts d'arbres décisionnels).

Les forêts aléatoires, Random Forest [128], sont une famille d'algorithmes d'apprentissage automatique robustes qui peuvent être utilisés pour la régression (Random Forest Regressor) ou pour la classification (Random Forest Classifier). Random Forest est un modèle de forêts aléatoires composé d'un grand nombre de petits arbres de décision, appelés estimateurs, qui produisent chacun leurs propres prédictions. Le modèle de forêts aléatoires combine les prédictions des estimateurs pour produire une prédiction plus précise. L'utilisation de Random Forest permet de réduire et contrôler le sur-apprentissage (overfitting) dans les arbres de décision et d'améliorer la précision. De plus la normalisation des données n'est pas nécessaire. Cependant, il requiert une grande puissance de calcul car en pratique de nombreux arbres sont à combiner pour améliorer les résultats, ce qui conduit à une combinatoire élevée et à des résultats difficiles à interpréter.

3.2.3.1.2. Calcul de la probabilité de dissuasion d'une arme

La probabilité de dissuasion d'une arme (PEA : Probabilité Effet Arme) se calcule de la manière suivante :

$$PEA = 1 - coeff_{menace} MOE_{arme}(D_{but}, G_a, V_a, I_a, V_E, t_n, V_n) \quad (5)$$

Cette formule se décompose en deux termes. Le premier, $coeff_{menace}$, représente le niveau de menace d'un navire en fonction de sa taille et de sa vitesse sous la forme d'un coefficient compris entre 0 et 1. Le tableau ci-dessous montre un exemple de coefficients de menace en fonction du type de navire :

Tableau 3-1 : Exemple de coefficient de menace en fonction du type de bateau

Type de navire	Coefficient de menace
Embarcation rapide	[0,6; 0,9[
Navire de plaisance	[0,3; 0,6[
Gros navire	[0; 0,3[

Plus le navire sera petit et plus il sera considéré comme menaçant car il sera en capacité de se déplacer à très grande vitesse.

Le second terme $MOE_{arme}(D_{but}, G_a, V_a, I_a, V_E, t_n, V_n)$, définit la probabilité d'efficacité de dissuasion de la menace par le dispositif choisi de vitesse V_E(m/s) et de temps de neutralisation t_n(s) en fonction de D_{but} (m) distance du début de l'attaque, G_a(°) gisement sous

lequel l'attaquant est détecté, V_a(m/s) vitesse de l'attaquant, I_a (°) inclinaison de l'attaquant par rapport à G_a et V_n(m/s) la vitesse du navire.

Dans le calcul de la probabilité de dissuasion, MOE_{arme} sera pondérée par le terme $coeff_{menace}$ pour tenir compte de la dangerosité du type d'embarcation. Plus un navire sera considéré comme menaçant ou dangereux et plus la probabilité de le dissuader sera faible. Par exemple, si la valeur de MOE_{arme} est de 0,6 pour une arme donnée et que le coefficient de menace est de 0,1, cela signifiera que cette arme aura 94% de chance de dissuader l'attaquant. Au contraire, si le coefficient de menace est de 0,9 cette arme n'aura que 46% de chance de dissuader l'attaquant.

3.2.3.1.3. Calcul du risque de collision

L'indicateur relatif au risque de collision est simplifié dans un premier temps et calculé à l'aide de la distance euclidienne entre le navire et ses obstacles voisins :

$$d = \sqrt{(x_{navire} - x_{obstacle})^2 + (y_{navire} - y_{obstacle})^2} \quad (6)$$

Un périmètre de sécurité autour des obstacles de rayon $r_{obstacle}$ et autour du navire de rayon r_{navire} sera défini tel qu'illustré Figure 3-5. Le rayon r_{navire} est égal à $3{,}6L$, où L est la longueur du navire et cette valeur est extraite du modèle de Fuji [129]. Le rayon $r_{obstacle}$ est supposé le même pour tous les obstacles et le risque est évalué à 100% (situation de danger), si la distance euclidienne est inférieure au plus grand des deux périmètres de sécurité :

$$\mathrm{d} < \max(r_{obstacle}, r_{navire}) \quad (7)$$

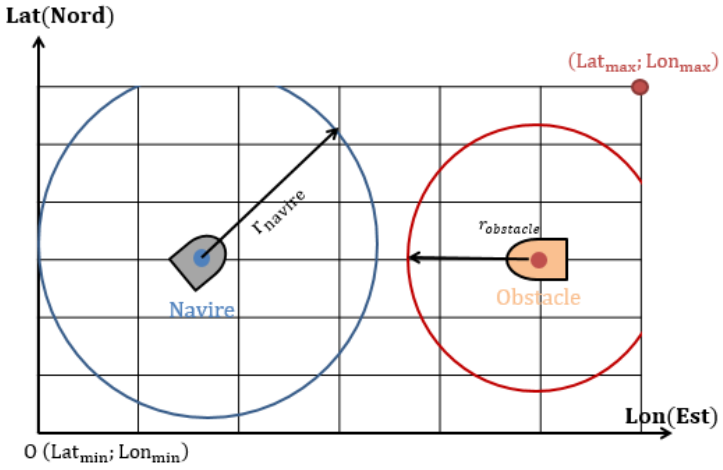

Figure 3-5: Représentations des périmètres de sécurité

3.2.3.2. Méthode(s) pour la simulation mentale

La simulation mentale peut être modélisée par une méthode basée soit sur un modèle prédictif tel que des chaînes de Markov cachées [130], des filtres de Kalman [131] qui fournissent une prédiction de la survenue des états futurs, soit sur des modèles statistiques qui fournissent des probabilités d'événements basées sur l'observation préalable de ces derniers, ou prédisent de nouveaux événements grâce aux événements observés précédemment.

Dans le cadre de cette thèse, la méthode utilisée pour la simulation mentale est l'apprentissage par renforcement profond basé sur un processus de décision Markovien.

Comme présenté dans le Chapitre 2, l'AR profond est adapté pour modéliser en grande partie la simulation mentale. Un PDM fournit une probabilité de passage à l'état suivant en fonction des actions choisies et donne donc une estimation sur la survenue d'un événement futur, assurant la fonction d'anticipation. La simulation mentale basée sur l'apprentissage par renforcement profond est décrite ci-dessous.

Les indices, la base d'actions possibles et les indicateurs introduits vont constituer les données d'entrée des agents entraînés par apprentissage par renforcement profond, qui en sortie vont choisir le ou les plans d'actions les plus adaptés. Chaque plan d'actions est déterminé grâce à la politique π qu'ils auront apprise lors de l'entraînement, constituant leur expérience ou mémoire. Cette fonction politique contenue dans les réseaux de neurones de chaque agent permet de projeter chaque action de la base d'actions possibles dans le futur et d'anticiper à court terme son effet. Elle cherche à déterminer l'action qui va maximiser pour chaque nouvelle situation ou état de l'agent la récompense cumulée de l'agent. Rappelons que la récompense cumulée ou gain permet de ne pas considérer uniquement la récompense immédiate qu'une action pourrait apporter mais d'anticiper les récompenses attendues dans les prochains états. L'agent est donc capable d'anticiper à court terme l'effet de ses actions immédiates lui permettant d'anticiper l'évolution de la situation. Les fonctions politiques de chaque agent sont apprises à l'aide de méthodes d'apprentissages, les méthodes sans modèle ou basées sur des modèles, introduites dans le Chapitre 2. Quel que soit la méthode choisie, celle-ci doit être modélisée comme illustrée par un exemple de pseudo code de l'algorithme PPO-clip [132] Figure 3-6.

Une fois les fonctions politiques de chaque agent déterminées grâce à l'apprentissage, ces agents sont comparables à des « décideurs expérimentés » et préconisent directement des actions dans un environnement changeant et inédit. Selon le degré de connaissance de l'environnement dynamique, la politique peut être déterministe auquel cas à chaque état appartenant à un ensemble S une ou plusieurs actions appartenant à un ensemble A sont associées, ou alors probabiliste de sorte que la politique $\pi : S \times A \rightarrow [0,1]$ est définie par $\pi(a,s) = Pr(a_t = a|s_t = s)$. Il s'agit alors de la probabilité que l'agent choisisse d'exécuter l'action a dans l'état s. Les politiques des agents sont déterministes dans notre cas.

Algorithme : PPO-clip

1 : Entrée : Initialisation des poids de la politique θ_0, initialisation des paramètres de la fonction valeur ϕ_0

2 : **Pour** k = 0,1, 2,… **faire** :

3 : Collecter le jeu de I trajectoires $D_k = \{\tau_i\}_{i=1\ldots I}$ en exécutant la politique $\pi_k = \pi(\theta_k)$ dans l'environnement, où $\tau = \{s_t, a_t, r_t\}$

4 : Calculer les récompenses cumulées totales à gagner $\hat{R}_t$

5 : Calculer l'estimation de la fonction avantage, $\hat{A}_t$, basée sur la fonction valeur V_{ϕ_k}

6 : Mettre à jour la politique en maximisant PPO-clip, typiquement via une montée du gradient stochastique avec l'algorithme Adam [133] , privilégié dans PPO-clip :

$$\theta_{k+1} = \underset{\theta \in \Theta}{\operatorname{argmax}} \left[\frac{1}{|D_k|T} \sum_{\tau \in D_k} \sum_{t=0}^{T} \min \left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_k}(a_t|s_t)} A^{\pi_{\theta_k}}(s_t, a_t),\ g\left(\epsilon, A^{\pi_{\theta_k}}(s_t, a_t)\right) \right) \right]$$

7 : Mettre à jour la fonction valeur en déterminant l'erreur quadratique moyenne minimale avec un algorithme de descente de gradient :

$$\phi_{k+1} = \underset{\phi}{\operatorname{argmin}} \left[\frac{1}{|D_k|T} \sum_{\tau \in D_k} \sum_{t=0}^{T} \left(V_\phi(s_t) - \hat{R}_t \right)^2 \right]$$

8 : **fin**

Figure 3-6 : Pseudocode d'une méthode d'apprentissage sans modèle basée sur la recherche de politique [11]

Lors de la conduite d'une mission navale le navire doit exécuter différentes actions, telles qu'engager une arme le plus rapidement possible ou changer de cap et de vitesse, relatives aux deux tâches principales d'une mission, à savoir la gestion de la route et la gestion des événements indésirables. C'est pourquoi le navire est décomposé en deux agents, un agent par tâche, et chaque agent est entraîné afin de préconiser des décisions sous forme d'actions.

Afin de faire apprendre la fonction politique par chaque agent, leur environnement d'apprentissage doit être configuré (Cf. Figure 3-7).

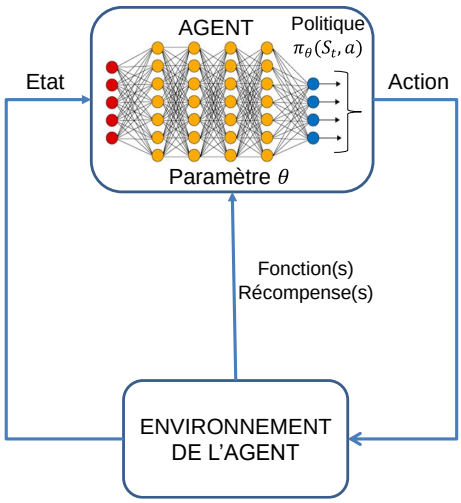

Figure 3-7 : Configuration de l'apprentissage par renforcement profond d'un agent

L'environnement de l'agent utilise des données et des informations issues des modules **environnement de simulation** et **connaissance situationnelle**. **L'environnement de simulation** est utile à l'agent pour connaître le contexte dans lequel il évolue, ici un environnement maritime. Le module **connaissance situationnelle** fournit à l'agent les indices sur la situation qui lui permettent de mettre à jour ses états en intégrant directement les indices ou alors en utilisant les indicateurs. Lorsque le type de mission change les états, actions et fonctions récompenses doivent être mis à jour. Pour des missions navales, d'un point de vue tactique les états modélisent toujours le comportement d'un navire et ses observations, c'est-à-dire les données recueillies via ses capteurs et les actions, qui doivent être adaptées aux capacités d'un navire (changement de cap, accélération, sélection d'une arme).
Les fonctions récompenses permettent aux agents d'apprendre des comportements souhaités pour exécuter les différentes tâches et objectifs d'une mission (cf : Chapitre 2), elles sont corrélées aux objectifs et au but de la mission. Ainsi chaque agent accomplie une tâche différente, par exemple gérer la route en atteignant le plus rapidement possible la zone d'arrivée, et chaque objectif de la tâche doit être retranscrit dans une fonction récompense, il y a souvent autant de fonctions récompenses que d'objectifs. Une fois les agents entraînés sur leurs tâches et objectifs respectifs ils sont sollicités sur chacune de leur tâche pour accomplir la mission dans sa totalité. De plus, ces fonctions récompenses peuvent contraindre l'agent dans son évolution à l'intérieur du domaine qui constitue son environnement, comme par exemple éviter d'aller dans une zone qui lui est interdite. Le but de la mission doit aussi être mis sous la forme d'une fonction récompense. La mise en place de ces fonctions est l'étape la plus importante et cruciale en AR pour qu'un agent apprenne à réaliser les objectifs souhaités et atteigne le but demandé. Si l'utilisateur souhaite changer les objectifs ou le but de la mission, le(s) agent(s) intelligent(s) doivent être ré-entraînés ou de nouveaux agents sont créés afin d'acquérir une nouvelle expérience.

3.2.3.3. Identification des décisions possibles

En sortie du module, les agents entraînés préconisent des décisions relatives à la gestion de la mission aux décideurs qui peuvent décider de suivre ou non la préconisation. Ces décisions peuvent être par exemple des valeurs de caps ou des préconisations d'armes. En plus de ces décisions, un ou plusieurs indicateurs peuvent être fournis à l'utilisateur pour rajouter une certaine explicabilité de la décision préconisée par l'agent, tels que ceux définis en 3.2.3.1. Ces indicateurs ne sont pas exhaustifs et d'autres peuvent être définis.

3.3. Définition de l'IHM

L'IHM est ici l'outil qui permet à l'opérateur de donner des instructions au SAD qui en retour renseigne sur l'état d'exécution de la mission. Dans ce cas, l'IHM a pour fonctions principales la gestion des données par l'utilisateur, l'affichage de celles-ci ainsi que l'affichage des décisions préconisées par le système en temps réel. La conception d'une IHM doit être élaborée en collaboration avec ses futurs utilisateurs, le SAD doit avoir un fonctionnement simple avec une interface facile à comprendre. Pour un utilisateur expert, une IHM plus

sophistiquée peut-être attendue, proposant plus de détails ainsi qu'une prise en compte de l'ergonomie via les Facteurs Humains. Dans tous les cas, l'IHM doit afficher les décisions préconisées par l'architecture cognitive, placées dans le champ visuel central de l'opérateur, pour lui fournir les informations nécessaires à l'évaluation et au traitement de la situation. Les décisions préconisées par l'architecture peuvent être appuyées avec l'affichage des indicateurs, rajoutant de l'explicabilité à la simulation affichée. Dans le cadre de notre système d'aide à la décision, l'IHM permet de reproduire la situation tactique du navire en 2D de manière simple sur laquelle sont superposées les décisions préconisées par l'architecture cognitive de manière dynamique.

L'IHM proposée est composée :

(1) D'un wrapper, c'est-à-dire d'une interface entre l'architecture cognitive et l'IHM matérialisée par un serveur web API composé d'un ensemble de règles définissant comment les données d'entrée et de sorties des modules transitent entre l'architecture cognitive et l'IHM. Le wrapper a donc vocation à connecter l'architecture cognitive et l'IHM et à gérer la communication entre elles.

(2) De la composante d'affichage qui va permettre à l'opérateur de visualiser les actions préconisées et d'interagir avec le système (mettre en pause, sélectionner de nouveaux points de départs de la mission, arrêter d'avoir ses préconisations).

3.4. Conclusion

Dans ce chapitre, la modélisation de notre architecture cognitive composant notre système d'aide à la décision reproduisant les étapes du RPD a été décrite. Notre architecture cognitive dont la structure s'inspire de celle de Kunde et Darken [82], [83], [84] est composée de trois modules : le module **environnement de simulation** modélisant l'environnement maritime de la mission et assurant la surveillance en environnement dynamique du RPD, le module **connaissance situationnelle** pour les étapes suivantes du RPD jusqu'à la reconnaissance de schémas déjà vus et le module **simulateur mental et composante décisionnelle** assurant toutes les autres étapes du RPD jusqu'à la préconisation d'actions. Dans ce module la modélisation de la simulation à l'aide de l'AR profond a été explicitée.

Une fois l'architecture configurée pour la conduite de missions navales, le système d'aide à la décision est prêt à être utilisé en temps réel. Avant de la faire fonctionner, l'utilisateur devra renseigner via l'IHM les données d'entrée, telles que le but et la zone maritime sur laquelle doit se dérouler la mission. Le système d'aide à la décision fonctionne en temps réel et en sortie de l'architecture cognitive les agents préconiseront des décisions sous la forme d'actions ou de plan d'actions. Ces décisions seront présentées via l'IHM à l'utilisateur qui aura le choix de les appliquer ou non. Certaines variables de l'environnement de simulation seront mises à jour et le processus de décision de l'architecture continuera jusqu'à ce que le but de la mission soit atteint.

Cependant une limite peut être déjà énoncée : la modélisation de la reconnaissance de schémas déjà vus a été simplifiée par rapport au RPD théorique.

Dans le chapitre suivant le prototypage de l'architecture cognitive sur un scénario de mission navale est décrit.

Chapitre 4 : Prototypage de l'architecture cognitive sur un scénario de mission navale

Ce chapitre présente le prototypage de l'architecture cognitive proposée dans le chapitre précédent. Pour rappel, notre SAD est composé d'une base de données, d'une architecture cognitive proposant des décisions et d'une IHM. Ces trois sous-systèmes communicants, fonctionnant de façon indépendante les uns des autres peuvent ainsi être prototypés séparément. Les bases de données associées à ce premier prototype exploratoire et une IHM développée afin d'évaluer le prototype sont également présentées.

Dans le cadre de cette thèse la base de données et l'architecture cognitive sont implémentées en Python 3. L'IHM, développée en JavaScript, a été réalisée en interne à NAVAL GROUP. Dans la suite, la constitution de la base de données, le prototypage de l'architecture cognitive et les spécifications de l'IHM sont décrits dans le but de venir en aide sur une mission navale. **Dans le cadre du premier démonstrateur développé, l'aide à décision porte sur deux tâches qui sont la gestion de la route optimale et la gestion des événements indésirables réduits aux menaces asymétriques.**

Avant de détailler le prototypage informatique, le scénario de mission navale sur lequel l'architecture cognitive sera prototypée est énoncé.

4.1. Description du scénario de mission navale

Suite à des entretiens avec d'anciens opérationnels, la mission choisie a été la suivante : un bâtiment de surface, plus précisément une frégate, doit se rendre dans une zone géographique pour y réaliser une action. Ce type de mission a été choisi car il représente 80% des missions tactiques navales.

Lors de ce type de missions, la frégate doit accomplir deux tâches principales :

(1) Emprunter la meilleure route pour atteindre la zone d'arrivée le plus rapidement possible.
(2) Gérer les menaces asymétriques auxquelles elle sera potentiellement confrontée.

Lors de la survenue de menaces asymétriques le but est de dissuader l'adversaire de poursuivre sa route de collision et non de le détruire directement. Les opérationnels se dotent toujours de mesures de dissuasion par l'utilisation d'armes non-létales, introduisant une action en deux temps :

- *Phase de dissuasion* : cette phase est composée d'un certain nombre de règles d'engagement impliquant des choix d'armes non-létales effectués selon la situation.

- *Phase de neutralisation :* La phase de neutralisation est divisée en règles d'engagement faisant intervenir des armes létales.

Pour cette première configuration de l'aide à la décision, un certain nombre d'hypothèses simplificatrices sont émises :

- L'aide à la décision ne porte que sur la gestion de la route et des menaces asymétriques.
- La superficie de la zone maritime sur laquelle se déroule la mission est bornée, de taille telle qu'il est possible de l'assimiler localement à un plan, et la zone maritime est une grille discrétisée en latitude et longitude.
- La mission s'effectue en eau non resserrée.
- Les mouvements de la frégate et des obstacles mobiles sont cinématiques.
- Les règles de navigation (COLREG) n'ont pas été prises en compte.
- Les conditions météorologiques ont seulement de l'influence sur la vitesse de la frégate.
- La zone d'arrivée est circulaire de rayon 20 km, toujours centrée en latitude et longitude maximale de la grille.
- L'objectif de la frégate est de minimiser son temps de trajet ce qui revient à optimiser sa vitesse. La vitesse de la frégate est supposée toujours égale à sa vitesse maximale admissible et le modèle prédictif de vitesse maximale admissible défini Chapitre 3 est utilisé.
- Les menaces asymétriques sont des navires de petites tailles par rapport au porteur, elles arriveront de manière convergente sur la frégate à grande vitesse dans le but de rentrer en collision avec elle.
- Pour simplifier le problème, la trajectoire de chaque menace asymétrique est considérée dans un premier temps comme convergente de façon radiale vers le navire, telle qu'illustrée Figure 4-1. La menace évoluera normalement dans le trafic maritime puis changera sa trajectoire brutalement à partir d'une certaine distance, appelée distance minimale de passage (distance à partir de laquelle le navire attaqué ne peut plus manœuvrer pour éviter l'attaquant), pour arriver à grande vitesse sur le navire exécutant la mission (Figure 4-1).

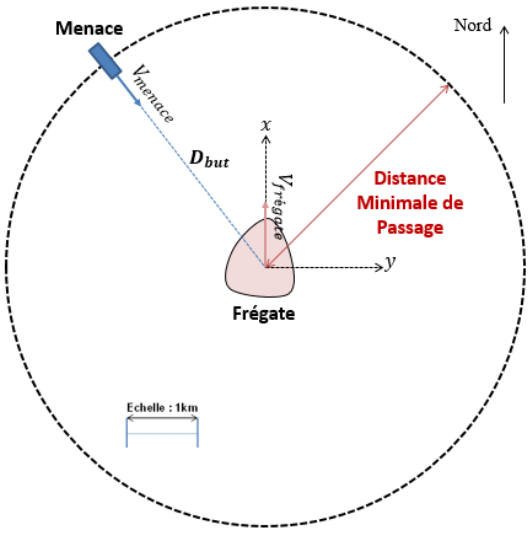

Figure 4-1: Exemple d'une attaque asymétrique

- L'aide à la décision ne porte que sur la phase de dissuasion, et les règles d'engagement ne sont pas prises en compte. Le but de l'aide à la décision lors de la survenue de menaces asymétriques sera de préconiser la meilleure arme non létale à utiliser pour

avoir le plus de chance de dissuader la menace. Il n'y a aucune restriction et règles sur l'utilisation des armes.
- Les réactions des menaces suite à l'engagement des armes ne sont pas modélisées. Il est supposé que toutes les menaces s'introduiront dans le périmètre de neutralisation mettant fin à la phase de dissuasion.

De plus lors de son transit la route de la frégate sera ponctuée par :

- Des conditions météorologiques variables en espace et en temps. Ces conditions météorologiques seront extraites d'une base de données décrite dans les sections suivantes.
- Des obstacles fixes : côtes, zones de danger, points à éviter et autres.
- Des obstacles mobiles : trafic maritime, dangers dérivants et autres.

Toutes les données relatives aux obstacles fixes et mobiles sont calculées à partir des données radars.

Une fois les hypothèses établies, l'architecture cognitive va être configurée pour fonctionner sur le scénario de mission.

4.2. Implémentation de l'architecture cognitive

Les trois modules de l'architecture cognitive sont réalisés en Python et suivent le formalisme de la programmation orientée objet (POO) [133]. La POO est un paradigme de programmation qui fournit un moyen de structurer les programmes de manière à ce que les propriétés et les comportements soient regroupés dans des objets individuels. Un objet est une entité ou une variable à laquelle sont attachées deux entités : des données et des fonctions qui agissent sur ces données. Les données sont appelées attributs de l'objet, et les fonctions, méthodes de l'objet. La structure d'un modèle commun pour un ensemble d'objets est appelée sa classe.

La Figure 4-2 ci-dessous, illustre le contenu des modules de notre architecture cognitive. Dans ce prototype, les modules **environnement de simulation** et **connaissance situationnelle** ont été regroupés au sein de la même classe car ils partagent des attributs communs et le module **simulateur mentale et composante décisionnelle** a été divisé en deux modules. Dans la suite, une description détaillée du prototypage de chacun des modules est réalisée.

