Explicabilité en Apprentissage par Renforcement : vers une Taxinomie Unifiée

Maxime Alaarabiou, Nicolas Delestre, Laurent Vercouter

Explicabilité en Apprentissage par Renforcement : vers une Taxinomie Unifiée

Maxime Alaarabiou[1] **Nicolas Delestre**[1] **Laurent Vercouter**[1]

[1] INSA Rouen Normandie, Normandie Univ, LITIS UR 4108, F-76000 Rouen, France

maxime.alaarabiou@insa-rouen.fr
nicolas.delestre@insa-rouen.fr
laurent.vercouter@insa-rouen.fr

Résumé

La problématique de l'explicabilité est à l'heure actuelle un enjeu important en intelligence artificielle, et plus spécifiquement en apprentissage par renforcement. Dans cet article, nous proposons une nouvelle classification des techniques d'explicabilité pour l'apprentissage par renforcement. Pour se faire, nous nous appuyons sur les aspects spécifiques de l'apprentissage par renforcement en définissant les concepts d'explication, de source d'explication et de technique d'explicabilité. En utilisant cette nouvelle classification, nous effectuons une analyse de l'état de l'art des techniques existantes dans ce domaine. Enfin, nous proposons une rétrospective de l'évolution des systèmes de classification, mettant en lumière les avancées et les tendances récentes en apprentissage par renforcement explicable.

Abstract

The issue of explainability is a current important concern in artificial intelligence, specifically in reinforcement learning. In this article, we propose a new classification of explainable reinforcement learning techniques. To do so, we rely on specific aspects of reinforcement learning by defining the concepts of explanation, source of explanation, and explicability technique. Using this new classification, we conduct an analysis of the state-of-the-art techniques existing in this field. Finally, we offer a retrospective on the evolution of classification systems, highlighting recent advances and trends in this domain.

1 Introduction

L'apprentissage par renforcement (*Reinforcement Learning*, RL) est un champ de l'intelligence artificielle (*Artificial Intelligence*, AI), et plus particulièrement de l'apprentissage machine, dans lequel on cherche à créer un agent qui maximise sa récompense dans un environnement [31]. L'intervention humaine se limite au strict minimum dans le processus d'apprentissage. L'agent doit donc développer une stratégie dans l'environnement en passant par une phase d'exploration non guidée. Pendant cette phase d'exploration, l'agent expérimente par essais-erreurs afin de mettre au point une fonction politique pertinente. Cette fonction a pour objectif d'établir une correspondance entre l'observation et les actions de l'agent, de sorte qu'une succession de décisions suivant cette politique maximise l'espérance de la somme des récompenses. Pour améliorer cette politique, on apprend la fonction de critique qui pour une observation donnée, prédit l'espérance de la somme des récompenses.

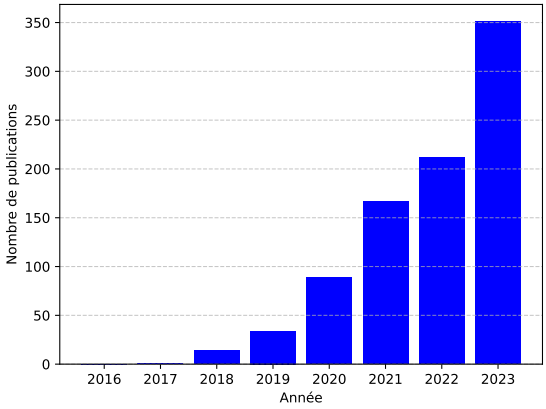

FIGURE 1 – Évolution du nombre de publications sur arXiv ayant le mot-clé "XAI" dans leur résumé.

Cette méthode d'entraînement a obtenu beaucoup de succès depuis l'émergence de l'apprentissage profond, car les réseaux de neurones permettent d'approximer efficacement les fonctions de politique et de critique [6, 20]. L'un des apports les plus marquants de cette méthode est la capacité d'un agent obtenu par RL à affronter une équipe de joueurs professionnels du jeu Dota [2].

L'utilisation à grande échelle d'agents ayant appris par renforcement rencontre toujours plusieurs défis, malgré les progrès significatifs réalisés dans le domaine [20, 11, 26, 17, 2, 29]. Cette réticence à adopter ces technologies repose principalement sur le manque de confiance que les humains accordent aux systèmes qu'ils ne peuvent

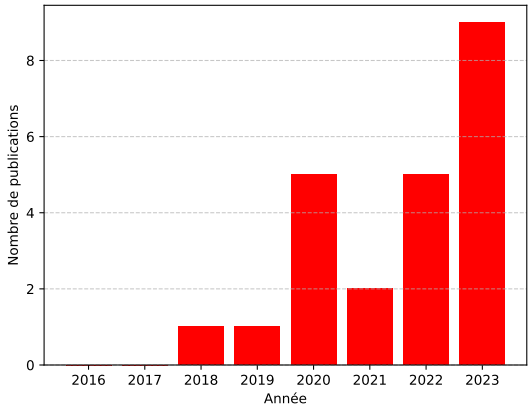

FIGURE 2 – Évolution du nombre de publications sur arXiv ayant le mot-clé “XRL” dans leur résumé.

pas expliquer [10]. Pour résoudre ce problème, un nouveau champ de recherche a émergé : l'intelligence artificiel explicable (*eXplainable Artificial Intelligence*, XAI [10]). Dans le cadre de l'apprentissage par renforcement, on parle plus spécifiquement d'apprentissage par renforcement explicable (*eXplainable Reinforcement Learning*, XRL).