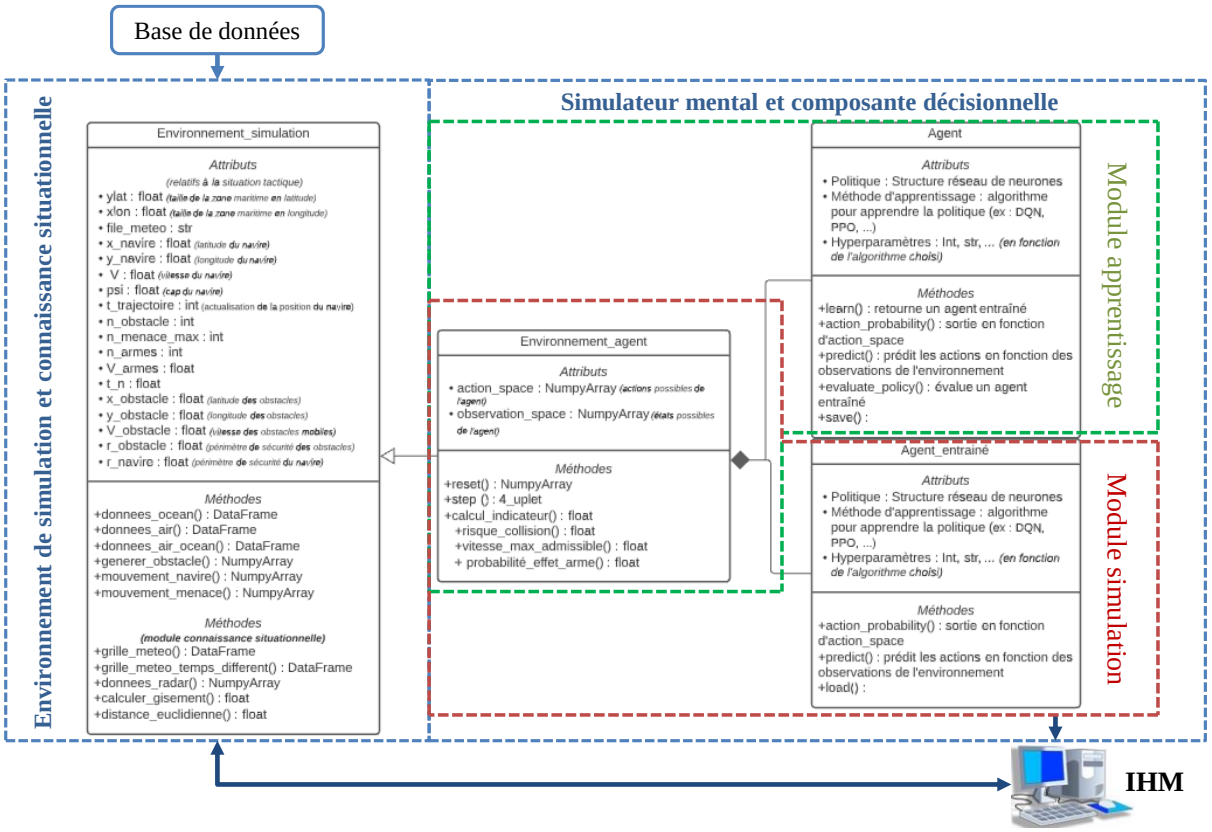

Figure 4-2 : Présentation du système d'aide à la décision sous forme d'un diagramme de classe UML

4.2.1. Implémentation du module environnement de simulation

Comme évoqué dans le Chapitre 3 le but du module **environnement de simulation** est de recréer l'environnement dynamique dans lequel se déroule la mission, ici un environnement maritime adapté à une mission navale au niveau tactique. L'environnement maritime ou situation tactique sera instancié sous la forme d'une classe, *Envrionnement_simulation*, illustrée Figure 4-2. Pour cette première implémentation, notre classe comporte des attributs relatifs à la zone maritime, au trafic maritime et au navire qui doit exécuter la mission**.**

L'avantage de modéliser la situation tactique sous forme de classe est de pouvoir facilement rajouter des attributs et des méthodes pour gagner en niveau de réalisme. Une fois les attributs instanciés, des méthodes vont être définies afin de manipuler les attributs de la classe, les bases de données, créer des instances de classe et les détruire, et de modéliser l'environnement du navire. Ces méthodes et attributs sont décrits ci-dessous.

Pour l'implémentation de l'environnement du navire tous les paramètres relatifs à la zone maritime dont une partie a été définie dans le Chapitre 3 telle que $\mathrm{Lat}_{min}, \mathrm{Lat}_{max}, \mathrm{Lon}_{min}, \mathrm{Lon}_{max}, \Delta\mathrm{lat}, \Delta\mathrm{lon}$ seront rajoutés en tant qu'attributs dans la classe *Environnement_simulation.* Le Tableau 4-1 ci-dessous résume l'ensemble des valeurs des attributs utilisées et qui resteront constantes tout au long de la mission.

Tableau 4-1: Tableau récapitulatif des attributs utilisés pour une zone maritime considérée

Attribut	Valeur
x_{lon}(taille de la zone maritime en latitude)	1,5°
y_{lat}(taille de la zone maritime en longitude)	1,5°
r_{navire}	430 m
$r_{obstacle}$	200 m
ΔLat	0,001°
ΔLon	0,001°
ΔLat_{meteo}	0,5°
ΔLon_{meteo}	0,5°
$t_{trajectoire}$	8 minutes
$n_{menaces\_max}$	4
$n_{obstacle}$	60
n_{armes}	3
V_{armes} (vitesse des armes)	Non spécifié
t_n (temps de neutralisation)	Non spécifié
t_{menace}	10 secondes
Distance_dissuasion	Non spécifié
Distance_neutralisation	Non spécifié

La zone maritime de la mission sera toujours une grille de 1,5° (x_{lon}) par 1,5° (y_{lat}) équivalent à une zone d'environ 168 km par 168 km maillée régulièrement en latitude (ΔLat) et longitude (ΔLon). Sur cette grille les données météorologiques sont discrétisées tous les 0,5° en latitude et longitude ($\Delta Lat_{meteo}, \Delta Lon_{meteo}$) et la frégate et les obstacles mobiles auront un mouvement discret avec leurs nouvelles positions actualisées toutes les 8 minutes ($t_{trajectoire}$). Lors de la mission, la frégate pourra être confrontée au maximum à 4 menaces asymétriques simultanées ($n_{menaces\_max}$) et 3 armes non-létales différentes (n_{armes}) pourront être préconisées. Les menaces évolueront normalement dans le trafic maritime, composé au maximum de 60 obstacles ($n_{obstacle}$), puis changeront leur trajectoire brutalement à partir d'une certaine distance, ici Distance_dissuasion, pour converger sur la frégate jusqu'à atteindre Distance_neutralisation. Distance_dissuasion ou distance minimale de passage est la distance à partir de laquelle la frégate ne peut plus manœuvrer pour éviter la collision avec les menaces. Pour des raisons de confidentialité, les caractéristiques des armes non-létales, la vitesse des armes (V_{armes}) et leur temps de neutralisation (t_n), ne sont pas explicitées. Lors de la confrontation à des menaces asymétriques, les positions des menaces, des obstacles mobiles et de la frégate seront actualisées toutes les 10 secondes (t_{menace}).

Une fois l'environnement maritime implémenté, le module **environnement de simulation** doit recueillir un certain nombre de données afin de pouvoir assurer la fonction de surveillance en environnement dynamique du RPD. Dans le cadre de la mission choisie et de ce premier prototype où l'architecture cognitive doit aider à la gestion de la route et de la menace asymétrique, les données impliquées seront les données météorologiques, les caractéristiques de la frégate et les données radars, telles que les coordonnées géographiques des navires voisins.

Ce choix de données réside dans le fait que les données météorologiques associées aux grandeurs propulsives de la frégate permettent d'avoir des informations sur les zones préférentielles à emprunter et sur la vitesse maximale admissible du navire et les données radars

permettent d'obtenir l'évolution de la situation tactique de manière dynamique. Les sections suivantes précisent la nature et l'origine de ces données.

4.2.1.1. Les données météorologiques

Les données météorologiques ont été téléchargées sur le site Web de l'ECMWF [134]. L'ECMWF ou European Centre for Medium-Range Weather Forecasts est le Centre européen pour les prévisions météorologiques à moyen terme, proposant un large choix de jeux de données. Pour cette première implémentation, le jeu de données ré-analysées ERA5 a été utilisé. Les données de ré-analyse météorologiques sont produites en combinant des mesures des observations réelles et des données de simulation à l'aide de techniques d'assimilation de données pour arriver à la description la plus réaliste possible des phénomènes météorologiques [134]. Ces données décrivent les conditions météorologiques d'une portion de la Terre, représentée par une carte repérée en latitude-longitude. Les données ont été rééchantillonnées à une résolution spatiale de 0,5 degrés ($\Delta\mathrm{Lat}_{\mathrm{meteo}}, \Delta\mathrm{Lon}_{\mathrm{meteo}}$), correspondant à environ 56 km. Les données sont disponibles à différentes résolutions horaires, telles que toutes les heures ou toutes les trois heures.

A l'aide du jeu de données ERA5, une base de données météorologiques a été constituée sur plusieurs périodes de l'année en différents lieux et moments de la journée. Les données sont actualisées toutes les 3 heures et seules les 5 variables météorologiques ayant le plus d'influence sur le déplacement d'un navire ont été retenues, à savoir la direction du vent et de la houle (°), la vitesse du vent (m/s), la hauteur de la houle (m) et la température de l'air (°C). La Figure 4-3 illustre la variable hauteur de houle disponible dans ce jeu de données en mer Méditerranée.

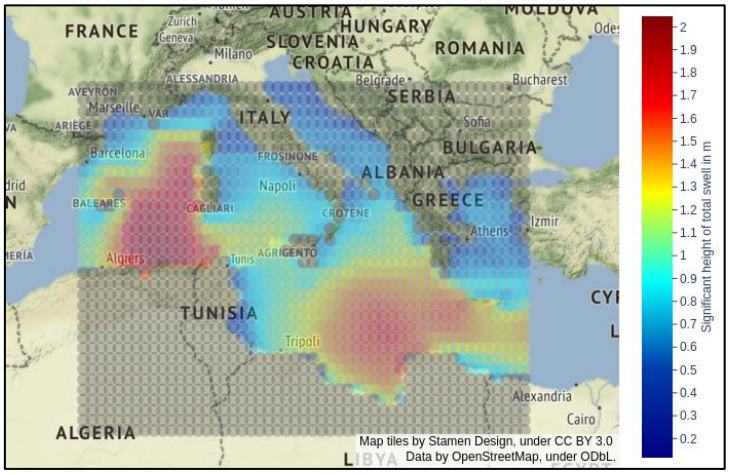

Figure 4-3 : Répartition de la hauteur de houle (m) dans une partie de la mer Méditerranée

N'ayant qu'une seule source de données provenant de l'ECMWF, le choix pour préparer et prétraiter ces données météorologiques s'est porté sur Python. Ce choix a été fait car (1) l'architecture modulaire étant entièrement implémentée à l'aide de Python, la communication entre base de données et architecture sera facilitée et (2) la librairie netCDF4 [135] utilisée pour traiter les jeux de données ERA5 téléchargés au format NetCDF, est en Python. La Figure 4-4 montre un exemple de DataFrame obtenu après prétraitement et mise en forme des 5 variables météo considérées :

lats	longs	mdww	mwd	shts	time	si10	t2m
39.0	11.5	320.079468	226.862000	0.741001	2013-04-01	7.698500	-4.019653
39.0	12.0	18.331650	244.127808	0.712566	2013-04-01	7.629588	-3.956329
39.0	12.5	10.668335	263.920593	0.650181	2013-04-01	7.546519	-3.908813
39.0	13.0	20.051086	277.978210	0.585773	2013-04-01	7.463451	-3.862915
39.0	13.5	15.996948	282.142242	0.572782	2013-04-01	7.380383	-3.815399

Figure 4-4 : Exemple d'un DataFrame obtenu après prétraitement

Avec :

- lats : latitude (°)
- longs : longitude (°)
- mdww : la direction du vent (°)
- mwd : la direction de la houle (°)
- shts : la hauteur de la houle (m)
- time : la date
- si10 : la vitesse du vent (m/s)
- t2m : la température de l'air (°)

4.2.1.2. Les caractéristiques du navire

Pour des raisons de confidentialité seules les grandeurs propulsives ont été considérées afin de décrire la dynamique du navire. Ces grandeurs propulsives sont réelles.

4.2.1.3. Les données radars

Les données radars permettent de recueillir entre autres des informations sur la position en latitude et longitude des obstacles présents dans la zone maritime mais également la vitesse des obstacles mobiles et leurs directions par rapport au navire qui exécute la mission. Dans notre démonstrateur, ces données sont simulées de manière pseudo-aléatoire selon une loi uniforme dans un intervalle de valeurs réalistes, en utilisant les fonctions Python associées au sein d'une méthode generer_obstacle(). Elles permettront de manière simplifiée de simuler l'évolution de la zone maritime dans laquelle transite le navire avec l'actualisation des positions des navires et des éventuelles menaces asymétriques. Cette actualisation des positions est implémentée dans *Environnement_simulation* sous la forme de méthodes décrites dans la suite.

4.2.1.3.1. Implémentation du mouvement des navires et de la trajectoire d'une menace asymétrique

Le mouvement des navires est modélisé numériquement sous la forme d'une méthode, mouvement_navire() permettant d'actualiser la position d'un navire au cours du temps en

fonction des données radars. Chaque navire est supposé évoluer en mouvement rectiligne uniforme, son cap et sa vitesse ne varient pas au cours de l'observation de son mouvement entre deux positions. Ainsi au cours du temps la position d'un navire actualisée est donnée $\forall t \in [0;T]$ et $\forall t_0 \in [0;T]$ avec $t \geq t_0$:

$$\begin{cases} x_{navire}(t) &= x_{navire}(t_0) + V_{navire}(t_0)\cos\left(\frac{\pi}{180}\psi\right)t \\ y_{navire}(t) &= y_{navire}(t_0) + V_{navire}(t_0)\sin\left(\frac{\pi}{180}\psi\right)t \end{cases} \quad (1)$$

Les coordonnées de chaque navire, exprimées en mètres seront ramenées en angles dont le calcul est détaillé en ANNEXE 1 : Représentation des données cartographiques.

La trajectoire d'une menace asymétrique sera implémentée au sein d'une méthode mouvement_menace() permettant d'actualiser la position d'une menace au cours du temps. Chaque menace est supposée évoluer en mouvement rectiligne uniforme. Ainsi au cours du temps la position d'une menace actualisée par mouvement_menace() est donnée par $\forall t \in [0;T]$ et $\forall t_0 \in [0;T]$ avec $t \geq t_0$:

$$\begin{cases} y_{menace}(t) &= y_{menace}(t_0) + V_{menace}(\mathrm{t}_0)\sin\left(\frac{\pi}{180}\psi\right)t \\ x_{menace}(t) &= a y_{menace}(t_0) + b \end{cases} \quad (2)$$

Avec a le coefficient directeur de la droite affine entre la position de la menace et du navire à l'instant t et b l'ordonnée à l'origine de la droite affine. Les coordonnées exprimées en mètres seront ramenées en angles.

4.2.2. Implémentation du module connaissance situationnelle

Le module **connaissance situationnelle** permet de filtrer les données recueillies par l'environnement de simulation. Dans ce premier prototype, seules la base de données météorologique et les données radars simulées devront être filtrées. Les fonctionnalités du module **connaissance situationnelle** vont être ajoutées sous forme de méthodes au sein de la classe *Environnement_simulation*. Quatre types de méthodes vont être ajoutées (cf : Figure 4-2) :

- Une ou plusieurs méthodes permettant d'obtenir les données météorologiques adaptées à la zone maritime, au jour et à l'heure de la mission. Deux méthodes ont été implémentées pour le moment, grille_meteo() permettant de recueillir les données météo associées à la zone maritime de la mission pour un jour donné et grille_meteo_temps_different() permettant de recueillir les données météo associées à la zone maritime de la mission pour un jour et une heure donnée.
- Une méthode permettant d'obtenir les données radars dans un rayon donné autour du navire. La méthode donnees_radar() recueille la position en latitude et longitude, la vitesse et le cap des obstacles dans un rayon équivalent à la portée du radar de navigation du navire.
- Une autre méthode permettant de calculer l'orientation des obstacles mobiles par rapport à la frégate, appelée calculer_gisement().

- Une dernière méthode distance_euclidienne() qui en fonction de la position de la frégate et des obstacles calculera la distance euclidienne les séparant.

Les données étant simulées ou téléchargées, il n'y aura pas de problème de données manquantes. La reconnaissance de schémas déjà vus se fera donc systématiquement et les données filtrées seront directement introduites dans le module suivant qui sera en capacité de proposer à chaque fois une décision.

4.2.3. Implémentation du module simulateur mental et composante décisionnelle

L'implémentation du module **simulateur mental et composante décisionnelle** dont la composition est rappelée Figure 4-2 est l'étape la plus importante et difficile à mettre en œuvre ; l'environnement d'apprentissage par renforcement profond va être mis en place et le bien fondé des décisions prises par les agents va dépendre en grande partie de la qualité de l'environnement d'apprentissage. Au sein de ce module, deux modules doivent être implémentés comme illustrés Figure 4-2 : le premier est dédié à l'entraînement des agents et le second aux agents entraînés et utilisés dans l'architecture cognitive pour préconiser les actions. Comme énoncé au début du chapitre, notre architecture cognitive doit venir en aide sur la tâche de gestion de la route et de la gestion de la menace asymétrique. Le module **simulateur mental et composante décisionnelle** va être adapté à ces deux tâches où deux agents vont être entraînés séparément, comme illustré Figure 4-5. Dans le module apprentissage, deux environnements vont être créés, (1) *Environnement_gestion_trajectoire* dédié à l'agent en charge de la trajectoire et (2) *Environnement_gestion_menace* dédié à l'agent en charge de la menace asymétrique.

Dans la suite, les indicateurs utilisés par les agents pour prendre leurs décisions et les deux modules sont décrits.

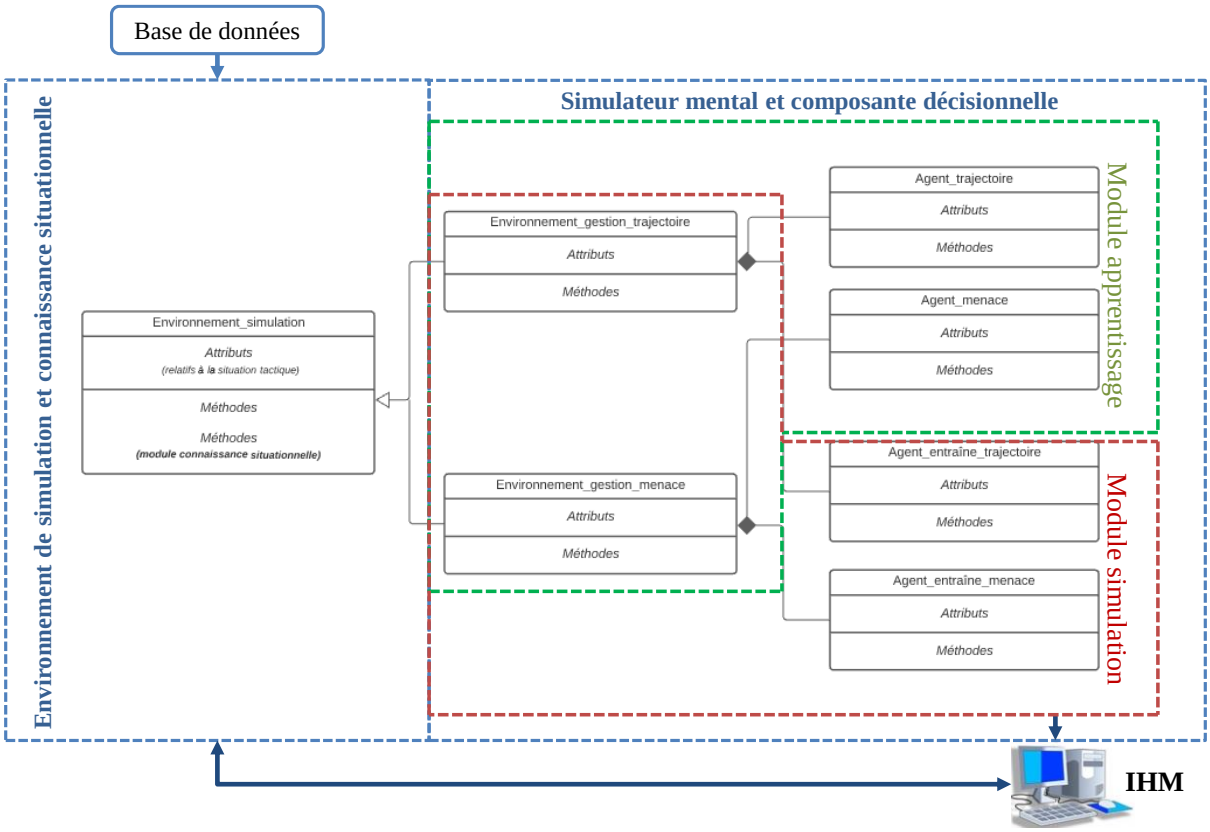

Figure 4-5 : Présentation du système d'aide à la décision pour le scénario envisagé sous forme d'un diagramme de classe UML

4.2.3.1. Implémentation des indicateurs

Chaque indicateur modélisé dans le Chapitre 3 sera implémenté sous la forme d'une méthode au sein des classes *Environnement_gestion_trajectoire* et *Environnement_gestion_menace*.

Le calcul de la vitesse maximale admissible sera implémenté au sein de la méthode vitesse_max_admissible() et fera intervenir l'algorithme Random Forest Regressor, décrit dans Chapitre 3, pour la prédiction de la vitesse en fonction des données météorologiques et des caractéristiques du navire. La modélisation a été réalisée à l'aide de l'algorithme *RandomForestRegressor*, issu de la librairie Python Scikit-learn [136], qui a été entraîné en 2,8 secondes sur 278 472 exemples, validé sur 68 921 exemples et testé sur 697 exemples avec une validation croisée à 10 blocs. Les hyperparamètres par défaut ont été utilisés sauf *n_estimators*, le nombre d'arbres composant la forêt qui était de 10 et *max_depth*, la profondeur maximum de chaque arbre qui était de 5. Ces deux hyperparamètres ont été déterminés à l'aide de la fonction GridSearchCV() de la bibliothèque *model_selection* de Scikit-learn, permettant sélectionner les meilleurs valeurs d'hyperparamètres. L'erreur moyenne absolue obtenue était d'environ 0,0028 nœuds, ce qui a permis de valider l'utilisation de ce modèle pour la prédiction de la vitesse maximale admissible réalisée en quelques millisecondes.

Le calcul de l'indicateur permettant à partir d'informations sur la menace, sur le navire et des caractéristiques d'une arme d'estimer la probabilité que cette arme dissuade la menace sera implémentée au sein de la méthode probabilité_effet_arme().

Le dernier indicateur, le risque de collision défini dans le Chapitre 3 faisant intervenir les périmètres de sécurité $r_{obstacle}$ et r_{navire} qui sont instanciés en tant qu'attributs de la classe *Environnement_simulation* sera implémentée au sein de la méthode risque_collision() dans *Environnement_gestion_trajectoire*.

Dans la suite, les modules apprentissage et simulation sont décrits.

4.2.3.2. Module apprentissage.

Le module d'apprentissage doit contenir les deux concepts de base de l'AR, rappelés Figure 4-5, à savoir :

- Un environnement, équivalent dans notre cas à la zone maritime sur laquelle évolue le navire,
- Un agent qui devra apprendre la politique souhaitée et évoluera dans l'environnement.

Comme énoncé en 4.2.3, le module d'apprentissage est composé de deux environnements avec des spécificités différentes, et d'un agent par environnement. L'implémentation générale de l'environnement d'un agent est tout d'abord explicitée dans la partie suivante puis les spécifiés de chacun des deux environnements seront introduites.

4.2.3.2.1. Implémentation de l'environnement d'un agent

L'environnement d'un agent sera toujours implémenté sous la forme d'une nouvelle classe *Environnement_agent* qui héritera de la classe *Environnement_simulation*, tel qu'illustré Figure 4-2. Les attributs et les méthodes instanciés au sein de la classe *Environnement_agent* suivront le formalisme OpenAI Gym [137], une librairie Python pour développer des environnements d'apprentissage par renforcement, basé sur la POO. OpenAI Gym ne fait aucune hypothèse sur la structure que doit avoir l'agent et elle est compatible avec toute bibliothèque de calcul numérique, telle que TensorFlow [138]. L'interface centrale de Gym est une classe *Env* qui représente l'environnement de l'agent, équivalente ici à *Environnement_agent* et qui doit contenir les éléments suivants, illustrés Figure 4-2 :

- Deux attributs de classe, action_space, l'espace des actions possibles de l'agent et observation_space, l'espace des états possibles. Nos agents assurant certaines fonctionnalités d'un navire, l'espace des états et des actions possibles doivent être en accord avec (1) ce que le navire est en capacité d'observer au sein de son environnement (coordonnées géographiques et vitesse des navires environnant) et (2) les actions que peut entreprendre un navire (changement de cap, de vitesse, utilisation d'armes).
- Une méthode reset() qui doit retourner une valeur comprise dans l'espace des états possibles (observation_space). Lors de l'apprentissage, cette fonction réinitialise

l'environnement à chaque épisode. Lors de la réinitialisation l'agent est repositionné au point de départ et une nouvelle situation tactique est générée avec de nouvelles données générées pseudo-aléatoirement. En réinitialisant à chaque fois une situation tactique différente la fonction reset() créée de l'aléa pour que l'agent puisse être robuste en situation inédite.

- Une méthode step(action), qui est la fonction de réponse de l'environnement. L'agent fournit l'action comprise dans l'espace des actions possibles (action_space) qu'il veut entreprendre, et l'environnement renvoie l'état dans lequel cette action l'a amené ainsi que la récompense immédiate accordée. La valeur retournée par cette méthode est un 4-uplet, dans l'ordre suivant :
    - *state*, le nouvel état de l'agent, n-uplet de valeurs comprises dans l'espace des états possibles.
    - *reward*, la récompense immédiate de l'agent suite à son action prise.
    - *done*, un booléen dont la valeur est TRUE si l'agent atteint le but souhaité ou est dans un état non souhaité, ou FALSE sinon.
    - *info*, un dictionnaire qui peut être utilisé pour la correction de bugs.

Dans la suite, l'implémentation des deux environnements d'apprentissage est décrite, à savoir *Environnement_gestion_trajectoire* pour apprendre à un agent à préconiser des décisions pour atteindre une zone d'arrivée le plus rapidement possible tout en évitant les obstacles statiques et mobiles, et *Environnement_gestion_menace* pour apprendre à un autre agent à préconiser des décisions sous la forme d'armes non létales à mettre en œuvre pour avoir le plus de chance de dissuader une ou plusieurs menaces asymétriques.