En raison de la pertinence du sujet et pour espérer une plus large utilisation du RL dans le monde réel un nombre croissant d'équipes de recherche travaillent à mettre au point de nouvelles méthodes d'explicabilité. Cet engouement pour l'XAI se traduit ces dernières années par une rapide augmentation du nombre de publications, comme illustré dans les figures 1 et 2. Face à cette croissance rapide, il est apparu crucial pour ce champ de se structurer et de proposer une taxinomie unifiée. Cela permettrait d'assurer la comparaison des futurs résultats entre eux, la compréhension des chercheurs via l'utilisation d'un vocabulaire unifié dans la communauté et une organisation structurée.

Pour permettre aux agents d'évoluer dans le monde réel, il est nécessaire qu'ils puissent interagir à travers des environnements partiellement observables et impliquant un grand nombre d'états. C'est pourquoi nous nous positionnons, dans le cadre de l'apprentissage par renforcement profond (*Deep Reinforcement Learning*, DRL) ainsi que dans un contexte d'environnements décisionnels markoviens partiellement observables (*Partially Observable Markov Decision Processes*, POMDP).

Dans cet article notre objectif est de nous appuyer sur deux articles de revue de l'état de l'art [19, 25] afin d'en synthétiser les informations et de produire une taxinomie unifiée. Celle-ci aura pour but de classifier de manière précise les méthodes existantes ainsi que celles à venir.

2 Présentation du système de classification

Les **explications** [1] décrivent à un **observateur** le fonctionnement d'un système d'intelligence artificielle. Cela lui permet de construire un **modèle mental** du système. Un modèle mental représente la compréhension de l'observateur du système ainsi que sa capacité à anticiper les comportements du système dans de nouvelles situations [9]. Pour obtenir des explications, en RL on utilise des **sources d'explication** sur lesquelles des **techniques d'explicabilité** sont appliquées, comme l'illustre la figure 3.

2.1 Sources d'explication

Nous dénombrons trois **sources d'explication** en RL : les politiques, les trajectoires et les fonctions de valeurs. Ces sources sont obtenues pendant ou après l'entraînement. Dans tous les cas, elles capturent des informations sur l'agent.

Les **politiques** sont des fonctions qui, pour une observation donnée, génèrent une distribution dans l'espace des actions. L'objectif de l'apprentissage par renforcement est de trouver une politique qui, pour chaque état, maximise la somme des récompenses à venir. Les politiques sont améliorées de manière itérative pendant la phase d'apprentissage [31].

Les **trajectoires** représentent une séquence d'interactions successives entre l'environnement et l'agent. En fonction de l'observation, l'agent sélectionne une action qui influence l'environnement, produisant ainsi une récompense et une nouvelle observation. Ce processus se répète jusqu'à la fin de la trajectoire [31]. Nous distinguons trois types de trajectoires : les trajectoires d'entraînement, les trajectoires post-entraînement et les trajectoires fictives. Les trajectoires d'*entraînement* sont celles obtenues par les interactions entre une politique en phase d'apprentissage et l'environnement. Les trajectoires *post-entraînement* sont générées par les politiques finales. Les trajectoires *fictives* résultent de l'interaction entre une politique et une approximation de l'environnement.

Les **fonctions de valeurs** génèrent une prédiction sur la suite de la trajectoire pour une observation donnée [32]. La fonction de valeur la plus couramment utilisée est la fonction de critique, qui fournit une prédiction sur la somme des récompenses obtenues dans le futur de la trajectoire. Cette fonction de critique joue un rôle important dans le RL. L'agent va apprendre à orienter ses actions afin de maximiser cette prédiction. Cependant, d'autres fonctions de valeurs peuvent être considérées afin de produire de l'explicabilité (voir la sous-section 3.3).

2.2 Explication

Une **explication** est une information interprétable, c'est-à-dire une information qui peut être observée, comprise et étudiée par un observateur [1]. C'est à partir de ces informations que l'observateur construit son modèle mental du fonctionnement du système. Nous proposons de caractériser une explication en RL par sa **forme** et son **sujet**.

Nous définissons la **forme** d'une explication comme la caractérisation de ce qu'elle décrit. Selon nos connaissances actuelles [19, 25], nous définissons que les explications peuvent prendre les formes suivantes :

Modèle de la politique : modèle interprétable de la fonction de politique d'un agent.

Modèle de l'interaction : modèle interprétable des interactions entre l'agent et l'environnement.

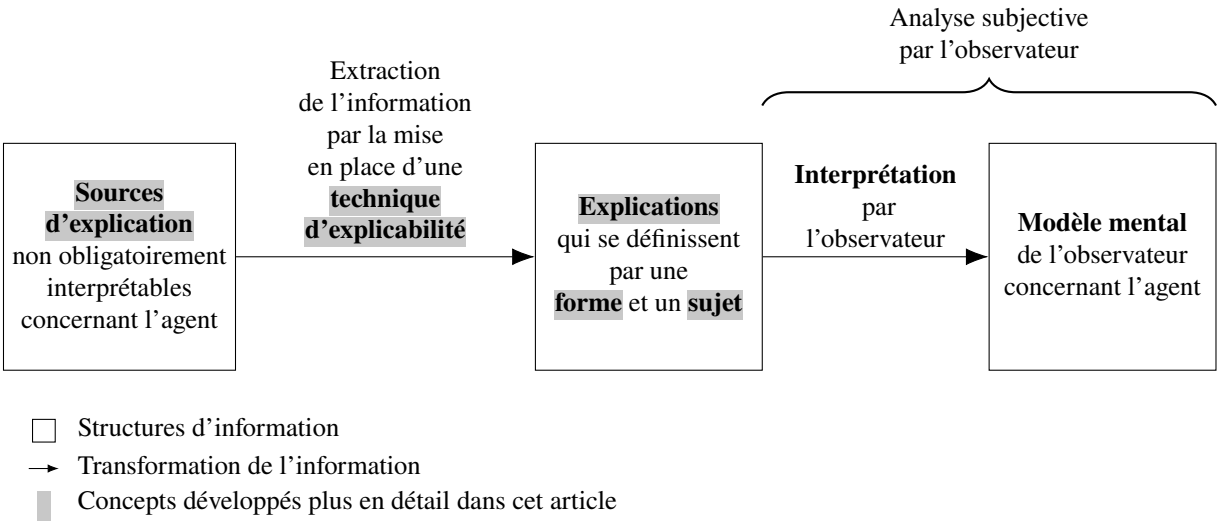

FIGURE 3 – Processus d'explication en apprentissage par renforcement explicable.