4.2.3.2.2. Environnement_gestion_trajectoire

Pour cette première application, la frégate étant supposée avancer à sa vitesse maximale admissible, l'agent ne proposera que des angles de caps à l'organe de commandement. L'étape suivante, comme évoqué précédemment, est de paramétrer *Environnement_gestion_trajectoire* en déterminant l'espace des actions (action_space), des états (observation_space), et en spécifiant les méthodes reset() et step(action). La Figure 4-6 résume les états, les actions possibles de l'agent et les récompenses mises en place au sein de step(action) pour que l'agent atteigne la zone d'arrivée le plus rapidement possible tout en évitant les obstacles.

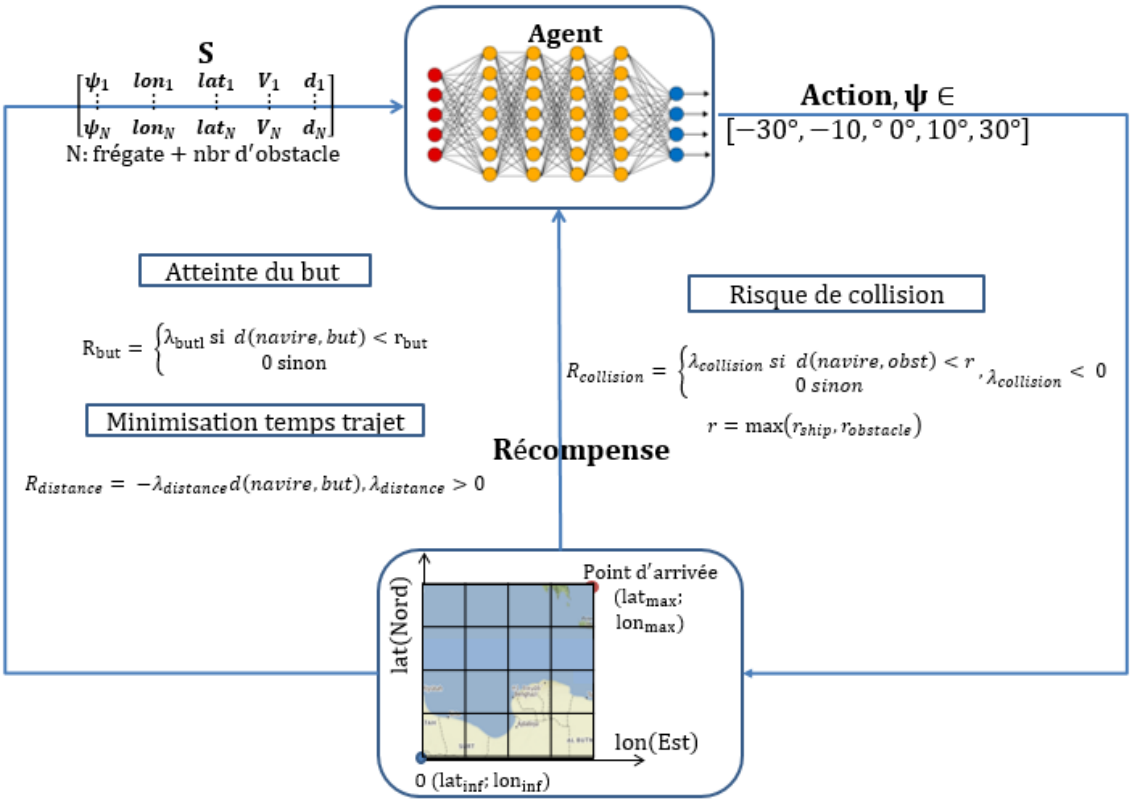

Figure 4-6 : Configuration de l'apprentissage de l'agent en charge de la gestion de la trajectoire

- **L'espace des actions (action_space)**

L'espace des actions de l'agent est équivalent aux décisions que proposera l'agent, à savoir une valeur d'angle de caps. Pour simplifier le problème dans un premier temps, l'espace d'actions est défini par un ensemble de cinq actions discrètes :

$$A = [-30°, -10°, 0°, 10°, 30°] \quad (3)$$

- **L'espace des états (observation_space)**

L'espace des états de l'agent va inclure les informations sur l'agent lui-même et sur ce qu'il est en capacité d'observer dans son environnement, issues des données radars simulées. Dans le cas de la gestion de la trajectoire, le vecteur d'état de l'agent sera composé en première ligne des informations relatives à l'état de la frégate avec son cap (ψ_1), ses coordonnées (lat_1, lon_1), sa vitesse maximale admissible (V_1) et la distance euclidienne (d_1) entre la frégate et la zone d'arrivée calculée à l'aide de la méthode distance_euclidienne(). Les autres lignes du vecteur d'état de l'agent représentent les informations sur les obstacles mobiles et statiques que peut recueillir la frégate à l'aide des données radars simulées. Ces informations sont le cap et la vitesses des obstacles, qui seront nuls si l'obstacle est statique, les coordonnées (latitude, longitude), et la distance euclidienne entre l'obstacle et la frégate calculée avec distance_euclidienne(). Le vecteur d'état de l'agent noté S sera de la forme :

$$S = \begin{bmatrix} \psi_1 & lat_1 & lon_1 & V_1 & d_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \psi_N & lat_N & lon_N & V_N & d_N \end{bmatrix}, \text{N : frégate + nbre d'obstacles} \quad (4)$$

Pour ce premier cas le nombre maximal d'obstacles statiques et mobiles sur la grille sera de 60.

- **Les méthodes reset() et step(action)**

Dans le cas du module trajectoire, le but de l'agent est d'atteindre une zone d'arrivée à partir d'un point de départ et d'une situation tactique. La méthode reset() permet de réinitialiser l'environnement à chaque épisode de l'apprentissage en sélectionnant pseudo-aléatoirement une zone maritime. Pour faciliter l'apprentissage de l'agent, les coordonnées de la zone maritime seront transformées en coordonnées relatives et la zone maritime sera toujours définie par $[0\,;\mathrm{x_{lon}}] \times [0\,;\mathrm{y_{lat}}]$. Sur cette zone maritime les données météo seront mises à jour grâce à la base de données créée ainsi que les méthodes grille_meteo() et grille_meteo_temps_different() ; et le point de départ, le nombre d'obstacles mobiles et statiques ainsi que leur positionnement seront générés pseudo-aléatoirement. La vitesse maximale admissible de la frégate sera réinitialisée à zéro et le cap de celle-ci sera choisi pseudo-aléatoirement entre 0 et 180°. En sortie de la méthode reset() l'état initialisé S de l'agent sera obtenu.

La méthode step(action), quant à elle, doit retourner un 4-uplet avec (1) *state*, le nouvel état de l'agent suite à l'action de cap qu'il aura choisie ; (2) *reward*, la récompense immédiate de l'agent suite à son action prise ; (3) *done*, un booléen dont la valeur sera TRUE si l'épisode se termine ou FALSE sinon et (4) *info*, un dictionnaire qui peut être utilisé pour la correction de bugs.

Pour (1), selon l'espace d'états et d'actions défini, la transition d'état de la frégate avec une action donnée $a \in A$, en suivant sa cinématique du mouvement est, $\forall\, t, t' \in [0; T]$, avec $t' \geq t$:

$$\begin{cases} \psi' & = & \psi + a \\ x_{fregate}(t') & = & x_{frégate}(t) + V(t)\cos(\psi')\, t_{trajectoire}\,, \\ y_{fregate}(t') & = & y_{frégate}(t) + V(t)\sin(\psi')\, t_{trajectoire} \end{cases} \quad (5)$$

Avec V la vitesse maximale admissible de la frégate.

La transition d'état pour les obstacles mobiles sera identique à l'équation (5) mais les angles de cap et les vitesses seront générés pseudo-aléatoirement à chaque pas de temps.

(2) et (3) sont décrits ci-dessous et (4) est un dictionnaire vide.

- ***Les fonctions récompenses***

En apprentissage par renforcement, l'élaboration des fonctions récompenses est une étape importante car ces dernières permettent à l'agent d'apprendre le comportement souhaité. Plus les fonctions récompenses seront adaptées au problème, et plus l'agent apprendra bien, c'est-à-

dire qu'il convergera rapidement lors de son apprentissage et ne restera pas « piégé » dans un minimum local. Pour la gestion de la trajectoire, les fonctions récompenses sont créées dans le but d'apprendre à l'agent à rallier la zone d'arrivée le plus rapidement possible tout en évitant les collisions. Les fonctions de récompenses suivantes pour la gestion de la route vont être implémentées :

Une fonction récompense relative à l'atteinte d'une zone. Cette fonction permettra de faire apprendre à l'agent à atteindre la zone de la mission. La zone d'arrivée sera représentée par un cercle dont le rayon représentera la zone à atteindre et le centre le point d'arrivée. L'agent recevra une forte récompense positive lorsqu'il atteindra cette zone, et la récompense sera de la forme :

$$\mathrm{R_{but}} = \begin{cases} \lambda_{\mathrm{but}} \text{ si } \mathrm{d(navire, but)} < \mathrm{r_{but}} \\ 0 \text{ sinon} \end{cases} \quad (6)$$

Où r_{but} sera fixé à 20 km autour du point d'arrivée et λ_{but} sera fortement positif et ajusté lors des différents apprentissages réalisés. r_{but} permet d'ajuster la taille de la zone d'arrivée, plus le rayon sera grand et plus il sera facile à l'agent d'apprendre où se situe la zone d'arrivée. La taille de la zone d'arrivée est à ajuster en fonction de la mission. Si par exemple le navire doit rallier un point GPS précis r_{but} devra être de l'ordre de quelques mètres. Ici 20 km a été choisi afin de diminuer le temps d'entraînement de l'agent. Si la condition de l'équation (6) est vérifiée, l'agent aura atteint la zone d'arrivée, le booléen *done* sera égal à TRUE mettant fin à l'épisode et si l'entrainement n'est pas terminé, un nouvel épisode sera réinitialisé à l'aide la méthode reset().

Une fonction récompense en rapport avec le risque de collision. Lors d'une mission, l'agent ne doit pas rentrer en collision avec les obstacles de son environnement. Pour lui apprendre à éviter les collisions une récompense négative lui sera renvoyée si le risque de collision tel que défini Chapitre 3 est de 100% :

$$\mathrm{R_{collision}} = \begin{cases} \lambda_{\mathrm{collision}} & \text{si} \quad \mathrm{d(navire, obstacle)} < \max(r_{navire}, r_{obstacle}) \\ 0 & \text{sinon} \end{cases} \quad (7)$$

Avec $\lambda_{collision}$ une constante négative pour punir l'agent et lui suggérer de ne pas prendre la même action si un scénario similaire se produit dans les futurs épisodes d'entraînement et $\mathrm{d(navire, obstacle)} < \max(r_{navire}, r_{obstacle})$, calculé avec la méthode risque_collision(). r_{navire} est égal à environ 430 m car la longueur d'une frégate est supposée être de 120 m et $r_{obstacle}$ à environ 200 m quel que soit le type d'obstacles. De même que précédemment $\lambda_{collision}$ sera fortement négatif et ajusté lors des différents apprentissages et si la condition de l'équation (7) est vérifiée, le booléen *done* sera égal à TRUE et l'agent recommencera un nouvel épisode.

En plus d'une fonction récompense relative aux risques de collision, une fonction pour apprendre à l'agent à évoluer dans une zone maritime bornée est implémentée :

$$R_{sortie} = \begin{cases} \lambda_{sortie} & \text{si} \quad lat_{agent} < 0 \text{ ou} > N_x \text{ ou si } lon_{agent} < 0 \text{ ou} > N_y \\ 0 & \text{sinon} \end{cases} \quad (8)$$

Si la condition de l'équation (8) est vérifiée l'agent aura pris une action l'amenant en dehors de la zone maritime dans laquelle il doit accomplir la mission, une récompense fortement négative λ_{sortie} lui sera renvoyée afin qu'il apprenne à ne plus sortir des limites de la zone. Le booléen *done* sera alors égal à TRUE et l'agent recommencera un nouvel épisode.

La dernière fonction récompense, équation (9), sera celle relative à l'objectif d'atteindre la zone d'arrivée le plus rapidement possible :

$$R_{distance} = -\lambda_{distance} d(navire, but) \quad (9)$$

Avec $\lambda_{distance}$ une constante négative pour punir l'agent proportionnellement à la distance qui le sépare de la zone d'arrivée. Cette récompense va inciter l'agent à choisir les zones où sa vitesse maximale admissible sera la plus élevée afin d'arriver le plus rapidement possible à la zone d'arrivée et de minimiser son nombre de récompenses négatives obtenues. Si les conditions des équations (6), (7) et (8) ne sont pas vérifiées alors l'agent recevra la récompense de l'équation (9), le booléen *done* sera égal à FALSE et l'épisode continuera tant que la condition de l'équation (6), (7) ou (8) n'est pas vérifiée.

4.2.3.2.3. Environnement_gestion_menace

Concernant la gestion de menaces asymétriques lors de la phase de dissuasion, les décisions proposées à l'organe de commandement seront des armes non-létales. Il est supposé que (1) toutes les menaces ne réagiront pas aux armes de dissuasion et (2) elles s'introduiront à chaque fois dans le périmètre de la distance de neutralisation. Comme précédemment *Environnement_gestion_menace* va être paramétré en spécifiant l'espace des actions (action_space), des états (observation_space), et les méthodes reset() et step(action). La Figure 4-7 résume les états, les actions possibles de l'agent et les récompenses mises en place au sein de step (action) pour que l'agent préconise la meilleure arme non-létale en fonction de la menace.

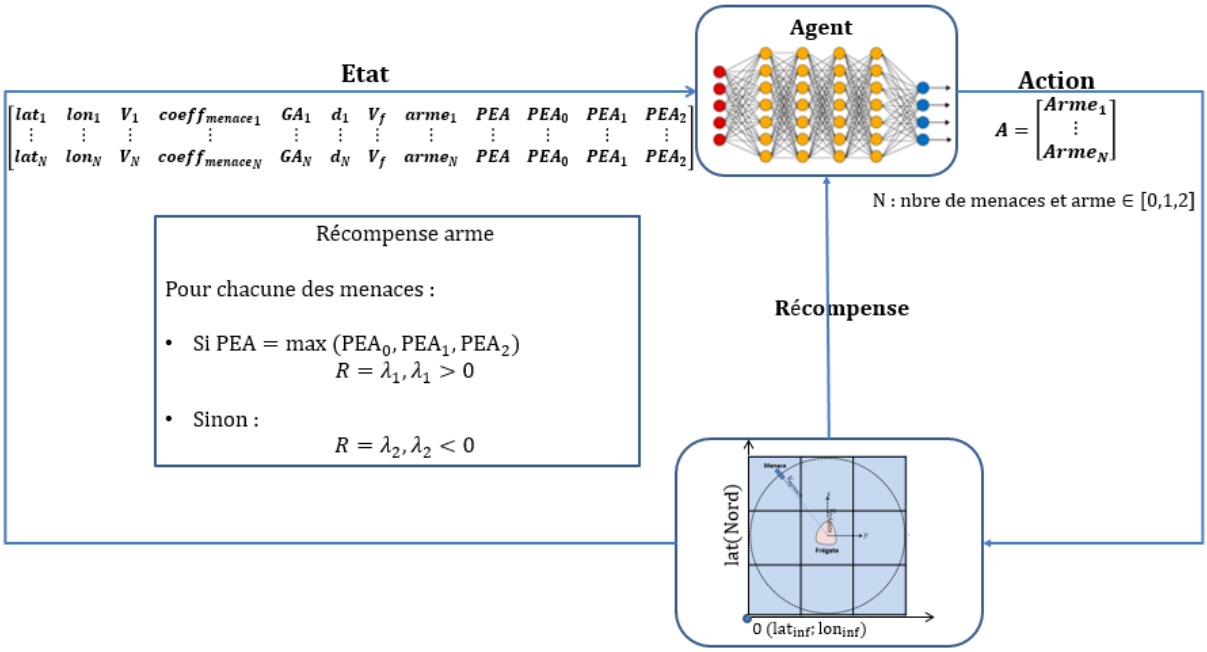

Figure 4-7 : Configuration de l'apprentissage de l'agent en charge de la gestion de la menace

- **L'espace des actions (action_space)**

L'espace des actions de l'agent est équivalent aux décisions que proposera l'agent, à savoir ici une arme non létale par menace. Pour ce type de mission l'agent pourra préconiser trois armes différentes sans limite d'utilisation, qui seront dénommées par 0 : tir de semonce, 1 : canon sonore et 2 : laser d'aveuglement. L'espace d'actions est alors défini par le vecteur suivant :

$$A = \begin{bmatrix} Arme_1 \\ \vdots \\ Arme_N \end{bmatrix} \quad (10)$$

Où N est le nombre de menaces, ici N peut varier de 1 à 4 et Arme ∈ [0,1,2] désigne l'arme préconisée par l'agent.

- **L'espace des états (observation_space)**

L'espace des états de l'agent va inclure des informations sur les menaces observées et sur lui-même. Le vecteur d'état sera composé de 1 à 4 lignes, où chaque ligne représente une menace différente avec possibilité pour l'agent d'observer jusqu'à quatre menaces simultanées. Chaque ligne est d'abord composée d'informations relatives aux menaces issues des données radars simulées avec leurs coordonnées(lat, lon), leur vitesse (V), leur coefficient de menace $(coeff_{menace})$, leur gisement par rapport à la frégate (GA) calculé à l'aide de la méthode calculer_gisement() et la distance euclidienne (d) entre la menace et la frégate calculée avec distance_euclidienne(). Des informations propres à la frégate se trouvent à la suite avec la vitesse de la frégate (V_f), l'arme choisie par l'agent $(arme)$, et les probabilités d'effet des armes $(PEA, PEA_0, PEA_1, PEA_2)$, où PEA représente la probabilité d'effet de l'arme choisie par l'agent et PEA_0, PEA_1, PEA_2 désignent respectivement les probabilités d'effets des armes

0,1,2 par rapport aux caractéristiques de la menace et de la frégate. Les probabilités d'effets des armes sont calculées à l'aide de la méthode probabilité_effet_arme().

Ainsi, le vecteur d'états de l'agent noté S sera de la forme :

$$S = \begin{bmatrix} lat_1 & lon_1 & V_1 & coeff_{menace_1} & GA_1 & d_1 & V_f & arme_1 & PEA & PEA_0 & PEA_1 & PEA_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ lat_N & lon_N & V_N & coeff_{menace_N} & GA_N & d_N & V_f & arme_N & PEA & PEA_0 & PEA_1 & PEA_2 \end{bmatrix} \quad (11)$$

Avec N, le nombre de menaces $\in [1, 2, 3, 4]$

- **Les méthodes reset() et step(action)**

Dans le cas du module menace, le but de l'agent est de préconiser lors de la phase de dissuasion, la meilleure arme non-létale à utiliser contre chacune des menaces asymétriques présente. Pour rappel, la méthode reset() permet de réinitialiser à chaque épisode l'environnement lors de l'apprentissage. La méthode reset() va générer pseudo-aléatoirement toutes les variables suivantes : une zone maritime, la position de la frégate, sa vitesse et son cap. Toujours au sein de reset(), le nombre de menaces sera ensuite choisi entre un et quatre et pour chacune des menaces, leur position initiale qui se trouvera sur un cercle de rayon $distance\_dissuasion$ autour de la frégate, leur vitesse comprise entre 30 et 40 nœuds et leur coefficient de menaces compris entre 0.6 et 0.9 seront générés. Les autres informations comprises dans l'état de l'agent telles que le gisement, les probabilités d'effet des armes seront directement calculées par les méthodes implémentées. En sortie de la méthode reset() l'état initialisé S de l'agent sera obtenu.

La méthode step(action), quant à elle, doit toujours retourner un 4-uplet avec (1) *state*, le nouvel état de l'agent suite aux armes choisies ; (2) *reward*, la récompense immédiate de l'agent suite à son action prise ; (3) *done*, un booléen dont la valeur sera TRUE si l'environnement atteint le but souhaité ou FALSE sinon et (4) *info*, un dictionnaire qui peut être utilisé pour la correction de bugs.

Pour *state*, suite à l'action donnée $a \in A$, les positions des menaces et de la frégate vont être mises à jour toutes les 10 secondes ($\mathrm{t_{menace}}$). Les positions des menaces vont être actualisées à l'aide de la méthode mouvement_menace() et celles de la frégate à l'aide de mouvement_navire(). Le nouveau gisement (GA), la nouvelle distance euclidienne (d) et les nouvelles probabilités (PEA, PEA_0, PEA_1, PEA_2) seront calculées en conséquence.

reward et *done* sont décrits ci-dessous et *info* est un dictionnaire vide.

- ***La fonction récompense***

Lors de la survenue de menaces asymétriques le navire n'a plus le temps de manœuvrer pour l'éviter, la seule solution est d'utiliser les armes pour se défendre. Dans cette configuration le but de l'agent va être d'aider l'organe de commandement à choisir la meilleure arme non-létale en fonction de la menace et de certaines caractéristiques de la frégate. La fonction récompense aura la forme suivante :

Pour chacune des menaces, la fonction récompense suivante sera renvoyée à l'agent :

$$R_{arme} = \begin{cases} \lambda_{arme} & \text{si} \quad PEA = \max(PEA_0, PEA_1, PEA_2) \\ -\lambda_{arme} & \text{sinon} \end{cases} \quad (12)$$

Où λ_{arme} est une constante positive pour récompenser l'agent d'avoir choisi la meilleure arme et l'inciter à la sélectionner de nouveau si un scénario similaire se présente. Le booléen *done* est initialisé à FALSE à chaque épisode de l'apprentissage. Il passera à l'état TRUE et l'épisode sera réinitialisé à l'aide de la méthode reset() dans deux cas de figures : (1) si l'équation (12) n'est pas vérifiée, ce qui signifiera que l'agent n'aura pas choisi la meilleure arme, et (2) si toutes les menaces se sont introduites dans le périmètre de neutralisation mettant fin à la phase de dissuasion.

Maintenant que les environnements du module apprentissage ont été paramétrés, les agents vont pouvoir être entraînés dans leurs environnements respectifs.

4.2.3.3. Implémentation des agents au sein des environnements

Une fois les environnements définis, les agents dédiés aux tâches (navigation et réponse à une menace) vont être implémentés sous forme de classe en relation avec leur environnement respectif, tel qu'illustré dans la Figure 4-5. Chaque agent doit apprendre une politique ou stratégie d'actions adaptée à une situation. Comme évoqué dans le Chapitre 2, il existe plusieurs types d'algorithmes ou méthodes d'apprentissage de la politique optimale. Les méthodes sans modèle étant actuellement les plus utilisées pour résoudre des problèmes du monde réel, nos choix algorithmiques vont se tourner vers l'une de ces méthodes. L'implémentation de l'une d'entre elles a été faite avec la librairie Python Stable-baselines [139]. Cette librairie a été choisie car (1) elle offre un ensemble d'algorithmes d'apprentissage par renforcement profond regroupant les méthodes sans modèle les plus utilisées aujourd'hui, telles que le Deep-Q-Network (DQN) [117] , ou la Proximal Policy Optimisation (PPO) [121] et (2) elle utilise une interface commune, basée sur le formalisme d'OpenAI, ce qui va nous permettre de facilement interfacer nos classes *Environnement_gestion_trajectoire* et *Environnement_gestion_menace* à l'une de ces méthodes.

Le choix de l'algorithme d'apprentissage va dépendre du problème à résoudre et aucune méthode n'existe pour déterminer le plus adapté. Cependant deux critères de sélection permettent de faire des choix d'algorithmes. Le premier repose sur le type des actions possibles : certains algorithmes ne sont adaptés qu'à un type d'actions (continue ou discrète) comme par exemple le DQN qui ne gère que les actions discrètes, ou le SAC [140] qui est limité aux actions continues. Le deuxième concerne la parallélisation de l'entraînement : certains algorithmes sont plus rapides à entraîner que d'autres et le choix va dépendre des machines sur lesquelles ils sont entraînés. **Pour ce premier prototype, les actions des deux agents seront discrètes et l'entraînement sera parallélisé. Le choix s'est donc orienté vers la méthode PPO, car elle est bien adaptée aux actions discrètes et aux entraînements parallèles.**

Une fois le choix de l'algorithme d'apprentissage effectué, le réseau de neurones de chaque agent doit être choisi et il faut régler les hyperparamètres ou utiliser ceux par défaut selon le problème.

Stable-baselines fournit un ensemble de réseaux de politiques par défaut qui peuvent être utilisés avec la plupart des espaces d'actions. Deux familles sont proposées, les *CnnPolicies* qui sont adaptées pour des images en entrée et les *MlpPolicies* pour les autres types d'entrée. Pour notre prototype, la politique *MlpPolicy* a été choisie, composée par défaut d'un perceptron multicouche de deux couches de 64 neurones pour les deux agents.
Pour le réglage des hyperparamètres, chaque agent doit être entraîné plusieurs fois avec différentes valeurs d'hyperparamètres pour constater comment ils affectent les performances. Il n'y a aucun moyen à priori de savoir si une valeur plus élevée ou plus faible d'un paramètre donné améliorera les récompenses totales. Pour obtenir un bon agent, des entraînements multiples devront être réalisés pour suivre les expériences, les données et tout ce qui est associé à l'entraînement des modèles.

Lorsque l'algorithme d'apprentissage, le réseau de politique et les premiers hyperparamètres ont été choisis, les agents sont connectés avec leur environnement. Avant d'exécuter un entraînement le nombre d'étapes sur lequel l'agent s'entraînera doit être défini.