Trajectoire de l'agent : représentation d'un ensemble d'interactions successives entre l'environnement et l'agent.

Prédiction sur une trajectoire : prédiction produite par une fonction de valeur sur le futur de la trajectoire.

Contrefactuelle : représentation de l'observation obtenue par la perturbation de l'observation originale dans le but de produire une action différente de la part de l'agent.

Carte de chaleur : pondération de chaque caractéristique des données d'entrée qui représente l'impact sur la sortie de la fonction de politique.

Le **sujet** d'une explication caractérise quelle facette du RL est expliquée. Le sujet d'une explication est entièrement déterminé par sa forme. Il existe trois catégories de sujets : l'*apprentissage*, la *décision* et la *stratégie* [19].

Une explication portant sur l'*apprentissage* offre des informations sur les évolutions itératives de la politique. Par exemple, elle pourrait détailler quelles expériences significatives ont permis à une politique de converger vers une stratégie spécifique ainsi que les grandes étapes de son développement.

Concernant la compréhension de la *décision* d'un agent, cela revient à identifier les éléments de l'observation qui sont déterminants dans le choix de l'action. Ce type d'explication se rapproche de l'apprentissage supervisé, car il ne tient pas compte de la dimension séquentielle de la prise de décision.

Enfin, une explication axée sur la *stratégie* fournit des informations sur une séquence de décisions. Elle cherche ainsi à résumer les choix faits par l'agent en analysant les interactions entre un ensemble de décisions et l'environnement.

2.3 Technique d'explicabilité

Une technique d'explicabilité est un **algorithme** qui utilise une ou plusieurs sources afin de générer des explications pour un humain. Il existe une grande variété de techniques, et leur présentation sera l'objet de la section 3. Ces techniques sont déterminantes dans le processus d'explication car elles définissent les sources utilisées ainsi que le sujet des explications.

3 Analyse des techniques d'explicabilité

Le but de cette section est de détailler le fonctionnement des techniques d'explicabilité (cf. tableau 1). Elles sont présentées en fonction des formes d'information qu'elles produisent. Chaque sous-section présente une **forme d'explication** ainsi que les **techniques** qui lui sont associées.

3.1 Modèles interprétables de la politique

Il existe une catégorie de modèles qui sont compréhensibles via la simple analyse de leur représentation ; on dit de ces modèles qu'ils sont intéprétables [1]. Ces modèles sont utilisés pour représenter la fonction de politique. Les principaux modèles considérés comme intéprétables sont : les arbres de décision [33], les algorithmes [34], la logique formelle [23] et les modèles linéaires [5]. Il est possible de mettre à profit l'interprétabilité intrinsèque de ces modèles pour produire de l'explicabilité dans le carde du RL.

Cette représentation directement intéprétable permet à l'observateur d'extraire de l'information. Une explication issue d'un modèle intépretable fournit principalement des informations sur le processus de décision, car ce qui est modélisé est la correspondance entre une observation et une action. Si le modèle est suffisamment simple ou bien organisé, il est possible pour l'observateur de se faire une représentation d'une séquence de décisions à travers l'environnement et ainsi, fournir des informations sur la stratégie de l'agent.

Pour obtenir un modèle interprétable de la politique, nous disposons de deux techniques d'explication : l'*apprentissage d'une politique interprétable* et l'*approximation de la politique par un modèle interprétable*.

Sujet de l'explication	Forme de l'explication	Technique d'explicabilités	Source de l'explication			
			Politique	Trajectoires d'entraînements	Trajectoires post-entraînements	Fonctions de valeurs
Apprentissage	Trajectoires d'entraînements	*Extraction par comparaisons multi-critiques*		✓		✓
Décision	Modèles interprétables de la politique	*Apprentissage d'une politique interprétable*	✓			
		Approximation de la politique par un modèle interprétables	✓			
	Cartes de chaleur de l'observation	*Perturbation de l'observation*	✓			
		Analyse du gradient de l'observation	✓			
		Architecture à base d'attention	✓			
	Contrefactuelles de l'observation	*Utilisation de l'espace latent pour la génération de contrefactuelles*	✓			
Stratégie	Prédictions sur la trajectoire de l'agent	*Fonctions de valeurs comme observation*				✓
		Enrichissement du critique				✓
	Modèles de l'interaction agent-environnement	*Apprentissage d'un réseau bayésien*			✓	
		Exploitation de l'espace latent pour la génération d'un MDP	✓			
	Trajectoires post-entraînement	*Décomposition hiérarchique*	✓			
		Extraction par analyse du critique			✓	✓

TABLE 1 – Taxinomie des techniques d'explications en apprentissage par renforcement

3.1.1 *Apprentissage d'une politique interprétable*

Cette technique a pour objectif l'apprentissage d'une politique représentée par un modèle **interprétable** [33, 34]. Il s'agit de l'approche la plus directe pour résoudre les problématiques posées par l'XRL. Ce type d'entraînement nécessite une modification du MDP afin de l'adapter à l'espace de recherche et au type de modélisation de la politique. De plus, cet espace de recherche doit souvent être minimisé afin de rendre l'apprentissage possible. Cette **réduction** nécessite l'intégration de **connaissances expertes** dans le processus d'apprentissage. Pour cette technique, c'est la *fonction de politique interprétable* qui est utilisée pour produire de l'explication à destination de l'observateur.

3.1.2 *Approximation de la politique par un modèle interprétable*

À l'aide d'un modèle interprétable, il est possible d'**approximer** une politique non interprétable [28]. Dans ce cas, on se rapporte à un problème d'**apprentissage supervisé**, où l'on utilise la politique non interprétable pour étiqueter chaque observation différente avec l'action choisie par la politique. En utilisant le jeu de données, on effectue un apprentissage supervisé à l'aide du modèle interprétable. Le *modèle interprétable* ainsi créé sera alors utilisé comme explication pour l'observateur. Si cette approximation est suffisamment fidèle, elle peut être utilisée dans l'environnement à la place de la politique, permettant ainsi d'avoir un maximum de contrôle sur l'agent.

3.2 Cartes de chaleur de l'observation

Les cartes de chaleur permettent d'identifier quelles parties du vecteur d'entrée sont déterminantes dans la sélection de la sortie (comme l'illustre la figure 4). Appliqué au cadre de l'apprentissage par renforcement, cela met en valeur les parties de l'espace d'observation déterminantes dans le choix de l'action. De par sa nature, cette forme d'explication fournit des informations sur la prise de décision de la politique. La compréhension simple de ce type de représentation permet une utilisation par un large public. Il existe trois techniques d'explicabilité pour obtenir cette explication : *perturbation de l'observation*, *analyse du gradient* et *architecture à base d'attention*.

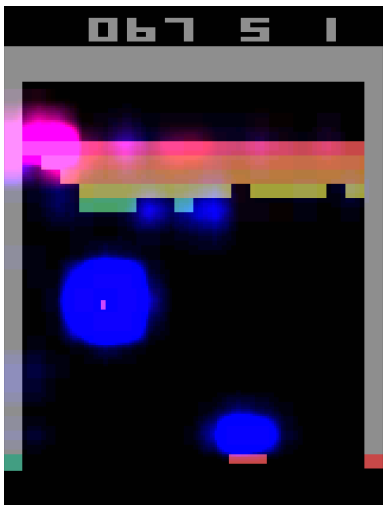

FIGURE 4 – Carte de chaleur par perturbation pour le jeu “Breakout” d’Atari [8]. En surbrillance, les zones les plus déterminantes pour le choix de la fonction de politique.

3.2.1 *Perturbation de l’observation*

L’analyse par perturbation consiste à appliquer différentes **modifications** à l’observation pour étudier leur **impact** sur la sortie de la **fonction de politique** [8]. Plus la variation de la sortie due à une perturbation sur une des composantes du vecteur d’entrée est **importante**, plus cette composante se voit attribuer une **pondération élevée**. La mesure de la perturbation s’effectue en calculant la distance entre la sortie originale et la sortie avec une perturbation en entrée. En utilisant cet ensemble de pondérations, on crée une *carte de chaleur* qui permet à l’observateur de visualiser quelles parties du vecteur d’observation sont sensibles pour le choix de l’action.

3.2.2 *Analyse du gradient de l’observation*

L’analyse du **gradient** de l’observation est un ensemble de techniques qui nécessitent une fonction de politique différenciable pour en extraire de l’information [30, 36, 13]. Pour se faire, on calcule le gradient de la fonction de politique pour l’ensemble des **composants d’un vecteur d’entrée**. Pour chaque composant du vecteur d’entrée, on obtient avec ce gradient une pondération qui représente son importance dans la sortie de la fonction politique. À partir de ces pondérations, on crée une *carte de chaleur* pour permettre à l’observateur d’identifier visuellement les parties du vecteur d’entrée les plus influentes dans la sortie de la fonction de politique.

3.2.3 *Architecture à base d’attention*

Les blocs d’**attention** constituent une architecture pour les réseaux de neurones permettant de pondérer l’importance relative d’une partie d’un vecteur par rapport aux autres parties de ce même vecteur [35]. La pondération se calcule en fonction du vecteur lui-même et est apprise par le réseau de neurones lors de sa phase d’apprentissage. Pour une observation spécifique, chaque partie du vecteur d’observation se voit attribuer un scalaire. On peut donc estimer que les parties avec les pondérations les plus élevées sont les plus importantes pour le choix de la sortie [21, 14]. C’est cette *carte de chaleur* des blocs d’attention qui sera communiquée à l’observateur comme explication.

3.3 Prédictions sur la trajectoire de l’agent

Effectuer des prédictions sur le futur d’une trajectoire permet de se représenter ce à quoi le système s’attend, fondé sur ses expériences d’entraînement (cf. figure 5). Ces prédictions sont réalisées à l’aide de fonctions de valeur. Ces fonctions permettent d’éclairer l’observateur sur le devenir de la trajectoire de l’agent. Par leur nature, cette forme d’explication renseigne sur la stratégie adoptée par l’agent.

La problématique de ce type d’approche réside dans la nécessité de s’assurer que les valeurs prédites par ces fonctions sont réellement déterminantes dans les actions sélectionnées par l’agent. Pour cela, on utilise les techniques suivantes : l’utilisation de *fonctions de prédiction comme observation* et l’*enrichissement du critique*.