Algorithme : PPO2

1 : Initialiser l'environnement d'apprentissage de l'agent
    env = Navigation_env ()
2 : Initialiser PPO2
3 : Initialiser la politique π_{old}, '*MlpPolicy*' de Stable-baselines
4 : Initialiser les hyperparamètres pour obtenir :
    model = PPO2('MlpPolicy', env, hyperparametres)
5 : Lancer l'apprentissage à l'aide de learn() : model.learn()
6 : **Pour** episode = 1 à N **faire** :
7 : Initialiser s_0, l'état initial de l'agent avec env.reset()
8 : **Pour** itération = 1 à I **faire** :
9 : **Si** done = False **faire** :
10 : Exécuter la politique π_{old} pour T étapes et enregistrer $\{s_t, a_t, r_t\}$:
11 : Choix d'une action a_t par l'agent en suivant sa politique
12 : Mise à jour de l'état s_t, de la récompense r_t et de *done* par step(action)
13 : step(action) : mise à jour des variables de env, de s_t, de r_t et de done
14 : Estimation de la fonction Avantage pour les T étapes
15 : Optimisation de la fonction politique perte ou erreur par une méthode de gradient en utilisant la fonction Avantage
16 : Mise à jour de la politique π
17 : $\pi_{old} = \pi$
18 : **Sinon :**
19 : Sortir de la boucle et commencer un nouvel épisode
20 : **Fin pour**
21 : **Fin pour**
22 : Fin apprentissage
23 : Sauvegarder la politique apprise avec la méthode save()

Figure 4-8 : Pseudocode pour l'entraînement de l'agent en charge de la gestion de la trajectoire à l'aide de Stable-baselines

Ce nombre d'étapes est subjectif et doit être assez élevé pour que l'agent ait le temps d'apprendre, de généraliser et que l'entraînement converge. L'agent est ensuite entraîné à l'aide de la méthode learn(). La Figure 4-8, illustre le déroulement de l'apprentissage de l'agent en charge de la gestion de la route à l'aide de l'algorithme PPO2 de Stable-baselines.

Après l'entraînement, une phase de validation doit être menée afin de s'assurer que les agents ont le comportement souhaité et sont capables de généraliser ce qu'ils ont appris sur des données d'entrée inédites. Différentes méthodes existent pour évaluer la politique apprise d'un agent. L'une des plus utilisée est de tester l'agent sur n épisodes inédits et d'observer la récompense moyenne. Stable-baselines fournit plusieurs fonctions pour réaliser la phase de validation de l'agent. Pour évaluer les performances en généralisation des agents, nous avons utilisé la fonction evaluate_policy() de Stable-baselines. A partir de l'agent entraîné, de son environnement et du nombre d'épisodes sur lequel on souhaite tester l'agent, la fonction evaluate_policy() renverra la récompense moyenne obtenue. En fonction de cette valeur, il est possible de savoir si un agent a bien appris ce qui lui était demandé. Si les performances en généralisation des agents sont satisfaisantes, les politiques apprises sont sauvegardées à l'aide la méthode save(). Dans la suite l'entrainement et la validation des deux agents sont explicités.

4.2.3.4. Entraînement et validation des deux agents

Dans cette partie, les hyperparamètres et les conditions d'entraînement des deux agents sont explicités sous la forme de deux tableaux récapitulant les informations importantes. Les variables γ et α ont été introduites Chapitre 2 et pour rappel une étape est le passage d'un état à l'autre de l'agent.

Tableau 4-2 : Valeurs des principaux hyperparamètres du module trajectoire

Hyperparamètres PPO2 module trajectoire	
Variables	*Valeurs*
γ	0,99
α	0,00025
Nombre d'étapes d'apprentissage	3 000 000
Nombre de CPU	16

Tableau 4-3 : Valeurs des principaux hyperparamètres du module menace

Hyperparamètres PPO2 module menace	
Variables	*Valeurs*
γ	0,99
α	0,00025
Nombre d'étapes d'apprentissage	4 000 000
Nombre de CPU	16

Une fois l'entraînement des deux agents terminé, il faut s'assurer que l'entraînement a bien convergé et évaluer les performances en généralisation de chacun des agents à l'aide de la fonction evaluate_policy() de Stable-baselines.

Pour l'agent en charge de la gestion de la trajectoire, la Figure 4-9 représente l'évolution de la récompense par épisode en fonction du nombre d'étapes. Pour rappel, une étape est le passage de l'état d'un agent à l'autre et un épisode est constitué de toutes les étapes qu'effectue un agent entre son état initial et son état final. La somme des récompenses collectées dans un

seul épisode s'appelle la récompense cumulative totale et le but de l'entraînement d'un agent est de la maximiser. Sur la Figure 4-9, le tracé gris foncé représente la moyenne mobile exponentielle de la récompense par épisode tandis que la valeur réelle de la récompense par épisode est représentée par les barres grises claires. Cette figure nous permet de nous assurer que l'entraînement a bien convergé puisque la récompense par épisode est maximisée sur la fin de l'entraînement à partir de 2,5 millions d'étapes.

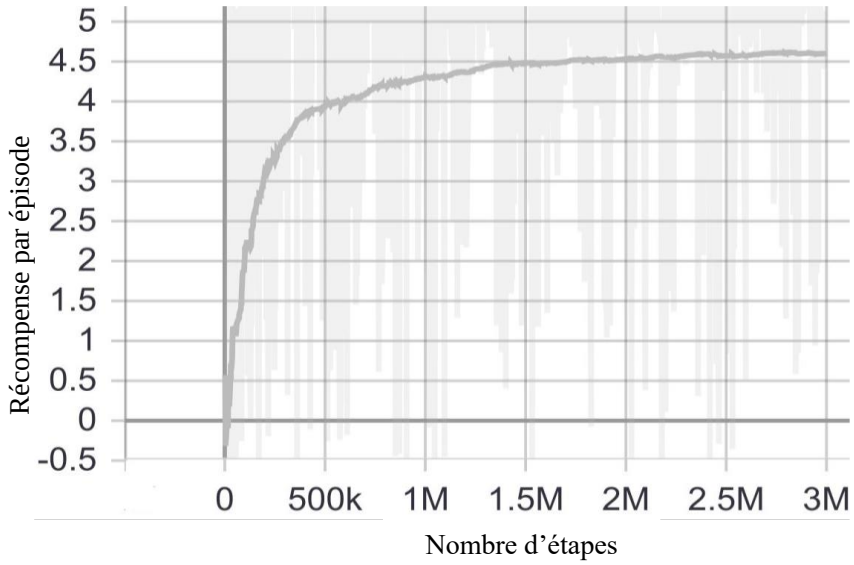

Figure 4-9 : Evolution de la récompense par épisode en fonction du nombre d'étapes pour la gestion de la route

L'agent a ensuite été évalué sur 500 épisodes différents générés pseudo-aléatoirement et a obtenu un succès moyen d'anticollision de 93 %. Les 7% d'échecs regroupent à la fois les collisions avec des obstacles statiques et mobiles et les sorties de la zone maritime bornée.

Concernant l'agent pour la gestion de la menace, la même démarche a été appliquée, la Figure 4-10 montre que l'entraînement à bien convergé puisque la récompense par épisode est maximisée sur la fin de l'entraînement à partir de 1,5 millions d'étapes.

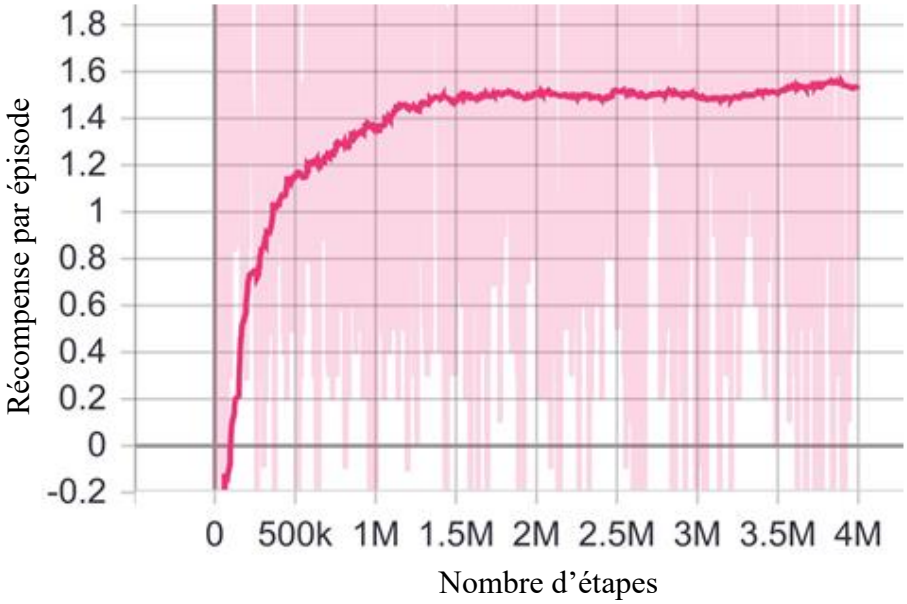

Figure 4-10 : Evolution de la récompense par épisode en fonction du nombre d'étapes pour la gestion de la menace

L'agent a aussi été évalué sur 500 épisodes différents générés pseudo-aléatoirement et a obtenu 87% de réussite, c'est-à-dire que dans 87% des cas la meilleure arme non-létale a été préconisée.

4.2.3.5. Module simulation

Une fois l'entraînement et la validation des deux agents terminés, les politiques apprises sont sauvegardées à l'aide de la méthode save() et sont téléchargées à l'aide de la méthode load() au sein du module simulation afin de préconiser des décisions lors de la conduite de la mission. De manière similaire que pour le module apprentissage, un module simulation regroupant les deux environnements et leur agent entraîné respectif est implémenté. Les environnements des agents sont identiques aux deux implémentés dans le module apprentissage, la seule différence concernera l'implémentation de la méthode reset(). Pour faire fonctionner l'architecture lors de la mission, les classes *Environnement_simulation, Environnement_gestion_trajectoire, Environnement_gestion_menace et Agent_entraine_trajectoire, Agent_entraine_menace* vont être instanciées. Lors de l'instanciation, la situation tactique ne sera plus générée pseudo-aléatoirement mais initialisée par l'utilisateur qui spécifiera les données d'entrée à *Environnement_simulation*. Ces données d'entrée seront introduites dans les deux nouvelles méthodes reset() des deux classes *Environnement_gestion_trajectoire, Environnement_gestion_menace*, qui retranscriront l'état initial au lieu de générer un état initial pseudo-aléatoire comme lors de la phase d'apprentissage. Il est supposé qu'au départ de la mission, il n'y aura aucune menace asymétrique. Dans cette première version, l'utilisateur ne peut choisir que la zone maritime et les points de départ et d'arrivée. Les autres données telles que les positions des obstacles sont encore générées pseudo-aléatoirement. Une fois la zone maritime spécifiée, les données météorologiques sont directement requêtées à partir de la base de données créée et après l'initialisation des méthodes reset(), l'architecture est lancée et des

décisions à la fois de cap et d'arme non-létale sont préconisées à l'utilisateur en fonction de l'évolution de la situation tactique jusqu'à la fin de la mission.

4.3. Implémentation de l'IHM

Après toutes ces étapes, l'architecture cognitive modulaire est prête à fonctionner de manière dynamique sur un scénario de mission [141]. Afin de faciliter l'utilisation de l'architecture, une IHM a été créée pour que l'utilisateur puisse solliciter quand il le souhaite le système, spécifier rapidement les données de la situation tactique et visualiser les décisions préconisées. Les spécifications de l'IHM ont été réalisées chez NAVAL GROUP dans le cadre d'un projet auquel cette thèse était rattachée et ne permet que de visualiser les décisions préconisées par l'agent entraîné pour la gestion de la route. Dans la suite, les spécifications de l'IHM sont décrites.

L'IHM est basée sur une solution à un seul écran, de préférence de grande taille, sur lequel est affiché un fond cartographique en 2D représentant la zone maritime d'évolution du navire. Les données initiales suivantes seront affichées au départ de la mission :

- Un fond cartographique,
- Un navire en vue 2D de dessus,
- Le temps courant,
- Les obstacles statiques naturels tels que les côtes ou les îles,
- Les données initiales du navire (i.e : agent entraîné) en accord avec les états et actions définis lors de la mise en place de l'apprentissage de l'agent. Ces données pourront être par exemple la position du navire en latitude et longitude, son cap, sa vitesse.
- Les données radars simulées permettant de recueillir la position (latitude, longitude) et la vitesse des mobiles.
- Une « heatmap » des données météorologiques.

La zone maritime de la mission sera bornée et définie par $[\mathrm{Lat}_{min}; \mathrm{Lat}_{max}] \times [\mathrm{Lon}_{min}; \mathrm{Lon}_{max}]$. L'IHM affichera de manière dynamique les sorties renvoyées par l'agent entraîné pour la gestion de la route et l'évolution de la situation tactique (données contenues dans le module **environnement de simulation)**. Des fonctionnalités d'interaction avec l'utilisateur ont aussi été ajoutées, l'utilisateur a la possibilité de mettre en pause l'architecture cognitive, de modifier les données d'entrée de celle-ci et de la relancer.

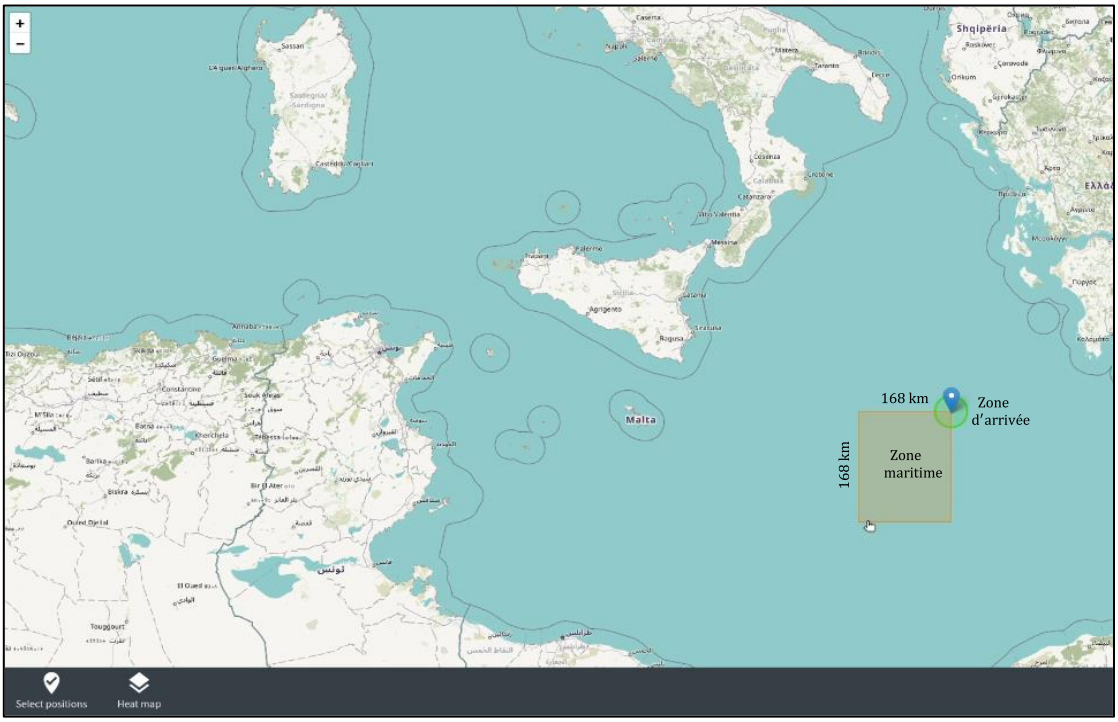

Figure 4-11 : Présentation de l'IHM – sélection zone d'arrivée

La Figure 4-11 représente l'aperçu de l'IHM avant le début de la mission lorsque l'utilisateur va devoir sélectionner les données d'entrée pour lancer l'architecture cognitive. Sa première tâche va être d'activer le bouton à gauche « Select positions » et de sélectionner en premier le point d'arrivée représenté en bleu. La zone d'arrivée de 20 km autour du point bleu va alors s'afficher automatiquement ainsi que la zone maritime de 168 km par 168 km dans laquelle se déroulera la mission.

Une fois cette première étape effectuée, l'utilisateur va choisir un point de départ au sein de la zone maritime toujours à l'aide de la fonction « Select positions », comme illustré Figure 4-12. Après la sélection du point de départ, la frégate s'affichera en position de départ et la portée du radar de navigation, ici de 50 km sera représentée par un cercle bleu autour de celle-ci.

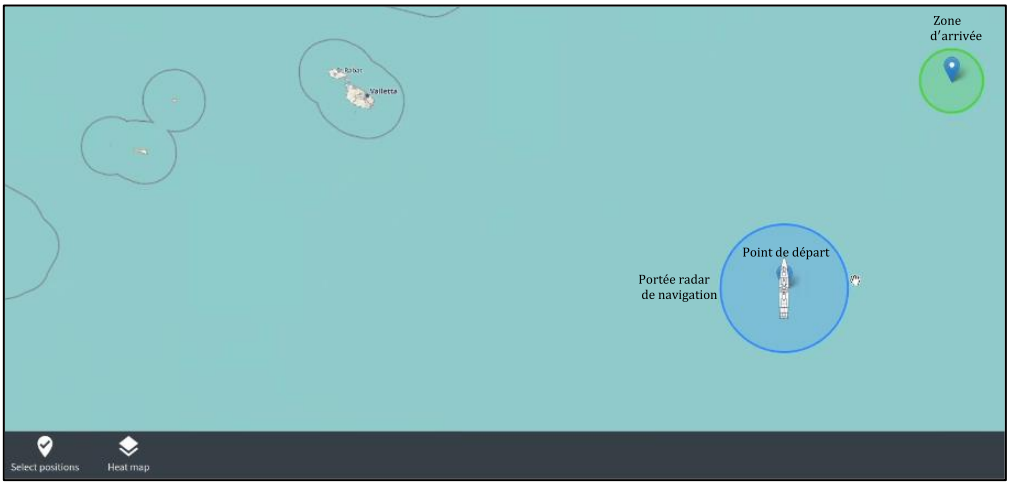

Figure 4-12 : Représentation de l'IHM-sélection du point de départ

Suite à la sélection du point d'arrivée et de départ, la position des obstacles ainsi que leurs données radars associées vont être générées pseudo-aléatoirement et la situation tactique va être représentée à l'utilisateur comme illustrée Figure 4-13. Les points de couleurs représentent les positions des différents obstacles. La situation tactique est initialisée et l'utilisateur va pouvoir utiliser différentes fonctions. Pour lancer le scénario, il devra activer la fonction « Play ». L'architecture cognitive va alors être lancée et l'agent commencera à préconiser des actions jusqu'à atteindre la zone d'arrivée. Si l'utilisateur souhaite voir le scénario en accéléré il pourra activer la fonction « Speed up » ou « Speed down » pour ralentir. Il aura aussi à sa disposition une timeline lui permettant de rejouer le scénario à partir d'un moment précis. Si la trajectoire préconisée par l'agent ne lui convient pas il pourra activer la fonction « Select cap » qui va lui permettre de mettre l'architecture cognitive en pause et de forcer l'agent à repartir d'une position avec un angle de cap donné.

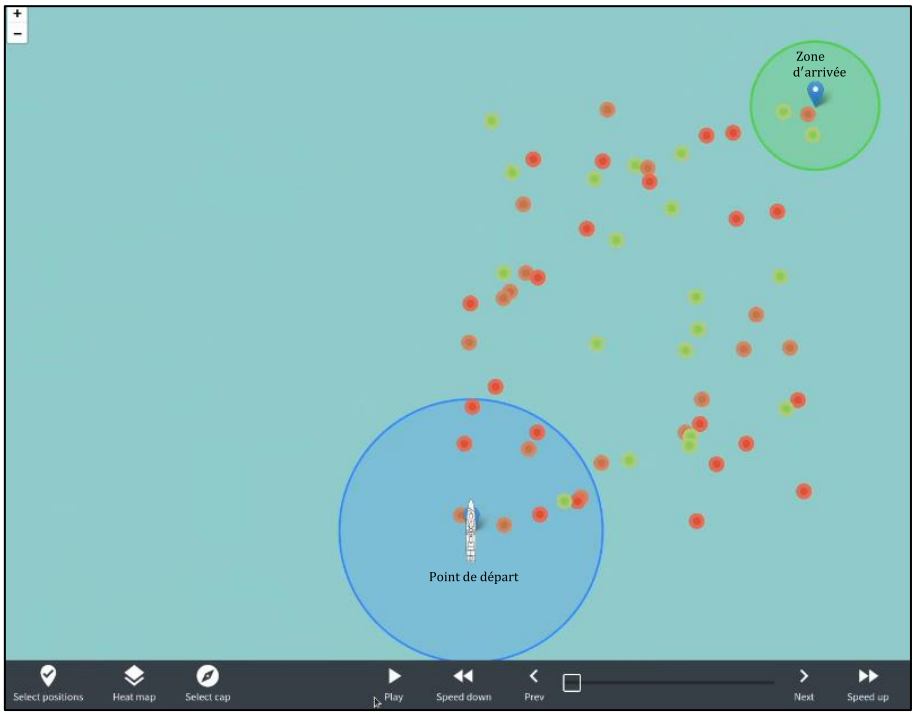

Figure 4-13 : Représentation de l'IHM - situation tactique initiale

La fonction « Play » activée, la frégate naviguera parmi les obstacles statiques et mobiles jusqu'à atteindre la zone d'arrivée. La Figure 4-14 montre une trajectoire préconisée par l'agent en bleu. Pour une meilleure visualisation il est supposé qu'entre deux positions discrètes le mouvement est rectiligne uniforme (cap et vitesse ne varient pas).

Figure 4-14 : Représentation de l'IHM - visualisation des préconisations de l'agent jusqu'à la zone d'arrivée

4.4. Conclusion

Dans ce chapitre, le prototypage de notre architecture cognitive d'aide à la décision a été détaillé. Les modules de l'architecture cognitive du Chapitre 3 sont implémentés sous forme de classes. Le module **environnement de simulation** est représenté par la classe *Environnement_simulation*. Les données non simulées contenues dans la base de données sont importées dans l'environnement de simulation et celles simulées sont directement générées au sein de celui-ci via des méthodes. Le module **connaissance situationnelle** est inclus dans *Environnement_simulation* sous forme de méthodes. Chacun des environnements du module **simulateur mental et composante décisionnelle** hérite de la classe *Environnement_simulation* et il est divisé en deux modules, un module apprentissage utilisé pour l'entrainement des agents et un module simulation permettant de solliciter les agents entraînés lors de la conduite d'une mission. Notre architecture cognitive a été prototypée sur un scénario de missions, où l'objectif principal était de rallier une zone d'arrivée le plus rapidement possible tout en gérant l'apparition d'éventuelles menaces asymétriques. Deux agents entraînés par apprentissage par renforcement profond pour la tâche de gestion de la route et de gestion des menaces ont été validés sur des scénarios jamais appris. Ainsi, l'architecture cognitive est capable de préconiser en temps réel des angles de cap pour que la frégate atteigne la zone d'arrivée le plus rapidement

possible couplés à des armes non létales si la frégate est confrontée à une ou plusieurs menaces asymétriques.

Pour utiliser le système, l'utilisateur saisit via l'IHM les données d'entrée de la mission qui sont envoyées à *Environnement_simulation.* En réponse *Environnement_simulation* renvoie de manière dynamique l'évolution de la situation tactique et les agents entraînés renvoient les décisions préconisées qui sont représentées par l'IHM.

Notre premier prototype informatique présente cependant des limites : (1) beaucoup d'hypothèses simplificatrices ont été faites qui devront être dans la suite retirées pour que le scénario gagne en réalité opérationnelle, (2) les agents entraînés par apprentissage par renforcement profond ont appris sur des données simulées supposées parfaites, ce qui peut être problématique lorsque les décisions s'appuient sur des données réelles, et (3) l'IHM conçue ne permet que de représenter la partie gestion de la route de la mission. Des pistes d'amélioration du prototype seront proposées dans le chapitre Conclusion et perspectives.

Chapitre 5 : Evaluation de l'architecture cognitive et de l'aide à la décision proposées

Nous avons présenté dans le chapitre précédent le prototypage d'une architecture cognitive fondée sur le RPD et l'apprentissage par renforcement profond. Pour rappel, cette architecture cognitive, composante essentielle d'un système d'aide à la décision, permet de préconiser des actions à un organe de commandement lors de la conduite de missions.

Ce prototype ne peut se substituer à un SAD finalisé et n'est pas utilisable dans des conditions opérationnelles. Ce prototype exploratoire a été principalement développé afin d'évaluer la simulation mentale par apprentissage par renforcement profond. Pour cela, les préconisations de décisions de l'architecture cognitive ont été comparées à (1) des résultats obtenus à l'aide d'un algorithme robuste et reconnu, et (2) des décisions prises par l'humain lors d'une campagne d'expérimentations en condition simulée, menée à NAVAL GROUP.

5.1. Evaluation de l'aide la décision par comparaison avec l'algorithme A*

L'agent entraîné pour la gestion de la route a été comparé à un algorithme de plus court chemin A* [142], [143] afin de confronter les temps d'exécution des algorithmes et les distances parcourues préconisées. A* est un algorithme largement utilisé dans la recherche de plus court chemins de graphes. Il consiste à tracer un chemin entre plusieurs nœuds d'un graphe en créant un arbre du chemin le moins coûteux entre le nœud de départ et le nœud d'arrivée. Pour chaque nœud du graphe, A* utilise une fonction $f(n)$ qui donne une estimation du coût total d'un chemin empruntant le nœud n :

$$f(n) = g(n) + h(n) \quad (1)$$

Avec $f(n)$ coût total estimé du chemin passant par le nœud n, $g(n)$ coût du chemin entre le nœud de départ et n et $h(n)$ coût estimé du noeud n à l'objectif et il s'agit de la partie heuristique de la fonction de coût.

A* est très utilisé car il trouvera toujours la solution optimale à condition qu'elle existe et n'est pas contraint à une recherche de chemin unidirectionnelle. Le choix de comparer notre agent à A* a donc été fait par rapport à l'optimalité du résultat trouvé par A* qui nous a permis de savoir si notre agent entraîné permettait comme souhaité de trouver un chemin optimal. Cependant A* n'est pas le meilleur algorithme pour chaque problème en terme de traitement requis et utilise beaucoup de mémoire pour tenir compte de chaque nœud créé. Cette limite de mémoire s'est confirmée avec la comparaison du temps d'exécution des deux algorithmes.

Pour cette première comparaison, les hypothèses suivantes ont été faites :

- Seuls les obstacles statiques ont été pris en compte.