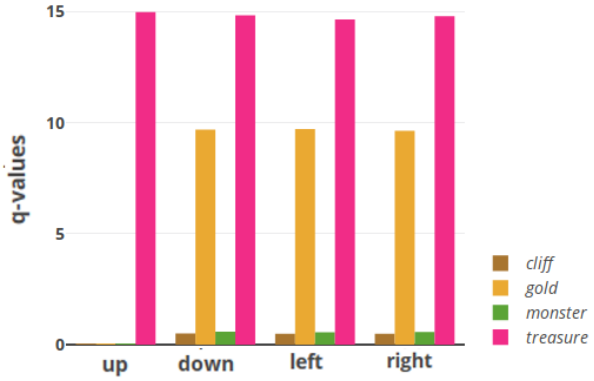

FIGURE 5 – Prédiction des sous-récompenses en fonction de l’action choisie par l’agent pour un états donné [15]. En fonction de l’action choisie par l’agent on peut déterminer quelle sous-récompense il cherche à maximiser.

3.3.1 *Fonctions de valeurs comme observation*

L’objectif est d’entraîner des **fonctions de valeurs interprétrables** afin de les utiliser comme observations pour l’agent. Si les caractéristiques prédites par les fonctions de valeurs ont du **sens** pour l’observateur, celui-ci sera capable de déterminer les aspects que l’agent cherche à maximiser par ses actions [16].

3.3.2 *Enrichissement du critique*

Afin d’obtenir des informations interprétables à partir des prédictions du **critique**, il est possible de **diviser** ce critique en **sous-fonctions** [15]. Ainsi, on décompose la fonction de récompense en sous-fonctions de récompenses, représentant ainsi des **sous-objectifs** de la tâche à accomplir. De manière similaire, pour chacune de ces sous-fonctions de récompenses, on crée des sous-fonctions de critiques qui leur sont associées, ainsi qu’une fonction de combinaison des critiques. L’agent sélectionnera ses actions de manière à maximiser cette fonction de combinaison des

sous-critiques lors de son apprentissage. Grâce à cette méthode, pour chaque observation et action de l'agent, on obtiendra une prédiction sur les sous-objectifs qu'il tente de maximiser. C'est cette *prédiction* qui servira d'explication pour l'observateur.

3.4 Trajectoires de l'agent

Les trajectoires représentent une séquence d'interactions entre l'agent et son environnement. En analysant ces trajectoires, l'observateur peut déduire la dynamique sous-jacente de l'évolution de l'agent au sein de son environnement. Selon le type de trajectoire utilisé (voir 2.2), le sujet de l'explication varie.

L'utilisation des trajectoires post-entraînement pour expliquer la stratégie est systématiquement employée afin d'obtenir une explication sur la stratégie de l'agent, même en dehors du cadre du XRL. Cette méthode simple, peut être perfectionnée grâce à des techniques telles que la *décomposition hiérarchique* et l'*extraction par analyse du critique*.

Pour expliquer le processus d'apprentissage d'un agent, on utilise des trajectoires d'entraînement ayant recours à la technique d'*extraction par comparaisons multi-critiques*.

3.4.1 *Décomposition hiérarchique*

Les agents hiérarchiques sont des agents composés d'un ensemble de **sous-politiques** spécialisées pour des tâches précises dans des sous-ensembles d'états de l'environnement. Cette méthode permet de répartir l'optimisation de la fonction de récompense sur plusieurs modèles et simplifie grandement l'apprentissage. Par ailleurs, ce **découpage** de la fonction de politique peut également apporter de l'explicabilité dans la stratégie de l'agent [3]. Si une sous-politique est choisie, cela signifie que l'agent va mettre en place un comportement spécialisé. Cette spécialisation dans le comportement nous donne des informations sur la stratégie mise en place par l'agent pour maximiser la fonction de récompense. La *trajectoire* de l'agent, ainsi que la *sous-politique* sélectionnée pour chaque action sont utilisées comme informations interprétables.

3.4.2 *Extraction par analyse du critique*

Dans cette technique, on utilise les prédictions du critique pour déterminer quelle **trajectoire présenter à l'observateur** [27]. Ainsi, on sélectionne un ensemble d'états provenant de multiples trajectoires. Pour un même état, on examine l'ensemble des actions possibles. L'état se voit attribuer comme valeur la différence entre la prédiction du critique avec l'action la plus basse et la prédiction du critique avec l'action la plus haute. Ensuite, on choisit de présenter à l'observateur les trajectoires qui contiennent l'état avec la valeur la plus élevée selon cette procédure. On considère que ces trajectoires représentent les moments **critiques** de l'agent et sont donc pertinents à soumettre à l'analyse de l'observateur. Pour cette technique les *trajectoires* sélectionnées représentent l'information intepétable pour l'observateur.

3.4.3 *Extraction par comparaisons multi-critiques*

Dans cette approche, l'objectif principal est de **pondérer les interactions** d'entraînement entre l'agent et l'environnement, selon leurs influences sur le processus d'apprentissage. Pour atteindre cet objectif, plusieurs fonctions de critique sont crées à l'aide de l'apprentissage *off-policy*. Un apprentissage *off-policy* est un processus d'apprentissage par renforcement qui permet à une politique d'apprendre à partir de données générées par une politique différente [7]. Chaque fonction de critique est entraînée sur un sous-ensemble du jeu de données.

La fonction de critique de référence est élaborée en exploitant l'intégralité du jeu de données. Ensuite, pour chaque interaction agent-environnement, une fonction de critique est de nouveau entraînée en excluant spécifiquement cette interaction. La **disparité de prédiction** entre une fonction de critique et la fonction de critique de référence représente l'impact de la suppression de cette interaction sur l'apprentissage. Par ce processus, une valeur est attribuée à chaque interaction en fonction de cette différence : plus la disparité est grande, plus la valeur attribuée à l'interaction est élevée.

Une valeur plus élevée indique que l'interaction est considérée comme plus importante dans le processus d'apprentissage. Ces *valeurs* sur les *interactions* agissent comme des explications pour l'observateur, offrant un moyen savoir l'importance de l'interaction agent-environnement dans le cadre du processus d'apprentissage.