- Les données météorologiques sont constantes sur la zone maritime afin de mettre en œuvre une fonction de coût A* simple.
- La distance euclidienne appelée $d(a, b)$ est utilisée comme fonction heuristique de A*. Cette distance permet à l'algorithme de proposer huit actions discrètes pour se déplacer sur la grille maritime : haut, bas, droite, gauche et les diagonales.
- L'espace des actions de l'agent en charge de la gestion de la route va être modifié pour concorder avec les actions proposées par A* et donc remplacé par un ensemble de huit actions discrètes : haut, bas, droite, gauche et les diagonales.

L'algorithme A* et l'aide à la décision ont été comparés sur deux cas d'application. Pour chacun des deux cas la zone maritime était la même en mer Méditerranée, définie par le couple $[33°; 34{,}5°] \times [15{,}5°; 17°]$, avec trente obstacles statiques pour le premier cas et soixante pour le second. Les obstacles statiques et le point de départ ont été générés pseudo-aléatoirement et comme les données météorologiques sont constantes tout au long du transit, la vitesse maximale admissible de la frégate est restée constante. Les résultats obtenus sont présentés dans le Tableau 5-1 , où le temps de calcul moyen et la distance moyenne parcourue ont été calculés pour dix scénarios différents dans chaque cas d'utilisation.

Tableau 5-1 : Comparaison entre A* et l'aide à la décision (AD) sur la partie gestion de la route

Nbre d'obstacles statiques	Algorithme/AD	Temps d'exécution moyen (s)	Distance moyenne parcourue (km)
30	A*	$3{,}80 \pm 0{,}78$	130 ± 29
	AD	$1{,}67 \pm 0{,}19$	$126{,}5 \pm 31$
60	A*	$4{,}25 \pm 1{,}04$	$133{,}5 \pm 32{,}8$
	AD	$1{,}91 \pm 0{,}10$	$130{,}5 \pm 33{,}5$

Nous remarquons que le temps d'exécution moyen (i.e : le temps pour préconiser un trajet) de l'aide à la décision est beaucoup plus rapide que celui de A* dans les deux cas d'utilisation. La différence entre les temps d'exécution moyens augmente lorsque le nombre d'obstacles statiques augmente, en effet la différence est de 2,13 s pour trente obstacles statiques et de 2,34 s pour soixante obstacles statiques. De plus, l'écart type du temps d'exécution augmente pour A* et diminue pour l'aide à la décision. Ces résultats peuvent laisser supposer que l'écart sera plus important lorsque les scénarios seront plus complexes. Cet écart n'est pas surprenant car les ressources informatiques à mobiliser pour les deux techniques sont significativement différentes. L'agent entraîné par apprentissage par renforcement profond peut nécessiter beaucoup de temps pour l'entraînement mais une fois entraîné, il est capable de prédire un chemin en quelques secondes, sans mobiliser d'importantes ressources informatiques. En revanche, l'algorithme A* prendra plusieurs minutes et secondes pour trouver une solution et utilise chaque fois beaucoup de ressources informatiques. Les temps d'éxecution étant fortement corrélés aux nombres d'obstacles, l'algorithme A* est donc moins bien adapté pour calculer rapidement une route que notre algorithme d'apprentissage par renforcement profond.

En ce qui concerne la distance moyenne parcourue, l'écart n'est pas significatif et reste à peu près constant, mais il pourrait être beaucoup plus important avec l'introduction d'obstacles mobiles. Cette première comparaison démontre que notre architecture cognitive est bien adaptée à la préconisation de decisions dans le cadre de la gestion de la route en temps contraint.

5.2. Evaluation de l'aide à la décision par une campagne d'expérimentations

5.2.1. Introduction

Un protocole d'évaluation a été mis place avec le service Facteurs Humains de NAVAL GROUP afin d'estimer les performances de l'aide à la décision proposée aux opérationnels lors de la conduite de mission. Des expérimentations ont été conduites avec des personnes ayant une expérience opérationnelle de niveaux différents afin d'évaluer (1) l'influence de l'aide à la décision sur le type de participants et (2) la performance des équipes assistées de l'aide à la décision.

En raison des conditions sanitaires liées à la crise du COVID-19, nous n'avons pas été en mesure de réunir un nombre d'opérationnels suffisant pour chaque groupe afin d'interpréter de façon exhaustive les résultats. Nous avons donc conduit dans un premier temps une campagne d'expérimentations à l'aide des membres du personnel de NAVAL GROUP. Des officiers réservistes de la Marine Nationale, des ingénieurs et chercheurs en R&D des systèmes de conduite de mission ont été mobilisés. Cette première phase d'évaluation nous a ainsi permis d'évaluer le potentiel de la démarche méthodologique et de l'architecture cognitive proposée, les fonctions du SAD et son ergonomie, le protocole d'évaluation et les premiers éléments concernant l'apport de l'aide à la décision. D'autres phases de tests impliquant un grand nombre d'opérationnels seront programmées dès que le contexte sanitaire le permettra afin de poursuivre l'évaluation du prototype dans des conditions quasi-opérationnelles.

L'évaluation de l'architecture cognitive a été faite avec deux groupes (G1 et G2). G1 est constitué de participants tandis que G2 est constitué de participants couplés au prototype d'aide à la décision. Les personnes sans expérience opérationnelle ont été affectées aux tâches de gestion de la route car notre premier prototype ne requiert pas de compétences expertes en navigation. Pour ces tâches, G1 et G2 étaient composés de quatre participants. Pour les tâches de gestion de menaces asymétriques, les personnes ayant une expérience opérationnelle ou une spécialité en lutte contre la menace asymétrique ont été sollicitées. Pour ces tâches, G1 et G2 étaient composés de trois participants. Tous les participants ayant des niveaux d'expérience à peu près identiques, les critères de comparaison et de validation sont tous directement comparés entre eux.

Chaque participant de G1 et G2 réalise **individuellement** les quatre tâches suivantes :

1. Dans la première tâche une route est gérée au sein d'une zone maritime en présence d'obstacles statiques et de conditions météorologiques variables.
2. Dans la deuxième tâche, la route est gérée en continue en tenant compte du scénario de la première tâche et des obstacles mobiles.

3. Dans la troisième tâche, la phase de dissuasion d'une menace asymétrique est gérée.
4. Dans la quatrième tâche, la phase de dissuasion de trois menaces asymétriques est gérée.

Dans les deux premières tâches le but du scénario est identique : rallier la zone d'arrivée le plus rapidement possible en supposant que la frégate transite toujours à sa vitesse maximale admissible. Dans les deux dernières tâches le but est de dissuader les menaces en choisissant à chaque fois les armes les plus efficaces. Le scénario de chaque tâche est le même pour tous les participants et un temps d'environ 1h30 est à prévoir par participant pour la réalisation des quatre tâches et l'entretien final. Lors de cet entretien une discussion aura lieu sur les fonctionnalités proposées de l'aide à la décision, les perspectives de son usage et la confiance/pertinence du prototype. Chaque participant réalise les tâches dans un Jupyter Notebook [144] où il a à sa disposition toutes les consignes et informations nécessaires pour leurs réalisations. Jupyter Notebook est une application web open source compatible avec Python qui permet de créer et partager des documents contenant du code, des équations, des visualisations et du texte. Deux Jupyter Notebook sont fournis aux participants, le premier permet d'effectuer les tâches de gestion de la route (tâches 1 et 2) et le deuxième les tâches de gestion de menaces asymétriques (tâches 3 et 4). Avant d'accomplir les expérimentations, les participants participeront à une phase de formation où ils manipuleront deux Jupyter Notebook pour se familiariser aux tâches demandées et ils pourront poser des questions.

Afin de comparer les performances entre les deux groupes et d'évaluer l'aide à la décision, des critères de comparaison et de validation, c'est-à-dire des grandeurs permettant de comparer les performances des équipes et des seuils à atteindre pour valider l'aide, sont préalablement définis et établis avec le service Facteurs Humains de NAVAL GROUP. Les critères de comparaison et de validation sont à définir selon la tâche à réaliser et doivent permettre d'avoir des informations sur (1) la gestion de la tâche par les participants, (2) le temps pour accomplir les consignes et (3) la pertinence des réponses.

Ci-dessous, une description exhaustive des tâches à réaliser.

5.2.2. Tâche 1 et 2 : gestion de la route

Avant de proposer le protocole expérimental pour les tâches 1 et 2, la procédure appliquée par les officiers pour gérer la route est rappelée.

5.2.2.1. Procédure appliquée par les officiers

Lors de la gestion de la route d'une mission, les opérationnels utilisent différentes informations telles qu'une cartographie de la zone maritime comprenant les « No Go area » et les rails commerciaux, les conditions météorologiques, les contraintes temporelles et les forces en présence.

Pendant la conduite de la mission, à l'aide des données fournies par les capteurs du navire (radars, sonars, infrarouge), les opérationnels obtiennent des informations sur la situation tactique, telles que la position des obstacles mobiles, et ajustent la route planifiée en

conséquence. Si le temps leur permet, ils peuvent prendre en compte aussi les caractéristiques principales du navire.

5.2.2.2. Protocole expérimental

En adéquation avec la procédure utilisée par les officiers, le protocole expérimental des tâches 1 et 2 pour valider les préconisations de l'architecture cognitive est décrit ci-dessous.

5.2.2.2.1. Tâche 1 : Gestion de la route parmi des obstacles statiques

Lors de la tâche 1, les participants de G1 et G2 doivent gérer une route au sein d'une zone maritime avec des obstacles statiques et des conditions météorologiques variables. Le but est de sélectionner la route permettant de rallier la zone d'arrivée le plus rapidement possible en supposant que la frégate transite toujours à sa vitesse maximale admissible.

Les participants des deux groupes ont à disposition plusieurs données, identiques à celles utilisées par l'aide à la décision, à savoir une visualisation de la situation tactique de la zone maritime avec le point de départ et d'arrivée de la frégate et les obstacles statiques, telle qu'illustrée Figure 5-1. Ils ont aussi à disposition les données météorologiques de la zone, actualisées toutes les 3 heures avec seulement les grandeurs utiles pour estimer la vitesse maximale admissible. Les données météorologiques sont discrétisées tous les 0,5° et entre deux positions discrètes, il est supposé qu'elles ne varient pas.

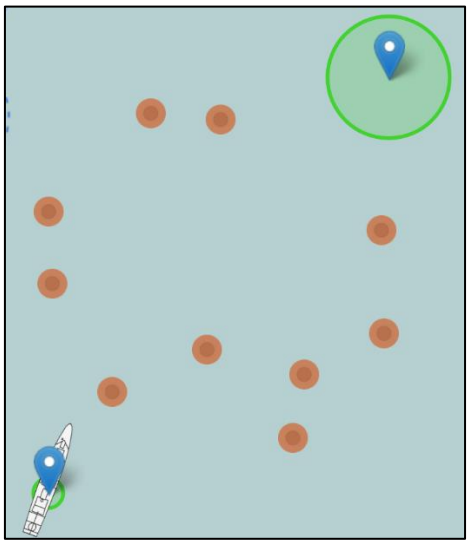

Figure 5-1 : Situation tactique tâche 1 pour les participants de G1 et G2

Les participants des deux groupes doivent sélectionner une route en cliquant directement sur l'IHM, illustrée Figure 5-2. La durée de l'expérimentation prend au maximum 2 minutes et dépend uniquement du temps nécessaire à chaque participant pour sélectionner une route. La Figure 5-2 montre aussi la différence d'IHM pour G1 et G2. Sur cette IHM :

- Les participants des deux groupes doivent cliquer dans le triangle autour de la frégate pour sélectionner les nouvelles positions de la frégate. Ce triangle a une ouverture de 60 ° pour être en adéquation avec les angles de cap que peut préconiser l'aide à la décision. La hauteur du triangle indique la distance maximale que peut parcourir la frégate au bout de 8 minutes, car l'aide à la décision prédit une nouvelle position toutes les 8 minutes ($t_{trajectoire}$). Entre chacune des positions discrètes le mouvement du navire est supposé rectiligne uniforme.
- Il n'est pas demandé aux participants de respecter les règles de navigation (COLREGs) [145] car pour l'instant le prototype n'en tient pas compte.
- La vitesse maximale admissible est directement approximée par le modèle prédictif introduit dans le Chapitre 3 en fonction de la position choisie et des conditions météorologiques.
- Les participants de G2 ont en plus à leur disposition la trajectoire préconisée par l'aide à la décision (Figure 5-2).

Une fois la zone d'arrivée atteinte la tâche se termine et les positions choisies ainsi que le temps mis pour décider du trajet sont enregistrés.

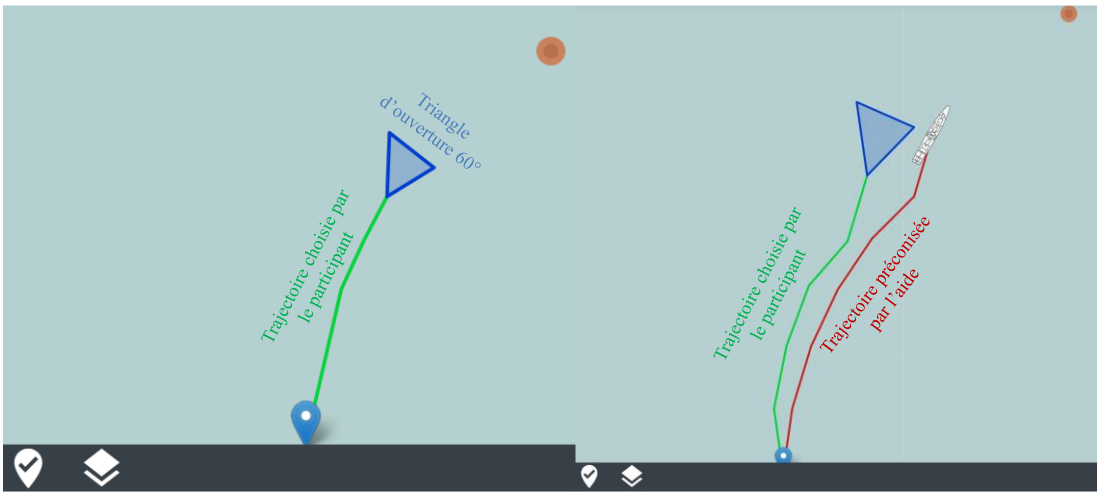

Figure 5-2 : IHM proposée aux participants de G1 (à gauche) et de G2 (à droite)

5.2.2.2.2. Tâche 2 : Gestion de la route au sein des obstacles statiques et mobiles

Lors de cette deuxième tâche, les participants de G1 et G2 doivent réaliser en temps accéléré la gestion de la route selon le scénario de la tâche 1 (cf : Figure 5-1) et en fonction des obstacles statiques et mobiles en présence. Le but est toujours de rallier la zone d'arrivée le plus rapidement possible. Les participants ont les données d'entrée de la tâche 1 et doivent sélectionner les nouvelles positions de la frégate de la même manière qu'en tâche 1.

5.2.3. Protocole d'évaluation

Pour l'évaluation de ces deux premières tâches, les critères de comparaison et de validation sont le temps total du trajet planifié, la distance totale parcourue et le temps de réalisation de la route.

Le temps total du trajet et la distance totale parcourue sont utilisés pour comparer l'écart de performance entre les participants de chaque groupe à niveaux d'expérience identiques et comparer les performances moyennes de chacun des groupes.

De la même manière, les temps de réalisation de chaque groupe sont calculés et comparés. Les écarts de performance entre les participants de chaque groupe à niveaux d'expérience identiques sont aussi comparés.

5.3. Tâche 3 et 4 : gestion de menaces asymétriques

De même que pour les tâches 1 et 2, la procédure appliquée par les officiers pour gérer la menace asymétrique est rappelée.

5.3.1. Procédure appliquée par les officiers

Des pistes radars sont utilisées afin d'obtenir des informations sur la position, la vitesse et le CPA (Closest Point of Approach) de la cible. Le CPA est un point estimé où la distance entre deux objets, dont l'un au moins est en mouvement, atteindra sa valeur minimale. Cette estimation est souvent utilisée pour évaluer le risque de collision entre deux navires. Les décideurs n'utilisent pas d'indicateurs pour prendre leurs décisions, et appliquent les ROE (règles d'engagement) quoiqu'il en coûte, selon une phase de dissuasion suivie d'une phase de neutralisation. Ils observent alors la réaction de la menace suite aux actions prises pour s'adapter. Même si le comportement d'une menace est irresponsable et non malveillant, les mêmes ROE sont réalisées, les militaires peuvent en déroger seulement si la menace est un drone sans passager. Le temps de prise de décision dépend de la vitesse de la menace et de la distance à laquelle elle se trouve.

Lors de la phase de dissuasion, utiliser l'arme non létale la plus efficace permet de lever plus rapidement l'ambiguïté sur les intentions du navire menaçant. Différentes actions sont associées à chaque arme non létale. Par exemple sur frégate, une arme non létale peut engager différentes actions : une action visuelle et sonore d'avertissement et une action visuelle, sonore de contrainte, etc. Une action d'avertissement est ponctuelle alors qu'une action de contrainte est continue.

5.3.2. Protocole expérimental

En adéquation avec la procédure utilisée par les officiers, le protocole expérimental des tâches 3 et 4 pour valider les préconisations de l'architecture cognitive est décrit ci-dessous.

5.3.2.1. Tâche 3 : Gestion d'une menace asymétrique

Lors de cette troisième tâche, les participants de G1 et G2 doivent gérer la phase de dissuasion lors de la survenue d'une menace asymétrique. Pour rappel la menace a une direction convergente sur la frégate. La phase de neutralisation et le respect des règles d'engagement (ROE) ne sont pas demandés car le prototype ne les gère actuellement pas et l'expérimentation s'achève à l'issue de la phase de dissuasion, c'est-à-dire lorsque la menace s'introduit dans le périmètre de neutralisation de la frégate. Il est supposé que toutes les menaces s'introduisent dans ce périmètre. Le but des participants de G1 et G2 est de choisir lors de la phase de dissuasion les armes non-létales les plus efficaces à disposition. Il n'y a aucune contrainte d'utilisation pour les armes.

Les participants doivent choisir les armes le plus rapidement possible et n'ont que dix secondes au maximum pour choisir l'arme, car toutes les dix secondes (t_{menace}) la menace asymétrique actualise sa position pour progresser en direction de la frégate en fonction de sa vitesse et de son cap.

L'aide à la décision qui gère actuellement la menace asymétrique ne tient compte que des données radars en entrée. Ainsi chacun des participants des deux groupes dispose seulement de données radars donnant des informations sur la menace telle que sa vitesse, sa position.

L'expérimentation est réalisée en temps accéléré car il est supposé dans l'expérimentation que les menaces ne réagissent pas aux armes engagées. La durée de l'expérimentation dépend de la vitesse de la menace et de la frégate et dure au maximum 2 minutes à partir de la distance de dissuasion jusqu'à celle de neutralisation. L'expérimentation conduite auprès des deux groupes est réalisée de la manière suivante :

Chaque participant dispose d'une vidéo du scénario de menace asymétrique, illustré Figure 5-3. Dans cette animation :

- Le navire menaçant est en position initiale à l'entrée du périmètre de dissuasion,
- Au-dessus de la menace s'affiche son identifiant, sa vitesse ainsi que son coefficient de menace. Pour le G2, il y aura en plus une annotation ENL (Equipement Non létal). Cette annotation correspond à l'arme préconisée par l'aide à la décision.
- Les participants ont le choix entre trois effecteurs différents : un tir de semonce (ENL 0), un canon sonore (ENL 1) et un laser d'aveuglement (ENL 2). Ces effecteurs sont préconisés aux participants de G2 par leur code. Avant de lancer l'animation, ils peuvent prendre connaissance des caractéristiques des armes disponibles.
- Une fois la menace dans le périmètre de neutralisation de la frégate, la phase de dissuasion est terminée.

Chaque participant de G2 peut tenir compte des préconisations de l'architecture comme il le souhaite.

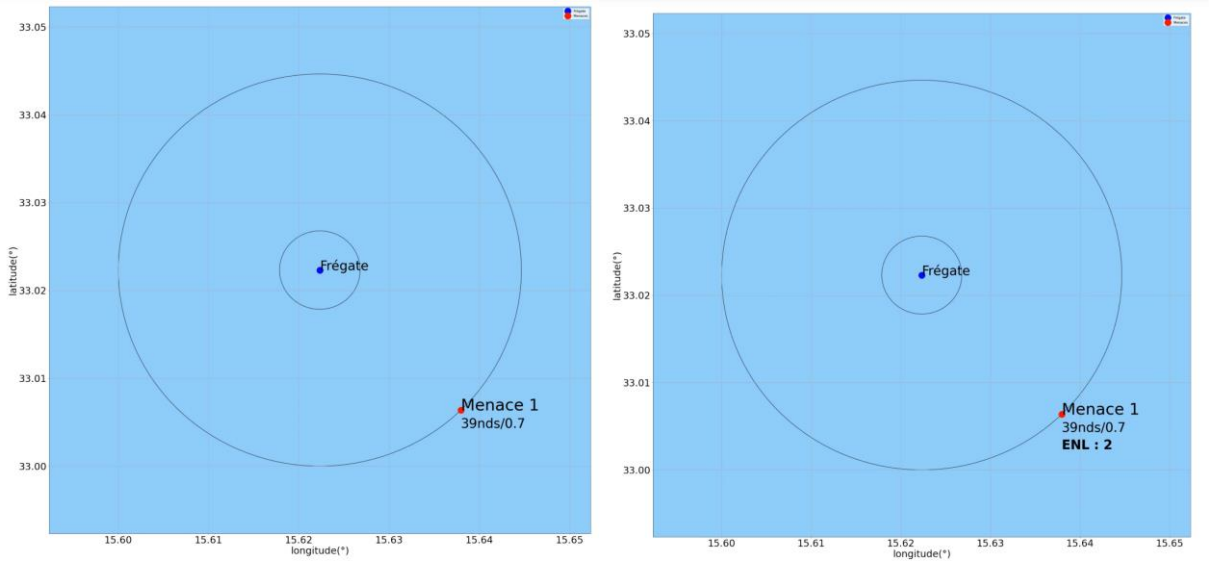

Figure 5-3 : Animations proposées aux participants de G1 (à gauche) et de G2 (à droite)

Dès que l'animation est lancée les participants doivent remplir un tableau, illustré Figure 5-4, où ils doivent cocher la case correspondante à l'arme choisie. La position de la menace s'actualise à chaque étape (toutes les dix secondes). Dans le tableau (Figure 5-4), T_0 correspond au début de la dissuasion lorsque l'animation commence. Une fois les dix secondes écoulées, un bip sonore informe les participants que les positions des menaces sont actualisées et qu'ils doivent cocher les cases de la ligne $T_0 + 10$. La tâche 2 se termine lorsque toutes les menaces se sont introduites dans le périmètre de neutralisation et les armes sélectionnées par les participants sont enregistrées.

Etapes	Menaces	Arme : 0 Tir de semonce	Arme : 1 Canon sonore	Arme : 2 Laser d'aveuglement
T_0	1			
	2			
	3			
	4			
$T_0 + n * 10$	1			
	2			
	3			
	4			

Figure 5-4 : Tableau fourni aux participants du G1 et G2

5.3.2.2. Tâche 4 : Gestion de plusieurs menaces asymétriques

Pour cette dernière tâche, les participants de G1 et G2 ont à gérer trois menaces en même temps. Ce choix a été fait afin de tester leurs réactions face à de multiples menaces. Le protocole expérimental et les données de cette tâche sont identiques à ceux de la tâche 3.

5.3.3. Protocole d'évaluation

Pour les tâches 3 et 4, les critères de comparaison et de validation sont le temps pris pour choisir un effecteur, le nombre de « non-réponses » et d'erreurs comptabilisées. Les erreurs sont obtenues en comparant l'arme sélectionnée par le participant à l'arme proposée par l'architecture, qui est soit la meilleure arme (une seule possibilité), soit une arme parmi les meilleures. Les meilleures armes à utiliser sont déterminées par les probabilités d'effet des armes calculées à l'aide de la méthode probabilité_effet_arme().

Différents estimateurs sont calculés pour chaque participant de chaque groupe : le temps moyen pour choisir un effecteur divisé par le nombre de menaces à gérer, le nombre de « non-réponses », le nombre moyen d'erreurs divisé par le nombre de menaces à gérer. Ces estimateurs permettent, d'évaluer (1) l'impact de l'aide à la décision sur le temps de prise de décision des participants et (2) d'évaluer la part de l'expérience du participant dans la performance de l'aide à la décision.

Afin d'évaluer la pertinence de l'aide à la décision, un ratio va être crée, (1) le nombre d'erreurs moyen de chaque groupe sera divisé par le temps de prise de décision moyen de chaque groupe. Ce ratio permettra de comparer le nombre moyen d'erreurs toutes les 10 secondes (une étape) de chaque groupe.

5.4. Interprétation des résultats

Une fois les deux tâches réalisées par les groupes G1 et G2, les résultats et les réponses aux entretiens finaux ont été enregistrés. La suite va expliciter et interpréter les résultats des expérimentations obtenus pour les deux tâches ainsi que les entretiens finaux des participants de G2.

5.4.1. Premiers résultats pour les tâches de gestion de la route

Le protocole d'évaluation décrit en 5.2.3 a été appliqué pour chacun des participants de chaque groupe où le temps du trajet, la distance parcourue et le temps de réalisation de la tâche ont été calculés. Le Tableau 5-2 ci-dessous illustre les valeurs moyennes de chacun des critères pour chaque groupe et chaque tâche.