3.5 Contrefactuelles de l'observation

Dans le contexte de l'apprentissage par renforcement, les explications contrefactuelles sont utilisées pour générer de nouvelles observations (cf. figure 6). Cette nouvelle observation distincte de la première conduit l'agent à adopter un comportement alternatif. Une explication contrefactuelle vise à expliquer le processus décisionnel de l'agent. La modification de l'observation renseigne l'observateur sur les ajustements nécessaires pour que l'agent adopte un comportement différent. La méthode dont on dispose pour créer des contrefactuelles est l'*utilisation de l'espace latent pour la génération de contrefactuelles*.

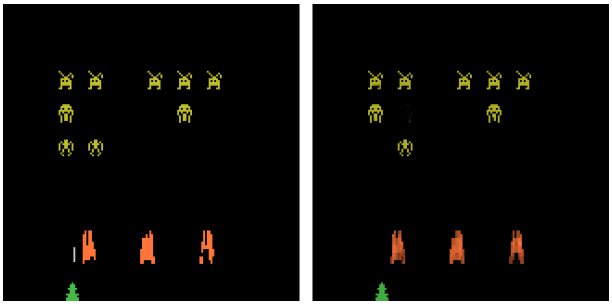

FIGURE 6 – Contrefactuelle sur le comportement de l'agent pour le jeu Space Invader [22]. Sur l'image de gauche l'observation provoque l'action "LEFT", tandis que pour l'observation de l'image de droite, l'action sélectionnée par l'agent est "RIGHT_FIRE".

Dans cette technique, on a recours à des **modèles génératifs** [4] pour exploiter la représentation latente. Il s'agit d'un type de modèle qui apprend à générer de nouvelles données en imitant les données d'entraînement. Dans cette technique, les modèles génératifs cherchent à produire des observations alternatives susceptibles de modifier les actions de l'agent [22].

La difficulté dans ce type de procédé réside dans le fait que pour servir d'explication, les contrefactuelles doivent être à la fois **proches** et **réalisables**. Une contrefactuelle proche signifie une contrefactuelle qui ne diffère pas trop de l'observation d'origine. Une observation contrefactuelle réalisable est une contrefactuelle qui fait sens comme observation dans l'environnement considéré.

3.6 Modèles de l'interaction agent-environnement

La **modélisation** des interactions entre un agent et son environnement permet de comprendre la manière dont les décisions de l'agent sont prises et quels impacts elles ont sur l'environnement (cf. figure 7). En incorporant ces interactions dans un modèle explicatif, on peut mieux appréhender les relations complexes entre l'agent et son environnement, identifiant ainsi les facteurs déterminants dans le processus décisionnel. De par la gestion de l'interaction **agent-environnement**, ce type d'explication a pour objectif d'expliquer la stratégie de l'agent. Les méthodes utilisées afin d'avoir une explication sous forme d'interactions agent-environnement sont : *apprentissage d'un réseau bayésien* et *exploitation de l'espace latent pour la génération d'un MDP*.

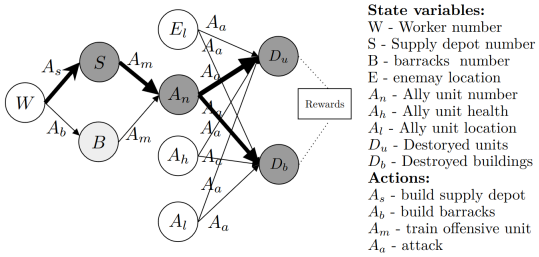

FIGURE 7 – Modèle de l'interaction agent-environnement pour le jeu vidéo Starcraft 2 [18]. Les sommets représentent des variables aléatoires. Les liens représentent les causalités entre les variables aléatoires.

3.6.1 *Apprentissage d'un réseau bayésien*

Il est possible d'observer l'évolution d'un agent dans un environnement sous la forme d'un ensemble de **variables aléatoires** liées entre elles par un **réseau bayésien** [18]. On définit des variables aléatoires comme représentant les aspects externes (l'environnement) et les aspects internes à l'agent (les actions). En étudiant les trajectoires, il est alors possible de déduire des liens de causalité entre toutes ces variables aléatoires. À la fin du processus, on obtient un graphe de liens de causalité pondérés qui servira d'explication pour l'observateur (cf. figure 7). C'est ce *réseau bayésien* représentant les interactions entre l'agent et l'environnement qui sera utilisé comme explication.

3.6.2 *Utilisation de l'espace latent pour la génération d'un MDP*

Chaque couche d'un réseau de neurones fonctionne comme une projection d'un espace vers un autre [37]. En parcourant les couches successives d'un réseau de neurones, l'information prend une forme de plus en plus abstraite. On peut supposer que deux vecteurs proches dans cet espace de projection sont également proches en terme de perception pour l'agent. De ce constat, il est possible de **redéfinir un MDP** en passant par la **représentation abstraite** des dernières couches d'un réseau de neurones. Une fois le nouveaux MDP finalisé, on obtient un modèle d'interaction entre l'agent et son environnement. Ce modèle représenté sous forme d'un *graphe*, sera utilisé comme explication.

4 Discussion

Pour l'XRL, les premières tentatives de classification datent du début des années 2020 [24, 12]. Pendant cette période, les techniques de classification reposaient en grande partie sur les approches XAI généralistes, à savoir la division entre deux catégories d'explication : "locale" et "globale". Les explications locales fournissent des informations spécifiques à une entrée particulière du système et répondent à la question suivante : quel est l'élément déterminant, dans cette entrée, pour le choix du système ? Les explications globales ont quant à elles pour objectif de fournir des informations sur la manière dont le système effectue ses choix, mais cette fois-ci sur l'ensemble des entrées possibles.