Tableau 5-2 : Valeurs des critères pour les tâches de gestion de la route

Participants: 4 par groupe		Temps du trajet moyen	Distance moyenne	Temps moyen de réalisation
Tâche 1 (sans obstacles mobiles)	**Sans aide (G1)**	4 h 34 min	190,5 km	59 s
	Avec aide (G2)	4 h 36 min	186,5 km	1 min 30 s
Tâche 2 (avec obstacles mobiles)	**Sans aide (G1)**	4 h 40 min	195,5 km	1 min 16 s
	Avec aide (G2)	4 h 49 min	195,5 km	1 min 34 s

Les résultats de ce tableau permettent de tirer de premières observations et interprétations sur l'influence de l'architecture sur les utilisateurs. Concernant le temps total du trajet moyen, que ce soit en gestion de la route avec ou sans obstacles mobiles celui-ci a été plus long de quelques minutes lorsque les participants ont disposé de l'aide. Cependant un écart de moins de 10 minutes représentant moins de 4% du trajet total, n'est pas significatif. Différentes hypothèses peuvent être faites : (1) la route a tout de même été correctement gérée malgré un temps de mission légèrement supérieur, par exemple le nombre de manœuvres (changements de cap) a peut-être été réduit, (2) d'autres estimateurs doivent être développés pour mieux évaluer l'aide à la décision, (3) le profil des participants n'est pas adapté et (4) les deux tâches demandées sont encore trop simples pour observer tous les bénéfices d'une aide à la décision. Cette dernière supposition pourra être confirmée à l'aide des entretiens finaux. Le même constat et les mêmes interprétations peuvent être émis concernant la distance moyenne parcourue, où il n'y a pas d'écart significatif entre les deux groupes et selon les tâches.

Concernant le temps moyen de réalisation de la tâche, celui-ci est à chaque fois plus élevé lorsque les personnes ont disposé de l'aide, laissant supposer qu'avec l'aide les participants tenaient compte des préconisations et réfléchissaient plus au trajet qu'ils souhaitaient emprunter, témoignant ainsi d'un intérêt pour les préconisations de l'architecture cognitive. Néanmoins il faudra rester vigilant à ce que la réflexion sur les préconisations de l'aide ne vienne pas engendrer une surcharge cognitive supplémentaire. On peut aussi remarquer qu'en gestion de la route sans obstacles mobiles l'écart de temps moyen de réalisation entre G1 et G2 est de 31 s alors qu'avec obstacles mobiles il est de 18 s, supposant que plus la situation tactique augmentera en complexité et plus cet écart diminuera jusqu'à sans doute s'inverser, allégeant ainsi la charge cognitive de l'opérationnel. Le temps moyen de réalisation est à surveiller avec vigilance car il permettra de savoir si l'aide à la décision est adaptée aux décideurs et leur permet de décider plus vite dans des situations contraignantes ou alors au contraire les pénalisent en sollicitant des ressources cognitives supplémentaires.

A l'issue de ces premières expérimentations, nous pouvons constater que les décisions préconisées par l'aide à la décision dans deux cas simples de gestion de la route ne modifient pas de façon majeure les décisions des participants et les temps de réalisation d'une tâche, ce qui confirme que l'architecture cognitive développée préconise comme prévu des actions proches de celles que choisiraient un opérationnel. On peut donc supposer que dans le cadre de scénarios très complexes de gestion de la route, l'architecture cognitive préconiserait rapidement des actions satisfaisantes que l'opérationnel aurait envisagées en un temps plus long, confirmant ainsi l'intérêt de développer une aide à la gestion de la route fondée sur une méthode de prise de décision en situation naturelle.

Dans la suite le résumé des entretiens finaux pour ces deux tâches est énoncé.

5.4.2. Bilan des entretiens finaux pour les tâches de gestion de la route

Les entretiens finaux n'ont concerné que les participants de G2 car les questions étaient relatives à l'aide à la décision proposée. Le ressenti des participants de G1 et G2 sur la difficulté de la réalisation des deux tâches a tout d'abord été recueilli et décrit dans la suite.

5.4.2.1. Difficulté des tâches 1 et 2 pour G1 et G2

Tous les participants des deux groupes ont trouvé la tâche 1 de gestion de la route très facile et très intuitive. Le temps de réflexion pour imaginer une route a été très court ne demandant que très peu de réflexion.

La tâche 2 de gestion de la route a demandé à tous les participants plus de réflexion à cause de l'introduction des obstacles mobiles. Tous les participants ont pris plus de temps pour réfléchir afin de ne pas entrer en collision avec un obstacle et de toujours rallier le plus rapidement possible la zone d'arrivée. Certains participants ont navigué très près de certains obstacles. Sur les huit participants seulement deux avaient réfléchi à une stratégie avant de réaliser la tâche. Le premier a dû changer de route au milieu du scénario car sa gestion initiale était trop risquée à cause des obstacles présents et le second qui avait décidé de réaliser son parcours en naviguant toujours derrière les obstacles mobiles a respecté cette stratégie jusqu'à la fin s'affranchissant de l'aide à la décision.

Malgré ces quelques difficultés rencontrées, les participants ont trouvé les tâches à réaliser très simples ce qui confirme notre hypothèse (4) citée en 5.4.1 où les deux tâches demandées sont encore trop simples pour observer tous les bénéfices d'une aide à la décision.

5.4.2.2. Résumé des entretiens finaux des participants de G2

Suite à la réalisation des tâches de gestion de la route, les participants de G2 ont répondu aux deux questions (1) sur la confiance/pertinence des préconisations faites par l'architecture cognitive pour savoir si elles étaient adaptées à leurs décisions, (2) sur les perspectives d'usage d'une telle aide. La question relative aux fonctionnalités proposées ne leur a pas été posée car elle n'est pas adaptée à un premier prototype.

Un résumé est d'abord fait concernant la tâche 1.

Sur les quatre participants de G2 trois ont eu confiance dans l'aide. Un des participants a dit avoir une « confiance absolue » car il manquait d'expérience dans la réalisation de la tâche et préférait donc se rassurer avec l'aide. Un autre a eu confiance car le prototype faisait des recommandations ce qui lui a permis de se sentir libre dans ses choix et le dernier car il comprenait la logique des préconisations de l'aide confirmant que les décisions préconisées étaient adaptées à ses choix envisagés. Le dernier participant n'a pas pu déterminer une confiance ou non dans l'aide car il n'a pas regardé les préconisations. Concernant la pertinence du prototype sur cette première tâche, les participants ont trouvé les préconisations de l'architecture adaptées à la tâche demandée, compréhensibles et la plupart du temps en accord avec leurs décisions envisagées. L'usage du prototype ne leur a pas sollicité de ressources cognitives supplémentaires.

Concernant les perspectives d'usage d'une telle aide pour la tâche 1, deux participants imaginent un usage potentiel pour les opérationnels, car l'architecture est capable de préconiser des décisions satisfaisantes et compréhensibles pour la tâche demandée. Un autre a pensé à une perspective plus précise où le prototype pourrait s'intégrer dans le navire du futur pour l'aide au routage.

Pour terminer, les entretiens de la tâche 2 sont résumés. Entre les deux tâches, les remarques faites par les participants concernant les deux points, confiance/pertinence et perspectives d'usage se ressemblent.

Sur les quatre participants, les trois ayant fait confiance au prototype lors de la tâche 1, ont réitérer cette confiance lors de la tâche 2, sans rajouter de remarques supplémentaires par rapport à celles faites pour la tâche 1. Pour cette deuxième tâche où il y a eu l'introduction des obstacles mobiles, certains participants ont changé d'avis quant à la pertinence du prototype. Trois participants ont trouvé l'aide plus pertinente pour la conduite, à cause de l'introduction d'obstacles mobiles qui a augmenté leur charge mentale. Deux ont été rassurés par celle-ci, car les décisions préconisées correspondaient à ce qu'ils avaient imaginé et un autre a ressenti une baisse de charge cognitive, ce qui prouve la pertinence d'une aide pour assister des opérationnels.

Tous les participants envisagent des perspectives d'usage de l'aide à la décision proposée pour cette seconde tâche. Un participant imagine un usage pour une mission où il y aurait plusieurs checkpoints à rallier, un autre pense que ce serait très utile dans des situations non prévisibles, dangereuses et inattendues. Un autre verrait le prototype s'intégrer dans le navire du futur mais cette fois-ci dans un mode « autopilote » pour permettre à l'organe de commandement de se concentrer sur d'autres tâches plus critiques et importantes.

Dans la suite les résultats obtenus pour les tâches de gestion de menaces asymétriques sont introduits.

5.4.3. Résultats pour les tâches de gestion de menaces asymétriques

De même que pour les tâches de gestion de la route, le protocole d'évaluation pour les tâches de gestion de menaces asymétriques décrit en 5.3.3 a été appliqué pour chacun des participants de chaque groupe où le temps de décision moyen par menace, le nombre de « non-réponse », le nombre d'erreurs moyen par menace et le ratio représentant le nombre d'erreurs moyen divisé par le temps de décision moyen par étape ont été calculés. Le Tableau 5-3 ci-dessous illustre les valeurs de chacun des critères pour chaque groupe et chaque tâche.

Tableau 5-3 : Valeur des critères pour les tâches de gestion des menaces après dépouillement

Participants: 3 par groupe		**Temps de décision moyen par menace (s/menace)**	**Nbre de "non-réponse"**	**Nbre d'erreurs moyen par menace (erreur/menace)**	**Ratio, nbre erreurs moyen /temps de décision moyen par étape (erreur /10s)**
Tâche 3 (une menace)	**Sans aide (G1)**	3,7	1	0,5	1,4
	Avec aide (G2)	2,3	0	0,3	1,4
Tâche 4 (trois menaces)	**Sans aide (G1)**	1,8	3	0,4	2
	Avec aide (G2)	2,02	0	0,3	1,7

Comme pour les tâches de gestion de la route, le faible échantillon de participants nous permet uniquement de faire des premières observations sur l'influence de l'aide à la décision sur les utilisateurs. Le premier critère du tableau qui est le temps de décision moyen par menace

permet d'observer l'impact de l'usage du prototype sur le temps de décision des participants. Lorsque les participants n'ont eu qu'une menace à gérer, les participants de G2 ont en moyenne pris les décisions plus rapidement que ceux de G1. Cependant avec trois menaces, les résultats sont sensiblement identiques. Ce premier résultat supposerait que lorsque les participants sont à l'aise avec une charge mentale très modérée, ils ont le temps d'analyser rapidement les préconisations de l'aide et de réduire en moyenne leur temps de prise de décision. Au contraire lorsqu'ils sont dans un état de charge mentale relativement élevée, l'introduction d'une aide ne semble ni réduire ni augmenter significativement leur temps de prise de décisions puisqu'un écart de deux dixièmes de seconde n'est pas significatif.
Les premiers résultats obtenus pour le nombre de « non-réponses » nous permettent de constater d'une part que plus le nombre de menaces à gérer augmente et plus les « non-réponses » augmentent pour les participants n'ayant pas l'aide. Cette tendance est très probablement corrélée à l'augmentation de la charge mentale. D'autre part, quel que soit le nombre de menaces à gérer, G2 comptabilise zéro « non-réponse », laissant supposer que les décisions préconisées par l'architecture semblent adaptées aux décideurs leur permettant de s'appuyer sur l'aide lorsqu'ils sont en surcharge cognitive. Le critère suivant, le nombre d'erreurs moyen par menace permet de comparer la pertinence des effecteurs choisis pour chaque groupe. Les résultats obtenus sont identiques lors de la gestion d'une menace pour G1 et G2. Ce résultat peut supposer soit un certain manque de confiance dû à un manque de formation des participants sur le prototype, soit une charge mentale très modérée impliquant que les décideurs ont confiance dans leur propre choix et préfèrent suivre leur intuition. Lors de la gestion de trois menaces, le G2 comptabilise en moyenne légèrement moins d'erreurs par étape par rapport au G1, laissant supposer qu'en situation stressante avec un besoin de prise de décision rapide, les décideurs qui ont à disposition une aide adaptée à leur processus de décision feront moins d'erreurs.

De même que pour les tâches de gestion de la route, ces premières expérimentations nous permettent de constater que les décisions préconisées par l'aide dans les deux cas de gestion de menaces asymétriques, ne modifient pas significativement le temps de prise de décision moyen par menace et diminuent le nombre de « non-reponses », confirmant que l'architecture cognitive développée préconise des décisions compréhensibles et adaptées aux décideurs tout en évitant la sollicitation de ressources cognitives supplémentaires. Ceci se confirme aussi avec le nombre d'erreurs moyen divisé par le temps de décision moyen par étape montrant que les décisions préconisées sont adaptées aux tâches demandées, appuyant encore l'intérêt de développer une aide à la décision fondée sur une méthode de prise de décision en situation naturelle.

5.4.4. Bilan des entretiens finaux pour les tâches de gestion de menaces asymétriques

Comme pour les tâches de gestion de la route, les entretiens finaux n'ont concerné que les participants de G2 car les questions sont relatives à l'aide à la décision proposée.

Avant de résumer les entretiens finaux, la charge mentale des participants a été évaluée après la réalisation de chacune des deux tâches. Dans la suite, la charge mentale ressentie par les participants des deux groupes est décrite.

5.4.4.1. Charge mentale ressentie par les participants de G1 et G2 pour les tâches 3 et 4

Concernant la tâche de gestion d'une menace asymétrique par les participants de G1, deux des participants ont ressenti une charge mentale très peu élevée. Très peu d'hésitations voire aucune n'a été constatée dans le choix des effecteurs. Le troisième participant ne s'est pas senti à l'aise dès le début de l'expérimentation puisque dès la deuxième étape il n'a pas eu le temps de cocher de case. Cela ne l'a pas déstabilisé pour la suite puisqu'à partir de la troisième étape sa charge mentale a fortement diminué jusqu'à la fin où il est redevenu à l'aise avec très peu d'hésitations. Ces hésitations auraient surement pu être évitées avec plus de formations en amont. Ce constat montre l'importance de l'accompagnement humain qui doit être mis à disposition de l'utilisateur d'un SAD avant (formation) et pendant la mission.

Lors de la gestion de trois menaces, tous les participants de G1 ont qualifié leur charge mentale d'élevée et ont trouvé la tâche beaucoup plus complexe avec plus d'hésitations dans leur choix. Pour tous les participants, la charge mentale est redevenue très modérée lorsqu'il ne restait à la fin qu'une menace à gérer. Un des participants a essayé d'adopter une stratégie en repérant vite la position des trois menaces et en se concentrant toujours sur la cible la plus proche de la frégate. Pour un autre participant, malgré une charge mentale assez élevée ses choix ont été faits rapidement dès le début de l'expérimentation. Cependant, au milieu de la tâche son niveau de stress et sa charge ont augmenté subitement, car une des menaces n'a pas agi comme il l'avait imaginée. Cet événement inattendu lui a causé une certaine confusion qui s'est manifestée par un choix d'effecteurs en trop sur deux étapes. Le dernier participant a été très l'aise sur les cinq premières étapes avec quasiment aucune hésitation sur le choix des effecteurs. A partir de la sixième étape, certaines menaces ont commencé à s'introduire dans le périmètre de neutralisation, ce qui l'a beaucoup déstabilisé. Il s'est alors trompé sur le numéro des menaces, a coché les cases ne correspondant pas aux bonnes menaces et a réalisé des « non-réponses ». A partir de là jusqu'à quasiment la fin du scénario, sa charge mentale a été très élevée, à plusieurs reprises il a changé ses choix ce qui dans une situation réelle n'aurait pas été possible.

Lors de la gestion d'une menace asymétrique, les participants de G2 ont eu dans l'ensemble une charge mentale très modérée. Tous les participants ont regardé les préconisations de l'aide et l'un d'entre eux a été déstabilisé par sa présence. Le participant en question a regardé l'aide dès le début du scénario et a eu un temps de réflexion plus important sur les deux premières étapes. Passé ces deux étapes, il est redevenu très rapide dans ses choix, s'est senti à l'aise avec une charge mentale très modérée. Cet événement confirme encore l'importance d'une phase de formation des participants sur le prototype. Les deux autres participants ont eu une charge mentale très peu élevée avec quasiment aucune hésitation dans leur choix.

Lors de la gestion de trois menaces, la charge mentale ressentie a été différente entre les participants et ils ont tous les trois plus tenu compte des préconisations de l'aide que lors de la tâche avec une seule menace. Un des participants a beaucoup regardé ce que l'aide lui proposait, lui permettant de qualifier sa charge mentale de modérée même si elle était plus élevée que pour une menace, montrant que l'architecture préconisait des décisions adaptées. Au contraire un autre participant a commencé l'exercice très à l'aise, mais rapidement il a été en surcharge cognitive l'empêchant de réaliser correctement l'exercice. Pour ce participant l'introduction en plus d'une aide ne l'a pas aidé, à cause du manque de formation car il s'interrogeait en

permanence sur le choix fait par l'aide, même si les décisions préconisées correspondaient à des choix envisagés. Le dernier participant a ressenti une charge mentale élevée, s'est appuyé sur l'aide et a réalisé la tâche sans problème, montrant que pour lui l'architecture préconisait des décisions adaptées, compréhensibles et respectant sa méthodologie de prise de décision.

5.4.4.2. Résumé des entretiens finaux des participants de G2

Suite à la réalisation des tâches de gestion de menaces asymétriques, les participants de G2 ont répondu à deux questions :

(1) Quelle est la confiance/pertinence des préconisations de l'architecture proposée ?
(2) Quelles sont les perspectives d'usage d'une telle aide ?

Le résumé regroupe les deux tâches de gestion de menaces asymétriques.

Concernant la confiance accordée dans le prototype, les trois participants ont eu des degrés de confiance différents. Un des participants a avoué avoir été influencé par l'aide et a eu complètement confiance car les préconisations l'ont aidé dans ses moments de stress et correspondaient à ses choix envisagés. Un autre participant a gardé une confiance neutre dans le prototype. Il a commencé les expérimentations sans aucun *a priori*, et au cours de celles-ci, a essayé d'augmenter sa confiance dans le prototype en essayant de comprendre les préconisations. Dans certains cas, ses choix étaient identiques à ceux proposés ce qui a augmenté sa confiance. Lorsqu'ils étaient différents, il a essayé de comprendre mais comme il n'a pas réussi à cause du peu de temps dont il disposait, il a préféré suivre sa propre intuition. Le participant confirme tout de même que s'il avait compris les choix de l'aide, il les aurait suivis et sa confiance aurait augmenté, montrant une fois de plus l'importance d'une formation avant utilisation du prototype.

Les participants ont trouvé l'architecture pertinente avec dans la plupart des cas des décisions proposées adaptées à celles qu'ils auraient envisagées. Un participant trouve que l'aide proposée permet de conforter un choix dans le cas où le décideur serait en état de stress ou de surcharge cognitive, prouvant encore que notre architecture cognitive semble adaptée au processus décisionnel en proposant des décisions satisfaisantes rapidement.

Les participants imaginent des perspectives d'usage sur le système de LCMA (Lutte Contre Menace Asymétrique) à bord des navires pour venir en aide aux opérationnels.

5.5. Conclusion

Dans ce dernier chapitre, une première comparaison de l'architecture cognitive sur la partie gestion de la route avec l'algorithme A* a été réalisée. Les résultats obtenus ont démontré que notre architecture cognitive est bien adaptée à la préconisation de decisions dans le cadre de la gestion de la route dans en temps contraint. Cependant, certaines fonctionnalités de l'aide n'ont

pas été prises en compte pour faciliter la comparaison avec l'algorithme A*, telles que la gestion des menaces asymétriques ou l'évitement des obstacles mobiles.

Dans une seconde partie, un protocole d'évaluation a été mis en place et validé en interne afin de s'intéresser à l'apport de l'aide à la décision aux opérationnels lors de la conduite de missions. Les expérimentations ont été conduites avec un faible échantillon de personnes internes à NAVAL GROUP à cause du contexte sanitaire lié à la crise du COVID-19. Du fait du manque d'expérience opérationnelle et du faible échantillon, les résultats des expérimentations nous ont permis de constater seulement des premiers éléments d'évaluation, et de recueillir des pistes judicieuses d'améliorations et d'applications concernant le déroulement du protocole.

Au vu des premiers résultats obtenus sur l'ensemble des tâches, nous pouvons constater que les décisions préconisées par l'aide à la décision ne modifient pas de façon majeure les décisions des participants ainsi que les temps de réalisation des tâches, confirmant que l'architecture cognitive développée semble préconiser comme prévu (1) des actions satisfaisantes proches de celles que choisiraient un opérationnel en temps contraint, (2) sans nécessiter à priori de ressources cognitives supplémentaires et (3) confirmant ainsi l'intérêt de développer un SAD fondé sur le RPD, une des méthodes de prise de décision en situation naturelle.

Les entretiens avec les participants ayant eu l'aide à la décision ont permis de valider certains des résultats obtenus. La plupart des participants ont trouvé les décisions préconisées adaptées aux choix envisagés et n'ont pas sollicité de ressources cognitives supplémentaires.

Ces entretiens ont permis de constater que l'acceptation d'un tel prototype était fonction de la confiance accordée à celui-ci. Cette confiance dépend essentiellement de la compréhension des décisions préconisées. Il a été constaté que le niveau de compréhension dépendait fortement de la formation dispensée aux participants avant utilisation du prototype.

Conclusion et perspectives

BILAN DE LA THESE

Le but de cette thèse était de proposer une aide à la décision pour la conduite de missions navales militaires au niveau tactique en respectant la méthodologie de prise de décision utilisée par les opérationnels. Le besoin de concevoir un système d'aide à la décision pour assister l'organe de commandement lors de la conduite de missions provient du constat suivant : en mission, le Commandant évalue en temps réel le succès de la mission à l'aide d'indicateurs construits et calculés par l'humain. Cependant, les situations tactiques sont aujourd'hui de plus en plus complexes et stressantes empêchant les humains soumis à leurs propres limites de traiter de manière fiable et rapide la quantité de données recueillies par les capteurs du navire. Dans cet environnement militaire incertain et contraint par le temps où les opérationnels doivent prendre des décisions rapidement, nous avons fait le choix de fonder notre système d'aide à la décision sur le RPD, méthode issue des NDM, adaptée à la prise de décision rapide en environnement contraint. Ce choix a été fait (1) pour proposer un système s'appuyant sur les informations pertinentes recueillies par les opérationnels pour décider, (2) pour être en adéquation avec le processus décisionnel mis en œuvre lors de la prise de décision rapide en situation opérationnelle et (3) pour que le système préconise des décisions proches de celles qu'envisageraient les opérationnels afin de ne pas solliciter de ressources cognitives supplémentaires, d'augmenter leur confiance dans le système et d'ainsi réduire leur charge mentale et leur temps de prise de décision.

Méthodes de gestion de la conduite de missions navales militaires

La première partie des travaux de recherche de cette thèse s'est intéressée aux méthodologies utilisées par la Marine Nationale pour la planification et la conduite de missions. La Marine s'appuie sur la MEDOT, une méthode analytique, et sur le processus d'évaluation ; mais à cause de la complexification des situations tactiques et du besoin de prises de décisions rapides, il est apparu que les opérationnels ne disposaient plus des ressources nécessaires en conduite pour appliquer de manière adéquate des méthodes analytiques. Ils se tournent alors vers des méthodes de prise de décision rapides, telles que le RPD issus des NDM. Pourtant, même si le RPD garantit une prise de décision beaucoup plus rapide que les méthodes analytiques, les décideurs restent soumis à la surcharge cognitive et à la rationalité limitée. Pour leur venir en aide, des systèmes d'aide à la décision ont donc été implémentés tentant de simuler le plus complètement le RPD, notamment la reconnaissance de schémas et la simulation mentale, les deux phases essentielles du RPD.

Conception du système d'aide à la décision

Après un état de l'art sur les SAD fondés sur le RPD, où un certain nombre d'avantages et de limites ont été identifié, nous avons fait le choix de nous inspirer de la structure de l'architecture cognitive de Kunde et Darken [82], [83], [84] pour concevoir notre SAD et d'utiliser l'apprentissage par renforcement profond pour modéliser au mieux la simulation mentale. L'apprentissage par renforcement profond est adapté pour modéliser le comportement humain dans un environnement dynamique et nous a permis de proposer une aide à la décision dynamique s'adaptant à des situations inédites, et robuste à la rareté des événements. Notre système d'aide à la décision est composé d'une base de données, d'une architecture cognitive et d'une IHM. C'est l'architecture cognitive proposée et composée des trois modules, **Environnement de simulation**, **Connaissance situationnelle** et **Simulateur mental et composante décisionnelle**, qui permet de reproduire les étapes du RPD. Chacun des modules est conçu de la même manière, (1) les données d'entrée sont définies auprès d'experts et recueillies, (2) des méthodes sont choisies pour assurer les fonctionnalités souhaitées et (3) les décisions attendues ou sorties du module sont aussi déterminées auprès d'experts. Grâce à cette conception modulaire, l'architecture peut être améliorée et de nouvelles données d'entrée et méthodes peuvent être facilement intégrées.

Prototypage de l'architecture cognitive

Une fois notre système d'aide à la décision modélisé, uniquement le prototypage informatique de l'architecture cognitive a été réalisé, en Python 3 en suivant un formalisme POO. Un premier prototype exploratoire a alors été conçu pour fonctionner sur un scénario de mission, où une frégate devait se rendre dans une zone géographique pour y réaliser une action. L'architecture cognitive avait pour rôle d'aider les décideurs sur deux tâches principales : (1) emprunter la meilleure route pour atteindre la zone d'arrivée le plus rapidement possible et (2) gérer la ou les menaces asymétriques auxquelles elle serait potentiellement confrontée. Les données associées ont alors été présentées et une IHM a été implémentée en JavaScript afin d'évaluer le prototype.

Les modules de l'architecture cognitive sont implémentés sous forme de classes permettant de rajouter facilement de nouvelles fonctionnalités. Le module **simulateur mental et composante décisionnelle** est divisé en deux modules, un module apprentissage utilisé pour entrainer un agent à gérer la mission envisagée et un module simulation permettant de solliciter un ou plusieurs agents entraînés pour la conduite d'une mission. Dans le module apprentissage, deux agents ont été entraînés, un pour gérer la trajectoire et un pour gérer la menace asymétrique. Une fois les deux agents entraînés, leurs performances ont été validées sur des scénarios inédits nous permettant de conclure qu'ils étaient capables de s'adapter à des situations inédites et de préconiser des décisions dans un environnement dynamique. Dans le module simulation, chaque agent entraîné interagit dans son environnement. Pour utiliser le prototype en conduite, l'utilisateur entre via l'IHM les données d'entrée de la mission et les agents entraînés renverront les décisions préconisées.

Evaluation de l'architecture cognitive et de l'aide à la décision proposées

Dans une dernière partie, une première comparaison de l'architecture cognitive sur la partie gestion de la route avec l'algorithme A* a été réalisée, démontrant que notre architecture congnitive est bien adaptée à la préconisation de decisions dans le cadre de la gestion de la route en temps contraint.

Une évaluation de l'architecture cognitive a ensuite été mise en place dans le but de s'intéresser à l'apport de celle-ci aux opérationnels lors de la conduite de mission en condition simulée. Un protocole d'évaluation a été créé et des premières expérimentations ont été conduites avec des personnes internes à NAVAL GROUP. Cependant, à cause du manque d'expérience opérationnelle des participants et du faible échantillon, les résultats de ces premières expérimentations ne nous ont permis que de tirer de premières conclusions. Les premières observations montrent que l'architecture cognitive développée semble préconiser comme prévu (1) des actions satisfaisantes proches de celles que choisiraient un opérationnel en temps contraint, (2) sans solliciter de ressources cognitives supplémentaires. Ces deux observations confirment donc l'intérêt de développer un SAD sur le RPD, une des méthodes issues des NDM. Grâce aux entretiens on a aussi pu conclure sur l'importance de la confiance des utilisateurs accordée au prototype qui dépend fortement des préconisations et de la formation qu'ils auront reçue.

PERSPECTIVES

Suite à la modélisation, le développement, la démonstration et l'évaluation de notre aide à la décision, nous avons pu identifier plusieurs perspectives pour l'amélioration de ces différentes étapes. Dans la suite nous présenterons les perspectives sur la modélisation et le développement du SAD puis nous ouvrirons sur des perspectives plus larges concernant la mise en place d'un protocole d'évaluation plus abouti, et l'application d'un tel système à d'autres domaines.

Perspectives de modélisation

La première perspective de modélisation concerne la modélisation du RPD. Dans notre modèle, la reconnaissance de schémas déjà vus a été simplifiée pour mettre l'accent sur la simulation mentale du RPD, qui est la plupart du temps partiellement modélisée dans les SAD existants. Une perspective importante pour ce travail serait donc de modéliser complètement la reconnaissance de schémas déjà vus qui se produit instantanément dans notre modèle quel que soit les données recueillies. Beaucoup de SAD existants proposent différentes méthodes abouties qui pourraient être réutilisées.

La modélisation d'une mission navale militaire tactique devra être améliorée pour se rapprocher de la réalité. La modélisation de la situation tactique est actuellement beaucoup trop simple dans nos travaux par rapport à la réalité opérationnelle. Seules les données radars et

météorologiques ont été considérées pour obtenir l'évolution de la situation. Dans la suite, d'autres sources de données devront être prises en compte telles que les données AIS, sonars, optroniques, et plus généralement toute source de données utiles pour se rapprocher le plus possible d'une situation tactique réaliste. De plus, comme introduit dans le Chapitre 1, beaucoup d'anomalies maritimes peuvent survenir en conduite nécessitant la prise rapide de nouvelles décisions. Seul le risque de collision et la menace asymétrique ont été modélisés de manière simplifiée. A court terme, la modélisation de ces anomalies devra être complexifiée pour gagner en réalisme et à long terme d'autres devront être rajoutées au sein d'un éventuel nouveau sous-module et modélisées éventuellement à l'aide d'ontologies [146]. Ce nouveau sous-module pourra aussi être enrichi avec des règles métiers, telles que les règles de navigation (COLREGs) qui n'ont pas été intégrées au modèle de décision. De même que pour les anomalies maritimes et les règles métiers, les règles d'engagements, telles que celles évoquées en Chapitre 4 pour la lutte contre une menace asymétrique, qui n'ont pas été prises en compte, pourront être intégrées. Le comportement des navires actuellement cinématique devra être remplacé par un comportement dynamique. Pour terminer, une mission a été découpée en deux tâches principales à réaliser, où un agent par tâche a été entraîné par apprentissage par renforcement profond. Peu de tâches sont aujourd'hui gérées par l'architecture et si le nombre de tâches venait à beaucoup augmenter il faudrait peut-être s'orienter vers du Hierarchical Reinforcement Learning (HRL) [147]. Le HRL décompose un problème d'apprentissage par renforcement ou par renforcement profond en une hiérarchie de sous-problèmes ou de sous-tâches de telle sorte que les tâches parentales de niveau supérieur invoquent les tâches enfant de niveau inférieur comme s'il s'agissait d'actions primitives. L'avantage de la décomposition hiérarchique est une réduction de la complexité de calcul si le problème global peut être représenté de manière plus compacte et si des sous-tâches réutilisables peuvent être apprises ou fournies indépendamment.

Actuellement notre prototype vient en aide aux décideurs en proposant des décisions s'appuyant sur certains indicateurs en fonction de l'évolution de la situation tactique. Seuls trois indicateurs ont aujourd'hui été modélisés permettant d'avoir la vitesses maximale admissible d'une frégate, la probabilité de dissuasion d'une arme en fonction de la menace et le risque de collision. Afin de faciliter la compréhension du décideur, plus d'indicateurs de confiance et de risque devront être établis à court terme pour justifier aux utilisateurs les décisions préconisées. Cette perspective est très importante car c'est en gagnant la confiance des utilisateurs et en leur préconisant des décisions adaptées et compréhensibles qu'un SAD pourra les aider sans solliciter de ressources cognitives supplémentaires. Cependant, il faut rester vigilant à ce que le système n'induise pas de sous charge mentale aux décideurs. En effet, les décideurs pourraient avoir un excès de confiance, augmentant l'inattention et donc le risque d'erreurs. Un compromis doit donc être réalisé et une vigilance est à tenir quant à l'automatisation de certaines tâches.

A court terme, les préconisations de l'architecture cognitive sont amenées à être certainement plus complexes qu'une proposition d'arme ou de changement de cap. Une préconisation plus complexe induira certainement un temps de réflexion plus long au décideur. Il faudra alors dans l'évolution du SAD être capable de quantifier le gain entre l'efficacité de l'action préconisée et le temps de réflexion. Une action plus complexe peut demander plus de temps de réflexion au décideur mais sera au final plus efficace qu'une action moins complexe et plus rapide à comprendre.

Une autre perspective du prototype serait de rajouter une interaction entre lui et l'utilisateur où l'architecture tiendrait compte des choix du décideur pour lui faire de nouvelles préconisations. Par exemple pour la gestion de menaces asymétriques, aujourd'hui

l'architecture préconise des armes sans tenir compte de l'arme choisie par l'utilisateur, une amélioration serait donc qu'elle préconise une arme en tenant compte de l'historique des actions de l'utilisateur.

Perspectives d'implémentation

Les perspectives d'implémentation concernent la base de données, l'architecture cognitive et l'IHM. Notre base de données constituée contient principalement des données simulées. Ainsi la prochaine étape sera de constituer une base de données réelles alimentée en temps réel. L'architecture cognitive devra alors être modifiée pour la gérer avec l'implémentation de méthodes de prétraitement de données, voire de fusion de données lorsque les données commenceront à être massives et de sources hétérogènes.

Les perspectives d'amélioration de l'architecture cognitive vont concerner l'implémentation des environnements et des agents d'apprentissage par renforcement profond. Une des premières améliorations à réaliser concernent les données sur lesquelles les agents ont été entraînées, qui ont été simulées et sont donc supposées parfaites. Un entraînement sur des données supposées parfaites ne permet pas de garantir que les agents seront capables de préconiser des décisions sur des données réelles même prétraitées. Dans la suite des travaux soit du bruit doit être rajouté sur les données simulées soit dans l'idéal des données réelles doivent être utilisées pour garantir le fonctionnement du SAD en condition réelle. Une deuxième amélioration à réaliser concerne la complexification de l'environnement et de l'agent qui fonctionnera de pair avec la modélisation plus complexe d'une mission évoquée précédemment. Par exemple, l'agent qui gère la route ne prédit que toutes les huit minutes une action discrète de cap parmi cinq, dans la suite il faudrait mettre en place des actions différentes plus continues et une prédiction plus courte. De plus, l'apprentissage par renforcement se base sur un processus de décision Markovien, où il est supposé que l'information utile pour la prédiction du futur est entièrement contenue dans l'état présent du processus et n'est pas dépendante des états antérieurs. Cependant dans la réalité l'information utile dans l'état présent n'est pas toujours certaine ni exhaustive c'est pourquoi afin d'avoir une prise de décision plus réaliste, un apprentissage par renforcement basé sur un processus de décision markovien partiellement observable (PDMPO) [148] devrait plutôt être implémenté. Dans un PDMPO, l'incertitude est double. Non seulement l'effet des actions entreprises est incertain, mais de plus, on ne dispose que d'indices pour connaitre l'état dans lequel on se trouve, et donc pour décider. Ces indices sont appelés des observations.

Concernant l'IHM, de grosses améliorations sont à prévoir. En l'état, l'IHM ne permet de représenter que la partie gestion de la route de la mission, c'est pourquoi les prochains travaux doivent s'intéresser à intégrer la gestion de la menace asymétrique avec la préconisation des armes. Un travail en interne devra être réalisé pour savoir si des normes particulières doivent être respectées et pour s'inspirer de ce qui existe déjà.

Perspectives d'évaluation

Afin d'évaluer l'architecture cognitive et de conclure sur son bien-fondé auprès des utilisateurs, une campagne expérimentale en interne a été réalisée. Il faudrait dans la suite des travaux, réitérer le protocole expérimental proposé avec un plus gros échantillon d'opérationnels avec des expériences différentes allant du novice à l'expert. Cet échantillon significatif permettrait de réaliser en plus des tests d'équivalence [149] afin de savoir si les décisions de notre architecture sont équivalentes à celles qu'auraient prises les décideurs. Un test conçu sur le modèle du test de Turing [64] pourrait aussi être réalisé afin de voir si des experts sont capable de faire la distinction entre les réponses humaines et celles de l'aide à la décision.

La première comparaison réalisée sur la partie gestion de la route de l'architecture cognitive et A* a été simplifiée avec notamment la seule présence d'obstacles statiques et des conditions météorologiques stationnaires sur la zone. Même si ces résultats ont démontré que notre architecture était bien adaptée à la préconisation de decisions dans le cadre de la gestion de la route dans en temps contraint, il faudra ajouter dans la suite de cette comparaison les obstacles mobiles et des conditions météorologiques variables.

Perspectives d'application

Cette thèse possède plusieurs perspectives d'application, tout d'abord pour NAVAL GROUP qui possède déjà une solide culture en Machine Learning. NAVAL GROUP ayant vocation à réaliser dans les années à venir les navires de combat du futur, c'est-à-dire des navires plus performants, plus sécurisés et plus respectueux de l'environnement. Les résultats de cette thèse seront valorisés au sein des systèmes de navigation et de commandement de ces navires par (1) une meilleure connaissance de l'environnent, (2) une meilleure capacité à faire évoluer le navire dans un trafic maritime toujours plus dense et (3) répondre plus efficacement aux nouvelles menaces. L'assistance au commandement, telle que présentée et détaillée dans le cadre de cette thèse permettra d'améliorer les capacités de décision du commandement en proposant des indicateurs et scénarios de mission lui permettant (1) d'optimiser la route à suivre et (2) d'augmenter les capacités d'intervention par une meilleure gestion des ressources à mettre en œuvre.
Plus concrètement, l'intérêt de ces travaux pour NAVAL GROUP est d'être en mesure d'installer à bord des navires de la Marine Nationale des services orientés IA à forte valeur ajoutée opérationnelle. Dans cette optique, ces travaux de thèse ont déjà fait l'objet d'une première livraison dans le lac de données de NAVAL GROUP.

Un système d'aide à la conduite de missions navales dédié à la formation des officiers de la Marine Nationale peut également être envisagé. L'élève-officier pourrait utiliser les préconisations du système afin de se perfectionner dans la gestion d'une mission.

En plus d'être exploitable dans les systèmes de navigation et de commandement des navires de la Marine, les travaux de thèse peuvent être transférables au sein des drones et navire autonomes et être utilisés pour d'autres types de mission avec des objectifs variés. Le système

pourrait aussi être transposé à des missions sous-marines et intégrer certaines fonctions d'aide à la décision déjà existantes en interne [150], telle que l'aide à la reconnaissance de signaux acoustiques sous-marins [151]. Outre l'application de ces travaux au domaine militaire ils peuvent être appliqués aux missions navales civiles mais aussi à d'autres domaines tels que l'aérien ou à la protection d'infrastructures offshore et côtière.

Bibliographie

[1] « ENSEMBLE NOUS SOMMES MARINS | Marine Nationale ». https://www.etremarin.fr/

[2] N. Chapon, « Méthodologie décisionnelle en situation de crise ». MDSC 1 : LE GEO-SPIELE V°1.2, 2021.

[3] J. C. of Staff, « Commander's Handbook for Assessment Planning and Execution. » CreateSpace Independent Publishing Platform, 2012.

[4] J. van der Pligt, « Decision Making, Psychology of », in International Encyclopedia of the Social & Behavioral Sciences, N. J. Smelser et P. B. Baltes, Éd. Oxford : Pergamon, p. 3309-3315, 2001.

[5] G. Klein, « A Recognition Primed Decision (RPD) Model of Rapid Decision Making », in Decision making in action : models and methods, 1993.

[6] G. Phillips-Wren, N. Ichalkaranje, et L. Jain, « Intelligent Decision Making : An AI-Based Approach », Studies in Computational Intelligence., vol. 97. 2008.

[7] P.-M. Riccio et D. Bonnet, Éd., « TIC et innovation organisationnelle : Journées d'étude MTO'2011. » Paris : Presses des Mines, 2013.

[8] B. Roy et D. Bouyssou, « Aide Multicritère à la Décision : Méthodes et Cas », Economica. 1993.

[9] J. Wiener, « The Diffusion of Regulatory Oversight », Globalization of Cost-Benefit Analysis in Environmental Policy, p. 123-141, sept. 2012.

[10] M. D. Crossland, « Decision Support Systems », Encyclopedia of GIS, p. 232, 2008.

[11] M. Anderson et S. L. Anderson, « Machine Ethics: Creating an Ethical Intelligent Agent », AIMag, vol. 28, n° 4, p. 15, déc. 2007.

[12] L. S. Iliadis, « A decision support system applying an integrated fuzzy model for long-term forest fire risk estimation », Environmental Modelling & Software, vol. 20, n° 5, p. 613-621, mai 2005.

[13] M. H. Hassoun, « Fundamentals of Artificial Neural Networks », A Bradford Book. 2003.

[14] S. Liu, K. C. See, K. Y. Ngiam, L. A. Celi, X. Sun, et M. Feng, « Reinforcement Learning for Clinical Decision Support in Critical Care : Comprehensive Review », Journal of Medical Internet Research, vol. 22, n° 7, p. e18477, 2020.

[15] G. Phillips-Wren et L. Jain, « Artificial Intelligence for Decision Making », in Knowledge-Based Intelligent Information and Engineering Systems, Berlin, Heidelberg, p. 531-536, 2006.

[16] https://www.marines.mil/

[17] C. Dietrich, « Decision Making : Factors that Influence Decision Making, Heuristics Used, and Decision Outcomes », Inquiries Journal, vol. 2, n° 02, 2010.

[18] R. S. Sutton et A. G. Barto, « Reinforcement Learning : an introduction », Second edition. Cambridge, Massachusetts : The MIT Press, 2018.

[19] W. Schultz, P. Dayan, et P. R. Montague, « A neural substrate of prediction and reward », Science, vol. 275, n° 5306, p. 1593-1599, mars 1997.

[20] C. Iphar, A. Napoli, et C. Ray, « An expert-based method for the risk assessment of anomalous maritime transportation data », Applied Ocean Research, vol. 104, p. 102337, nov. 2020.

[21] C. Iphar, C. Ray, et A. Napoli, « Data integrity assessment for maritime anomaly detection », Expert Systems with Applications, vol. 147, p. 113219, juin 2020.

[22] A. Vandecasteele, R. Devillers, et A. Napoli, « A semi-supervised learning framework based on spatio-temporal semantic events for maritime anomaly detection and behavior analysis », in CoastGIS 2013 - The 11th International Symposium for GIS and Computer Cartography for Coastal Zone Management, Victoria, Canada, 4 pages, juin 2013.
[23] M. Riveiro, G. Pallotta, et M. Vespe, « Maritime anomaly detection : A review », Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery, vol. 8, p. e1266, mai 2018.
[24] J. Roy et M. Davenport, « Categorization of maritime anomalies for notification and alerting purpose », NATO workshop on data fusion and anomaly detection for maritime situational awareness, La Spezia, Italy, 2009.
[25] C. Iphar, « Formalisation d'un environnement d'analyse des données basé sur la détection d'anomalies pour l'évaluation de risques : Application à la connaissance de la situation maritime », These de doctorat, Paris Sciences et Lettres (ComUE), 2017.
[26] L. Brumley, C. Kopp, et K. Korb, « The Orientation step of the OODA loop and Information Warfare », Proceedings of the 7th Australian Information Warfare and Security Conference, p. 18-25, janv. 2006.
[27] J.-F. Lebraty, « La maîtrise de l'information source de souveraineté : le cas des C4iSR », Prospective et stratégie, vol. Numéro1, n° 1, p. 103, 2010.
[28] L. J. Savage, « The foundations of statistics. » Oxford, England : John Wiley & Sons, 1954.
[29] J. Rodman, « Cognitive biases and decision making : a literature review and discussion of implications for the US Army. », Human Dimension Capabilities Development Task Force (HDCTF), White Paper. 2015.
[30] D. Dubois, H. Prade, et P. Smets, « Representing partial ignorance », Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, vol. 26, p. 361-377, juin 1996.
[31] B. Bouchon-Meunier et C. Marsala, « Logique floue : principes, aide à la décision. » Hermès-Lavoisier, 2003.
[32] H. A. Simon, « Rational choice and the structure of the environment. », Psychological Review, vol. 63, n° 2, p. 129-138, 1956.
[33] H. A. Simon, « The New Science of Management Decision. » USA: Prentice Hall PTR, 1977.
[34] J.-C. Pomerol, « Scenario development and practical decision making under uncertainty », Decision Support Systems, vol. 31, n° 2, p. 197-204, juin 2001.
[35] C. E. Zsambok et G. Klein, « Naturalistic Decision Making. » Psychology Press, 2014.
[36] W. Hardy, « Human Dimension Capabilities Development Task Force (HDCDTF) Mission Command » - Capabilities Development and Integration Directorate (CDID) 806 Harrison Drive Building 470 Fort Leavenworth, KS 66027-2302 913-684-452, p. 32, 2016.
[37] W. Hardy, « Social intelligence : assessment and training. » Human Dimension Capabilities Development Task Force- Capabilities Development Integration Directorate - Mission Command Center of Excellence (MC CoE), 2016.
[38] J. Schmitt et G. Klein, « A Recognitional Planning Model », klein associates inc fairborn oh, janv. 1999.
[39] K. G. Ross, G. A. Klein, P. Thunholm, J. F. Schmitt, et H. C. Baxter, « The Recognition-Primed Decision Model », army combined arms center fort leavenworth ks military review, août 2004.

[40] J. Sokolowski, « Representing Knowledge and Experience in RPDAgent », Proceedings of the 12th Conference on Behavior Representation in Modeling and Simulation (BRIMS), p. 5, mai 2003.
[41] L. Dodd, J. Moffat, J. Smith, et G. Mathieson, « From Simple Prescriptive to Complex Descriptive Models : An Example from a Recent Command Decision Experiment », 8th International Command and Control Research & Technology Symposium, p. 41, juin 2003.
[42] G. Klein, « Sources of Power : How People Make Decisions », Leadership and Management in Engineering., vol. 1. 2001.
[43] M. Slavkovik et G. Boella, « Recognition-primed group decisions via judgement aggregation », Synthese, vol. 189, n° 1, p. 51-65, déc. 2012.
[44] J. Rodman, « Social intelligence : introduction and overview for the Army's human dimension initiative. » Disponible sur : US Army Combined Arms Center (USACAC) Repository, 2016.
[45] M. J. F. Schmitt, « How We Decide », MCA. https://mca-marines.org/blog/gazette/how-we-decide/, 1995.
[46] J. Simonsen, « Herbert A. Simon : Administrative Behavior -How Organizations Can Be Understood in Terms of Decision Processes », Computer Science, Roskilde University, p. 12, 2004.
[47] J. P. Kahan, D. R. Worley, et C. Stasz, « Understandig commanders' information needs. » Santa Monica, Calif: rand arroyo center, 1989.
[48] J. Payne, J. Bettman, et E. Johnson, « Behavioral Decision Research: A Constructive Processing Perspective », Annual Review of Psychology, vol. 43, p. 87-131, nov. 2003.
[49] J. E. Hummel et K. J. Holyoak, « A symbolic-connectionist theory of relational inference and generalization », Psychological Review, vol. 110, p. 220-264, 2003.
[50] I. L. Janis et L. Mann, « Decision making: A psychological analysis of conflict, choice, and commitment. » New York, NY, US: Free Press, 1977.
[51] J. Yen, X. Fan, S. Sun, et M. McNeese, « Agent-based collaborative recognition-primed decision-making », US8442839B2, 14 mai 2013
[52] G. Kaempf, G. Klein, M. Thordsen, et S. Wolf, « Decision Making in Complex Naval Command-and-Control Environments », Human Factors, vol. 38, p. 220-231, 1996.
[53] M. Merad, N. Dechy, L. Dehouck, et M. Lassagne, « Les décisions face aux risques majeurs. Retours d'expériences et pistes d'amélioration », Congrès Lambda Mu 19 de Maîtrise des Risques et Sûreté de Fonctionnement, oct. 2014.
[54] J. Laird, C. Lebiere, et P. Rosenbloom, « A Standard Model of the Mind: Toward a Common Computational Framework across Artificial Intelligence, Cognitive Science, Neuroscience, and Robotics », AI Magazine, vol. 38, p. 13, déc. 2017.
[55] Wikistat, « Statistique & Machine Learning de Statisticien à Data Scientist », http://wikistat.fr/pdf/st-m-explo-classif.pdf, 2016.
[56] Q. Liu et Y. Wu, « Supervised Learning », in Encyclopedia of the Sciences of Learning, N. M. Seel, Éd. Boston, MA: Springer US p. 3243-3245, 2012.
[57] L. A. Zadeh, K.-S. Fu, K. Tanaka, et M. Shimura, « Fuzzy sets and their applications to cognitive and decision processes », 1st Edition edition. New York: Academic Press, 1975.
[58] J. Sokolowski et C. Banks, « Handbook of Real-World Applications in Modeling and Simulation », VMASC Books. Hoboken, NJ: John Wiley & Sons, Inc., 2012.
[59] W. Warwick, S. Quesada, R. Hutton, et P. Mcdermott, « Developing computational models of recognition-primed decision making », Proceedings of the Tenth Conference on Computer Generated Forces, 2001.

[60] B. Gevarter, « MoCogl: A Computer Simulation of Recognition-Primed Human », Decision Making, p. 22, 1991.
[61] J. Hiles, M. VanPutte, et B. Osborne, « Innovations in Computer Generated Autonomy at the MOVES Institute », Naval Postgraduate School Monterey CA, NPS-MV-02-002, déc. 2001.
[62] J. A. Sokolowski, « Enhanced Military Decision Modeling Using a MultiAgent System Approach », SIMULATION : Transactions of The Society for Modeling and Simulation International, p. 79(4) : 232-242, 2003.
[63] S. E. Gordon, « Cognitive Task Analysis Using Complementary Elicitation Methods », Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 39, nº 9, p. 525-529, 1995.
[64] A. M. Turing, « I.—COMPUTING MACHINERY AND INTELLIGENCE », Mind, vol. LIX, nº 236, p. 433-460, 1950.
[65] C. Adam et B. Gaudou, « BDI agents in social simulations: a survey », Knowledge Engineering Review, vol. 31, nº n° 3, p. 207-238, juin 2016.
[66] P. Busetta, R. Ronnquist, A. Hodgson, et A. Lucas, « JACK Intelligent Agents - Components for Intelligent Agents in Java », Melbourne, Australia, 1999.
[67] R. H. Bordini, J. F. Hübner, et M. Wooldridge, « Programming Multi-Agent Systems in AgentSpeak Using Jason (Wiley Series in Agent Technology). » Hoboken, NJ, USA: John Wiley & Sons, Inc., 2007.
[68] K. Hindriks, F. Boer, W. Hoek, et J. Meyer, « Agent programming in 3APL », Autonomous Agents and Multi-Agent Systems, vol. 2, p. 357-401, 1999.
[69] F. Lui, R. Connell, J. Vaughan, D. Jarvis, et J. Jarvis, « An Architecture to Support Autonomous Command Agents for OneSAF Testbed Simulations », 2002.
[70] S. N. Danial, J. Smith, B. Veitch, et F. Khan, « On the realization of the recognition-primed decision model for artificial agents », Hum. Cent. Comput. Inf. Sci., vol. 9, nº 1, p. 36, déc. 2019.
[71] L. De Raedt, K. Kersting, S. Natarajan, et D. Poole, « Statistical Relational Artificial Intelligence: Logic, Probability, and Computation », Synthesis Lectures on Artificial Intelligence and Machine Learning, vol. 10, p. 1-189, mars 2016.
[72] J. Smith, « The effect of virtual environment training on participant competence and learning in offshore emergency egress scenarios », Master of Engineering Faculty of Engineering and Applied Science, Memorial University of Newfoundland, 2015.
[73] S. Mueller, « A Bayesian Recognitional Decision Model », Journal of Cognitive Engineering and Decision Making, vol. 3, p. 111-130, 2009.
[74] S. A. Converse, J. A. Cannon-Bowers, et E. Salas, « Team Member Shared Mental Models: A Theory and Some Methodological Issues », Proceedings of the Human Factors Society Annual Meeting, vol. 35, nº 19, p. 1417-1421, sept. 1991.
[75] X. Fan, J. Yen, et R. A. Volz, « A theoretical framework on proactive information exchange in agent teamwork », Artificial Intelligence, vol. 169, nº 1, p. 23-97, nov. 2005.
[76] X. Fan, « R-CAST: Integrating Team Intelligence for Human-Centered Teamwork », Conference: Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence, p. 7, juill. 2007.
[77] D. Serfaty, E. E. Entin, et J. H. Johnston, « Team coordination training. », in Making decisions under stress: Implications for individual and team training., Washington, DC, US: American Psychological Association, , p. 221-245, 1998.
[78] X. Fan, J. Yen, et R. A. Volz, « A theoretical framework on proactive information exchange in agent teamwork », Artif Intell, vol. 169, nº 1, p. 23-97, nov. 2005.