Cette première catégorisation ne saisit pas les spécificités de l'apprentissage par renforcement, créant ainsi des imprécisions dans le système de classification. Pour remédier à cette lacune, une nouvelle génération d'articles, parus en 2023 [19, 25], a intégré les particularités de l'apprentissage par renforcement dans son système de classification.

L'article de 2023 proposé par Milani et al. [19], suggère que la classification doit être fondée sur ce que les méthodes cherchent à expliquer. Trois catégories de sujets d'explication sont décrites : "importance des caractéristiques" (*Feature Importance*, FI), "explication au niveau politique" (*Policy Level*, PL), et "explication du processus d'apprentissage et du processus décisionnel de markov" (*Learning Process and markov decision process*). Ce type de classification est en grande partie une adaptation de l'approche XAI classique pour le domaine du XRL : les catégories "locale" et "globale" sont remplacées dans cette classification par FI et PL. Comme différence majeure, on peut noter l'ajout d'une nouvelle catégorie "processus d'apprentissage et du MDP" qui regroupe un ensemble disparate d'éléments, à savoir : les explications de l'apprentissage ainsi que toutes les méthodes qui tirent leur explication du critique. Cette dernière catégorie s'inscrit difficilement dans la logique initialement proposée par les auteurs puisqu'elle contient en elle-même des explications se référant au MDP, alors que jusqu'à présent, la classification se faisait uniquement par rapport au sujet des explications.

L'article de Qing et al. [25] quant à lui utilise des méthodes de classification fondées sur les sources. Pour cette équipe de recherche, les grandes catégories à considérer

sont : les explications du "modèle de l'agent" (*Agent Model-Explaining*), les explications des "récompenses" (*Reward-Explaining*), les explications des "états" (*State-Explaining*) et les explications des "tâches" (*Task-Explaining*). Cependant, encore une fois, des approximations fragilisent la structure de classification. Par exemple, la catégorie "explication des états" n'est pas réellement une source d'explication : on ne peut rien expliquer de la représentation d'un état en lui-même. C'est l'interprétation faite par la politique ou par le critique de l'état qui produit de l'explicabilité.

Ces deux modèles présentent quelques imprécisions dans leurs classifications lorsqu'ils sont examinés individuellement. Cependant, cette situation s'explique par le fait qu'ils ont adopté chacun une seule approche pour leurs classifications. En intégrant les différentes approches proposées par ces deux articles, nous avons élaboré une classification plus complète qui prend en compte et intègre les points de vue respectifs de chacun.

Une observation notable émanant de l'analyse de l'état de l'art, à travers le prisme de notre système de classification, réside dans le fait que très peu de techniques se donnent pour mission d'expliciter le processus d'apprentissage lui-même. Cette lacune peut être attribuée à l'idée que, dans une perspective d'application concrète dans le monde réel, de telles explications peuvent sembler peu pertinentes, étant donné qu'elles ne fournissent pas d'informations détaillées sur le comportement de la politique adoptée. Cependant, nous jugeons pertinent d'explorer cet aspect. En effet, identifier les stratégies viables dans un environnement donné et comprendre les raisons pour lesquelles une stratégie est privilégiée par rapport aux autres peut offrir des informations significatives. Bien que ces informations ne soient pas directement utilisables pour l'intégration opérationnelle des agents, elles peuvent servir à comparer les différents algorithmes d'apprentissage par renforcement entre eux.

5 Conclusion

Notre article a mis en lumière les enjeux cruciaux à propos de l'apprentissage par renforcement explicable, dont le but est de faciliter l'intégration d'agents formés par renforcement dans des situations du monde réel. Il souligne également la nécessité de structurer ce domaine de manière efficiente, compte tenu de sa dynamique croissante. À cette fin, nous avons élaboré une nouvelle taxinomie en fusionnant deux perspectives existantes [19, 25].

Cette classification repose sur des aspects spécifiques de l'apprentissage par renforcement, mettant en exergue les concepts de **source d'explication**, d'**explication** et de **sujet d'explicabilité**. L'état de l'art présenté dans cet article démontre la pertinence de notre nouveau système de classification. Cet outil méthodologique offre la possibilité d'organiser de manière plus efficace les futures recherches dans ce domaine.

Cependant, un élément central demeure à discuter en détail : le type d'audience auquel une explication est destinée. Comme illustré dans la figure 3, une explication est élaborée pour être interprétée par un observateur. Généralement, on peut classer un observateur dans l'une des trois grandes catégories d'audience : les experts en intelligence artificielle, les spécialistes du domaine d'application du modèle d'IA, et les non-initiés, c'est-à-dire des observateurs qui ne possèdent aucune connaissance ni en intelligence artificielle ni dans le domaine d'application [19].

Pour évaluer l'impact d'un processus explicatif sur l'un de ses publics cibles, il est impératif que la communauté de chercheurs établisse des protocoles d'évaluation visant à mesurer la pertinence d'une explication. Ce domaine a été exploré par Milani et al. [19], dont le travail permet l'introduction d'une classification des différentes méthodes d'évaluation ainsi que diverses métriques permettant de quantifier l'efficacité d'une explication. Des recherches approfondies dans cette direction sont nécessaires, notamment pour comparer différentes approches d'explicabilité et évaluer leur pertinence en fonction du public visé.

Références

[1] Matthieu Bellucci, Nicolas Delestre, Nicolas Malandain, and Cecilia Zanni-Merk. Towards a terminology for a fully contextualized xai. *Procedia Computer Science*, 192 :241–250, 2021. **Article**.

[2] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv :1912.06680*, 2019. **Article**. **Vidéo**.

[3] Benjamin Beyret, Ali Shafti, and A. Aldo Faisal. Dot-to-dot : Explainable hierarchical reinforcement learning for robotic manipulation. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, November 2019. **Article**.

[4] Sam Bond-Taylor, Adam Leach, Yang Long, and Chris G Willcocks. Deep generative modelling : A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. *IEEE transactions on pattern analysis and machine intelligence*, 2021. **Article**.

[5] Damien Garreau and Ulrike von Luxburg. Explaining the explainer : A first theoretical analysis of lime. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1287–1296. PMLR, 26–28 Aug 2020. **Article**.

[6] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. **Livre**.

[7] Omer Gottesman, Joseph Futoma, Yao Liu, Sonali Parbhoo, Leo Anthony Celi, Emma Brunskill, and Finale Doshi-Velez. Interpretable off-policy evaluation in reinforcement learning by highlighting influential transitions, 2020. **Article**.

[8] Sam Greydanus, Anurag Koul, Jonathan Dodge, and Alan Fern. Visualizing and understanding atari agents, 2018. **Article**. **Code**.

[9] David Gunning and David Aha. Darpa's explainable artificial intelligence (xai) program. *AI magazine*, 40(2) :44–58, 2019. **Article**.

[10] David Gunning, Eric Vorm, Yunyan Wang, and Matt Turek. Darpa's explainable ai (xai) program : A retrospective. *Authorea Preprints*, 2021. **Article**.

[11] Matteo Hessel, Joseph Modayil, Hado van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. Rainbow : Combining improvements in deep reinforcement learning, 2017. **Article**.

[12] Alexandre Heuillet, Fabien Couthouis, and Natalia Díaz-Rodríguez. Explainability in deep reinforcement learning, 2020. **Article**.

[13] Tobias Huber, Dominik Schiller, and Elisabeth André. Enhancing explainability of deep reinforcement learning through selective layer-wise relevance propagation. In *KI 2019 : Advances in Artificial Intelligence : 42nd German Conference on AI, Kassel, Germany, September 23–26, 2019, Proceedings 42*, pages 188–202. Springer, 2019. **Article**.

[14] Hidenori Itaya, Tsubasa Hirakawa, Takayoshi Yamashita, Hironobu Fujiyoshi, and Komei Sugiura. Visual explanation using attention mechanism in actor-critic-based deep reinforcement learning, 2021. **Article**.

[15] Zoe Juozapaitis, Anurag Koul, Alan Fern, Martin Erwig, and Finale Doshi-Velez. Explainable reinforcement learning via reward decomposition. In *IJCAI/ECAI Workshop on explainable artificial intelligence*, 2019. **Article**.

[16] Zhengxian Lin, Kim-Ho Lam, and Alan Fern. Contrastive explanations for reinforcement learning via embedded self predictions, 2021. **Article**. **Code**.

[17] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments, 2020. **Article**.

[18] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. Explainable reinforcement learning through a causal lens. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 2493–2500, 2020. **Article**.

[19] Stephanie Milani, Nicholay Topin, Manuela Veloso, and Fei Fang. Explainable reinforcement learning : A survey and comparative review. *ACM Computing Surveys*, 2023. **Article**.

[20] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning, 2013. **Article**.

[21] Alexander Mott, Daniel Zoran, Mike Chrzanowski, Daan Wierstra, and Danilo Jimenez Rezende. Towards interpretable reinforcement learning using attention augmented agents. *Advances in neural information processing systems*, 32, 2019. **Article**.

[22] Matthew L Olson, Roli Khanna, Lawrence Neal, Fuxin Li, and Weng-Keen Wong. Counterfactual state explanations for reinforcement learning agents via generative deep learning. *Artificial Intelligence*, 295 :103455, 2021. **Article**. **Code**.

[23] Ali Payani and Faramarz Fekri. Inductive logic programming via differentiable deep neural logic networks, 2019. **Article**. **Code**.

[24] Erika Puiutta and Eric MSP Veith. Explainable reinforcement learning : A survey, 2020. **Article**.

[25] Yunpeng Qing, Shunyu Liu, Jie Song, Huiqiong Wang, and Mingli Song. A survey on explainable reinforcement learning : Concepts, algorithms, challenges. *arXiv preprint arXiv :2211.06665*, 2022. **Article**. **GitHub**.

[26] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. **Article**.

[27] Pedro Sequeira and Melinda Gervasio. Interestingness elements for explainable reinforcement learning : Understanding agents' capabilities and limitations. *Artificial Intelligence*, 288 :103367, November 2020. **Article**.

[28] Alexander Sieusahai and Matthew Guzdial. Explaining deep reinforcement learning agents in the atari domain through a surrogate model, 2021. **Article**.

[29] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. Mastering chess and shogi by self-play with a general reinforcement learning algorithm, 2017. **Article**.

[30] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks : Visualising image classification models and saliency maps, 2014. **Article**. **Code**.

[31] Richard S Sutton and Andrew G Barto. *Reinforcement learning : An introduction*. MIT press, 2018. **Livre**.

[32] Richard S Sutton, Joseph Modayil, Michael Delp, Thomas Degris, Patrick M Pilarski, Adam White, and Doina Precup. Horde : A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 761–768, 2011. **Article**.

[33] Nicholay Topin, Stephanie Milani, Fei Fang, and Manuela Veloso. Iterative bounding mdps : Learning interpretable policies via non-interpretable methods. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9923–9931, 2021. **Article**.

[34] Dweep Trivedi, Jesse Zhang, Shao-Hua Sun, and Joseph J. Lim. Learning to synthesize programs as interpretable and generalizable policies, 2022. **Article**. **Code**.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. **Article**.

[36] Laurens Weitkamp, Elise van der Pol, and Zeynep Akata. Visual rationalizations in deep reinforcement learning for atari games, 2019. **Article**.

[37] Tom Zahavy, Nir Ben Zrihem, and Shie Mannor. Graying the black box : Understanding dqns, 2017. **Article**.