[79] T. L. Pors, « Simulation cognitive de la prise de décision d'experts ; application au trafic maritime. », Phd thesis, Université de Bretagne Sud, 2010.
[80] T. Le Pors, T. Devogele, et C. Chauvin, « Multi agent system integrating Naturalistic Decision Roles: application to maritime traffic », IADIS International Conference Intelligent Systems and Agents, p. pp.100-107, juin 2009.
[81] S. Fournier, C. Claramunt, et T. Devogele, « TRANS : A Tractable Rolebased Agent Prototype for Concurrent Navigation System », The first European workshop on Multi-Agent Systems, janv. 2003.
[82] D. Kunde et C. J. Darken, « A Mental Simulation-Based Decision-Making Architecture Applied to Ground Combat », Proceedings of BRIMS 2006, 2006.
[83] D. Kunde, « Event Prediction for Modeling Mental Simulation in Naturalistic Decision Making », Phd thesis, Naval PostGraduate School, Monterey CA, 2005.
[84] D. Kunde, G. Army, et C. J. Darken, « Event prediction for modeling mental simulation in naturalistic decision making », Proceedings of BRIMS 2005, 2005.
[85] B. Sericola, « Chaînes de Markov: théorie, algorithmes et applications », Lavoisier. 2013.
[86] C. C. Ong, « Analysis of Cognitive Architecture in the Cultural Geography Mode », Homeland Security Digital Library, 1 septembre 2012.
[87] P. Sotirios et S. Papadopoulos, « Reinforcement learning : a new approach for the cultural geography model », Naval PostGraduate School, Monterey CA, 2010.
[88] C. Watkins et P. Dayan, « Q-learning », Mach Learn, vol. 8, nº 3, p. 279-292, mai 1992.
[89] T. H. Cormen, « Introduction à l'algorithmique: cours et exercices. » Paris: Dunod, 2004.
[90] F. Moisescu, M. Bo, G. Prelipcean, et M. Lupan, « INTELLIGENT AGENTS IN MILITARY DECISION MAKING », Science & Military, p. 7, 2010.
[91] T. Phung, M. Winikoff, et L. Padgham, « Learning Within the BDI Framework: An Empirical Analysis », Knowledge-Based Intelligent Information and Engineering Systems, 9th International Conference, p. 282-288, 2005.
[92] S. Sardina, L. de Silva, et L. Padgham, « Hierarchical Planning in BDI Agent Programming Languages: A Formal Approach », 5th International Joint Conference on Autonomous Agents and Multiagent Systems, p. 8, 2006.
[93] M. H. Hassoun, « Fundamentals of Artificial Neural Networks », Proceedings of the IEEE, vol. 84, nº 6, p. 906-, juin 1996.
[94] Y. Rocha et T.-Y. Kuc, « Mental Simulation for Autonomous Learning and Planning Based on Triplet Ontological Semantic Model », The 1st International Workshop on the Semantic Descriptor, Semantic Modelingand Mapping for Humanlike Perceptionand Navigation of Mobile Robots toward Large Scale Long-Term Autonomy (SDMM19), nov. 2019.
[95] J. K. Alt, « Learning from Noisy and Delayed Rewards The Value of Reinforcement Learning to Defense Modeling and Simulation. » Naval Postgraduate School, Monterey, California, 2012.
[96] L. Matignon, « Synthèse d'agents adaptatifs et coopératifs par apprentissage par renforcement.Application à la commande d'un système distribué de micromanipulation. », Phd thesis, U.F.R des sciences et techniques de l'université de Franche-Comté, 2008.
[97] B. F. Skinner, « Science And Human Behavior. » Simon and Schuster, 1965.
[98] I. P. Pavlov, « Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex. » Oxford, England: Oxford Univ. Press, p. xv, 430, 1927.
[99] E. L. Thorndike, « Animal intelligence: experimental studies », Macmillan Press. 1911.

[100] B. F. Skinner, « The Behavior of Organisms: An Experimental Analysis. » B. F. Skinner Foundation, 2019.
[101] D. Michie et R. A. Chambers, « BOXES : AN EXPERIMENT IN ADAPTIVE CONTROL », Department of machine intelligence and perception, univeristy of Edinburgh, 2013.
[102] A. G. Barto, R. S. Sutton, et C. W. Anderson, « Neuronlike Adaptive Elements That Can Solve Difficult Learning Control Problems », in Artificial Neural Networks: Concept Learning, IEEE Press, p. 81-93, 1990.
[103] R. S. Sutton, « Learning to predict by the methods of temporal differences », Machine Learning, vol. 3, n° 1, p. 9-44, août 1988.
[104] J. Cohen, S. M. McClure, et A. Yu, « Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration », Philosophical transactions of the Royal Society of London. Series B, Biological sciences, vol. 362, p. 933-42, 2007.
[105] G. Story, I. Vlaev, B. Seymour, A. Darzi, et R. Dolan, « Does temporal discounting explain unhealthy behavior? A systematic review and reinforcement learning perspective », Frontiers in behavioral neuroscience, vol. 8, p. 76, 2014.
[106] W. Krämer et D. Kahneman, « Thinking, Fast and Slow », Statistical Papers, vol. 55, 2014.
[107] V. Dutt, « Explaining Human Behavior in Dynamic Tasks through Reinforcement Learning », Journal of Advances in Information Technology, vol. 2, août 2011.
[108] S. J. Russell et P. Norvig, « Artificial intelligence : a modern approach. » Third edition. Upper Saddle River, N.J. : Prentice Hall, ©2010, 2010.
[109] M. Han, « Reinforcement Learning Approaches in Dynamic Environments », Phd thesis, Télécom ParisTech, 2018.
[110] A. G. Barto, S. J. Bradtke, S. P. Singh, T. T. R. Yee, V. Gullapalli, et B. Pinette, « Learning to Act using Real-Time Dynamic Programming », Artificial Intelligence, vol. 72, p. 81-138, 1995.
[111] « Optimization by Simulated Annealing | Science ». https://science.sciencemag.org/content/220/4598/671, 1983.
[112] S. Ray et P. Tadepalli, « Model-Based Reinforcement Learning », in Encyclopedia of Machine Learning, C. Sammut et G. I. Webb, Éd. Boston, MA: Springer US, p. 690-693, 2010,.
[113] R. Bellman, « Dynamic programming. » Princeton, NJ: Princeton Univ. Pr, 1984.
[114] C. Watkins, « Learning From Delayed Rewards », Phd thesis, King's College, Cambridge, Massachusetts, 1989.
[115] R. J. Williams, « Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning », Machine Learning, vol. 8, p. 229-256, 1992.
[116] C. Lemaréchal, « Cauchy and the Gradient Method », vol. Extra Volume ISMP (2012) 251–254. 2012.
[117] V. Mnih et al., « Playing Atari with Deep Reinforcement Learning », arXiv:1312.5602v1 [cs.LG], 19 décembre 2013.
[118] H. van Hasselt, A. Guez, et D. Silver, « Deep Reinforcement Learning with Double Q-learning », arXiv:1509.06461 [cs], déc. 2015.
[119] B. Manela, « Deep Reinforcement Learning for Complex Manipulation Tasks with Sparse Feedback », arXiv:2001.03877 [cs, stat], janv. 2020.
[120] V. Mnih et al., « Asynchronous Methods for Deep Reinforcement Learning », arXiv:1602.01783 [cs], juin 2016.

[121] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, et O. Klimov, « Proximal Policy Optimization Algorithms », arXiv:1707.06347 [cs], août 2017.
[122] K. Shao, D. Zhao, N. Li, et Y. Zhu, « Learning Battles in ViZDoom via Deep Reinforcement Learning », 2018 IEEE Conference on Computational Intelligence and Games (CIG), p. 1-4., août 2018.
[123] J. Myung et M. Pitt, « Model Comparison Methods », Methods in enzymology, vol. 383, p. 351-66, févr. 2004.
[124] V. Novák, I. Perfilieva, et J. Mockor, « Mathematical Principles of Fuzzy Logic » , Springer Science&Business Media. 1999.
[125] C. Dumortier, « Amélioration de la performance énergétique des bateaux civils et militaires », These de doctorat, Université de Bordeaux, 2020.
[126] F. Dupriez-Robin, « Dimensionnement d'une propulsion hybride de voilier, basé sur la modélisation par les flux de puissance », Thèse de Doctorat de l'Université de Nantes, école Polytechnique de l'université de Nantes, St Nazaire, 2010.
[127] C. Caplier, « Etude expérimentale des effets de hauteur d'eau finie, de confinement latéral et de courant sur les sillages et la résistance à l'avancement des navires. », Thèse de doctorat, UFR des sciences fondamentales et appliquées Pôle poitevin de recherche pour l'ingénieur en mécanique, matériaux et énergétique - PPRIMME, Poitiers, 2015.
[128] Tin Kam Ho, « Random decision forests », in Proceedings of 3rd International Conference on Document Analysis and Recognition, vol. 1, p. 278-282, août 1995.
[129] Z. Feng, H. Yang, X. Li, L. Yan, Z. Liu, et W. Liu, « Real-Time Vessel Trajectory Data-Based Collison Risk Assessment in Crowded Inland Waterways », IEEE 4th International Conference on Big Data Analytics (ICBDA), p. 134, mars 2019.
[130] A. van den Bosch, « Hidden Markov Models », in Encyclopedia of Machine Learning, C. Sammut et G. I. Webb, Éd. Boston, MA: Springer US, p. 493-495, 2010.
[131] R. E. Kalman, « A New Approach to Linear Filtering and Prediction Problems », Journal of Basic Engineering, vol. 82, n° 1, p. 35-45, mars 1960.
[132] « Proximal Policy Optimization — Spinning Up documentation ». https://spinningup.openai.com/en/latest/algorithms/ppo.html, 2018.
[133] E. Kindler et I. Krivý, « Object-oriented simulation of systems with sophisticated control », International Journal of General Systems - INT J GEN SYSTEM, vol. 40, p. 313-343, janv. 2005.
[134] « ECMWF », ECMWF. https://www.ecmwf.int/
[135] « netCDF4 API documentation ». https://unidata.github.io/netcdf4-python/
[136] « About us — scikit-learn 0.24.2 documentation ». https://scikit-learn.org/stable/about.html
[137] G. Brockman et al., « OpenAI Gym », arXiv:1606.01540 [cs], juin 2016.
[138] « API Documentation | TensorFlow Core v2.5.0 ». https://www.tensorflow.org/api_docs
[139] « Welcome to Stable Baselines docs! - RL Baselines Made Easy — Stable Baselines 2.10.2 documentation ». https://stable-baselines.readthedocs.io/en/master/
[140] T. Haarnoja, A. Zhou, P. Abbeel, et S. Levine, « Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor », arXiv:1801.01290 [cs, stat], août 2018.
[141] E. ARTUSI, « Ship path planning based on Deep Reinforcement Learning and weather forecast », in 2021 22nd IEEE International Conference on Mobile Data Management (MDM), , p. 258-260, 2021.

[142] P. E. Hart, N. J. Nilsson, et B. Raphael, « A Formal Basis for the Heuristic Determination of Minimum Cost Paths », IEEE Transactions on Systems Science and Cybernetics, vol. 4, nº 2, p. 100-107, juill. 1968.
[143] E. ARTUSI, F. Chaillan, et A. Napoli, « Path planning for a maritime surface ship based on Deep Reinforcement Learning and weather forecast », San Diego, United States, sept. 2021.
[144] « Project Jupyter ». https://www.jupyter.org
[145] « Convention on the International Regulations for Preventing Collisions at Sea », Centre for International Law. https://cil.nus.edu.sg/databasecil/1972-convention-on-the-international-regulations-for-preventing-collisions-at-sea/, 1972.
[146] A. Vandecasteele, R. Devillers, et A. Napoli, « From Movement Data to Objects Behavior Using Semantic Trajectory and Semantic Events », Marine Geodesy, vol. 37, nº 2, p. 126-144, avr. 2014.
[147] B. Hengst, « Hierarchical Reinforcement Learning », in Encyclopedia of Machine Learning, C. Sammut et G. I. Webb, Éd. Boston, MA: Springer US, 2010, p. 495-502.
[148] M. T. J. Spaan, « Partially Observable Markov Decision Processes », in Reinforcement Learning: State-of-the-Art, M. Wiering et M. van Otterlo, Éd. Berlin, Heidelberg: Springer, p. 387-414, 2012.
[149] J. Rogers, K. Howard, et J. Vessey, « Using significance tests to evaluate equivalence between two experimental groups », Psychological bulletin, vol. 113, p. 553-65, juin 1993.
[150] E. Chauveau, C. Lesire, et F. Chaillan, « Integration of Autonomous Heterogeneous Systems for Decision Making Autonomy in Naval Defence: A Position Paper », in OCEANS 2019 - Marseille, p. 1-6, juin 2019.
[151] E. Artusi et F. Chaillan, « Automatic recognition of underwater acoustic signature for naval applications », 1st Maritime Situational Awareness Workshop MSAW 2019, oct. 2019.

ANNEXE 1 : Représentation des données cartographiques

Les données cartographiques sont exprimées en latitude et longitude selon un standard connu tel que par exemple le WGS84. Le domaine marin étudié est défini par $[\mathrm{Lat}_{min}; \mathrm{Lat}_{max}] \times [\mathrm{Lon}_{min}; \mathrm{Lon}_{max}]$.

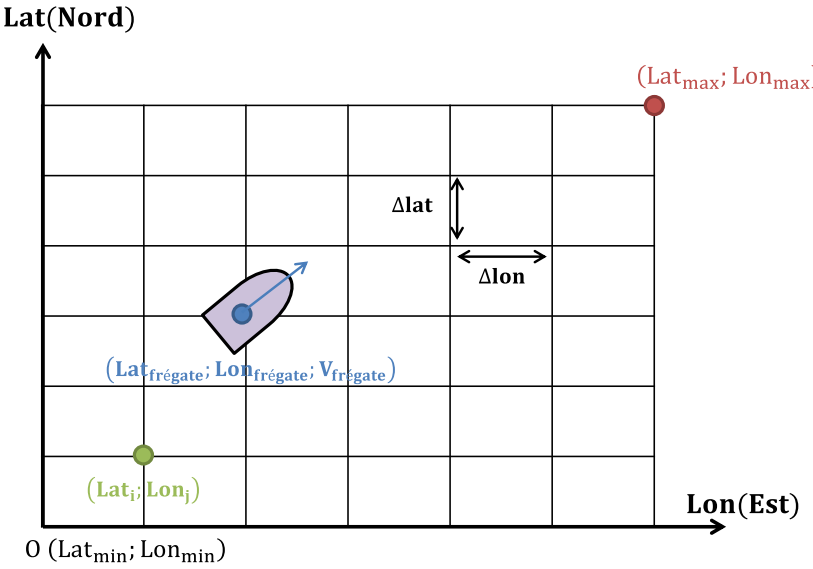

Figure Annexe-1 : Grille de déplacement de la frégate

Ce domaine est maillé régulièrement de sorte que $\forall i = 1 \dots N_{\mathrm{Lat}}$ et $\forall j \in 1 \dots N_{\mathrm{Lon}}$, tel qu'illustré par la Figure Annexe-1:

$$\begin{cases} \mathrm{Lat}_i &= \mathrm{Lat}_{min} + (i-1)\Delta\mathrm{Lat} \\ \mathrm{Lon}_j &= \mathrm{Lon}_{min} + (j-1)\Delta\mathrm{Lon} \end{cases} \quad (1)$$

Les pas de maille $\Delta\mathrm{Lat}$ et $\Delta\mathrm{Lon}$ sont des paramètres donnés, si bien que le nombre de points N_{Lat} et N_{Lon} du maillage sont respectivement donnés par :

$$\begin{cases} N_{\mathrm{Lat}} &= \left\lfloor \dfrac{\mathrm{Lat}_{max} - \mathrm{Lat}_{min}}{\Delta\mathrm{Lat}} + 1 \right\rfloor \\ N_{\mathrm{Lon}} &= \left\lfloor \dfrac{\mathrm{Lon}_{max} - \mathrm{Lon}_{min}}{\Delta\mathrm{Lon}} + 1 \right\rfloor \end{cases} \quad (2)$$

Dans ces conditions, il existe $N_{\mathrm{Lat}} \times N_{\mathrm{Lon}}$ positions possibles dans le domaine étudié. Les positions de la frégate sont représentées dans ce repère.

Les données météo sont généralement données avec un pas de maille plus large que celui du maillage du domaine, de sorte que :

$$\begin{cases} \Delta\mathrm{Lat}_{meteo} &> \Delta\mathrm{Lat} \\ \Delta\mathrm{Lon}_{meteo} &> \Delta\mathrm{Lon} \end{cases} \quad (3)$$

Dans ces conditions, une valeur de donnée météo $\mathcal{M}(\mathrm{Lat}_{meteo}; \mathrm{Lon}_{meteo})$ est positionnée sur la grille du repère par interpolation au plus proche voisin :

$$\begin{cases} i_{meteo} & = & \underset{i=1\ldots N_{\mathrm{Lat}}}{\operatorname{argmin}} \left[\left| \mathrm{Lat}_{min} + (i-1)\Delta\mathrm{Lat} - \mathrm{Lat}_{meteo} \right| \right] \\ j_{meteo} & = & \underset{j=1\ldots N_{\mathrm{Lon}}}{\operatorname{argmin}} \left[\left| \mathrm{Lon}_{min} + (j-1)\Delta\mathrm{Lon} - \mathrm{Lon}_{meteo} \right| \right] \end{cases} \quad (4)$$

De sorte que la donnée $\mathcal{M}(\mathrm{Lat}_{meteo}; \mathrm{Lon}_{meteo})$ est représentée dans la grille du domaine par la maille $(i_{meteo}; j_{meteo})$ avec :

$$\begin{cases} \mathrm{Lat}_{meteo} & \approx & \mathrm{Lat}_{min} + (i_{meteo} - 1)\Delta\mathrm{Lat} \\ \mathrm{Lon}_{meteo} & \approx & \mathrm{Lon}_{min} + (j_{meteo} - 1)\Delta\mathrm{Lon} \end{cases} \quad (6)$$

La frégate est supposée évoluer localement à l'intérieur du plan tangent au point de contact$(\mathrm{Lat}_{min}; \mathrm{Lon}_{min})$. Ce plan est muni d'un repère local $(O; x; y)$ lié à la Terre dans le référentiel terrestre, dont l'origine O coïncide avec le point de contact, et d'un repère lié à la frégate dans le référentiel terrestre noté $(G; x_G; y_G)$, présenté en Figure Annexe-2.

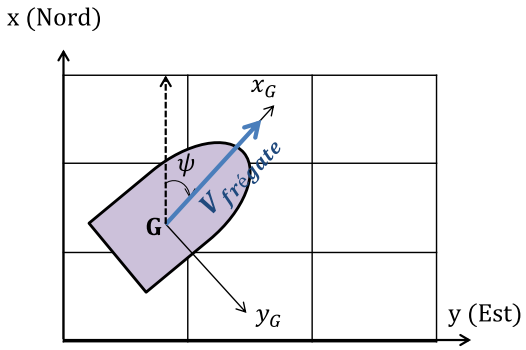

Figure Annexe-2 : Convention à utiliser pour l'affichage des données initiales

Dans ce repère localement euclidien, si ψ est le cap en degrés de la frégate et que la vitesse est donnée par le vecteur $\vec{V}_{frégate}$, alors :

$$\vec{V}_{frégate} = \begin{bmatrix} V_{frégate} \cos\left(\frac{\pi}{180}\psi\right) \\ V_{frégate} \sin\left(\frac{\pi}{180}\psi\right) \end{bmatrix} \quad (7)$$

La frégate est supposée évoluer en mouvement rectiligne uniforme par morceau à partir du point de départ de coordonnées dans le repère local $(x_{origine}; y_{origine})$, si bien qu'au cours du temps sa position est donnée $\forall t \in [0; T]$ par :

$$\begin{cases} x_{frégate}(t) & = & x_{origine} + V_{frégate} \cos\left(\frac{\pi}{180}\psi\right) t \\ y_{frégate}(t) & = & y_{origine} + V_{frégate} \sin\left(\frac{\pi}{180}\psi\right) t \end{cases} \quad (8)$$

Pour ramener les coordonnées de la frégate du repère local exprimées en mètres vers le repère natif exprimées en angles, nous utilisons la règle de conversion locale suivante qui à partir de tout couple $(\mathrm{Lat}\,;\mathrm{Lon})$ établit la correspondance approximative suivante :

$$\begin{cases} 1°_{\mathrm{Lat}} & \approx & 111110\ m \\ 1°_{\mathrm{Lon}} & \approx & \cos\left(\mathrm{Lat}\dfrac{\pi}{180}\right) \times 111110\ m \end{cases} \quad (9)$$

Ce qui conduit en pratique à l'approximation de la position de la frégate sur la grille :

$$\begin{cases} \mathrm{Lat}(t) & \approx & \dfrac{x(t)}{111110} \\ \mathrm{Lon}(t) & \approx & \dfrac{y(t)}{\cos\left(\frac{\mathrm{Lat}(t)\pi}{180}\right) \times 111110} \end{cases} \quad (10)$$

Ainsi, la position de la frégate en tout point de la grille du domaine est exprimée à partir de sa position dans le repère local par interpolation au plus proche voisin :

$$\begin{cases} i_{frégate}(t) & = & \underset{i=1\ldots N_{\mathrm{Lat}}}{\mathrm{argmin}} \left[\left|\mathrm{Lat}_{min} + (i-1)\Delta\mathrm{Lat} - \mathrm{Lat}(t)\right|\right] \\ j_{frégate}(t) & = & \underset{j=1\ldots N_{\mathrm{Lon}}}{\mathrm{argmin}} \left[\left|\mathrm{Lon}_{min} + (j-1)\Delta\mathrm{Lon} - \mathrm{Lon}(t)\right|\right] \end{cases} \quad (11)$$

De sorte que la position courante de la frégate est représentée dans la grille du domaine par la maille $\left(i_{frégate}(t); j_{frégate}(t)\right)$ avec :

$$\begin{cases} \mathrm{Lat}_{frégate}(t) & \approx & \mathrm{Lat}_{min} + \left(i_{frégate}(t) - 1\right)\Delta\mathrm{Lat} \\ \mathrm{Lon}_{frégate}(t) & \approx & \mathrm{Lon}_{min} + \left(j_{frégate}(t) - 1\right)\Delta\mathrm{Lon} \end{cases} \quad (12)$$

RÉSUMÉ

La conduite de missions de la Marine Nationale nécessite d'intégrer simultanément un bon nombre d'informations relatives au navire et à l'environnement dans lequel il évolue. L'organe de commandement d'un navire analyse ces informations afin de prendre les décisions adaptées à la situation pour mener à bien la mission. Cependant, les capacités humaines soumises à leurs propres limites ne sont plus suffisantes pour appliquer de manière adéquate et rapide les méthodologies mises à leur disposition dans un environnement incertain et contraint par le temps. Pour faire face à cette situation, la conception et le développement d'un système d'aide à la décision permettrait grâce à la combinaison des sciences cognitives et de la puissance de calcul des moyens informatiques modernes de réduire les temps (1) d'analyse de la situation et (2) de sélection d'une action appropriée à la situation. Les travaux de thèse vont ainsi se fonder sur une méthode de prise de décision issue du Natural Decision Making appelée Recognition Primed Decision (RPD), utilisée pour la prise de décision dans l'urgence.

L'objet de cette thèse a donc été de formaliser un système d'aide à la décision et plus particulièrement une architecture cognitive fondée sur le RPD et capable de réagir dans un environnement dynamique avec des situations inédites. La conception de l'aide s'est inspirée de l'architecture cognitive de Kunde et Darken et s'est appuyée sur l'apprentissage par renforcement profond.

MOTS CLÉS

Architecture cognitive, Recognition Primed Decision, apprentissage par renforcement profond, mission navale tactique, système d'aide à la décision, Intelligence Artificielle

ABSTRACT

Information analysis related to a ship and its environment is required in order to make the appropriate decisions during naval missions. However, human capabilities are no longer sufficient to reliably and rapidly apply methodologies at their disposal in an uncertain and time-constrained environment. Decision-makers face with uncertain situations where they lack time and resources. To deal with this situation, the design and the development of a decision support system would make it possible, thanks to the combination of cognitive sciences and the computing power of modern IT resources, to reduce the time required (1) to analyse the situation and (2) to select an appropriate action for the situation. The Phd work is based on a decision-making method from Natural Decision Making (NDM) called Recognition Primed Decision (RPD), used for emergency decision-making.

My proposal is therefore to formalize a decision support system and more specifically a cognitive architecture based on RPD and capable of reacting in a dynamic environment with novel situations. The design of the aid will be inspired by the cognitive architecture of Kunde et Darken and will be based on deep reinforcement learning.

KEYWORDS

Cognitive architecture, Recognition Primed Decision, deep reinforcement learning, tactic naval mission, decision support system, Artificial Intelligence