Étude de la motivation intrinsèque en apprentissage par renforcement

Arthur Aubret, Laëtitia Matignon, Salima Hassas

Étude de la motivation intrinsèque en apprentissage par renforcement

A. Aubret[1] L. Matignon[1] S. Hassas[1]

[1] Univ Lyon, Université Lyon 1, CNRS, LIRIS, F-69622, Villeurbanne, France

arthur.aubret@liris.cnrs.fr

Résumé

Malgré les nombreux travaux existants en apprentissage par renforcement (AR) et les récents succès obtenus notamment en le combinant avec l'apprentissage profond, l'AR fait encore aujourd'hui face à de nombreux défis. Certains d'entre eux, comme la problématique de l'abstraction temporelle des actions ou la difficulté de concevoir une fonction de récompense sans connaissances expertes, peuvent être adressées par l'utilisation de récompenses intrinsèques. Dans cet article, nous proposons une étude du rôle de la motivation intrinsèque en AR et de ses différents usages, en détaillant les intérêts et les limites des approches existantes. Notre analyse suggère que la notion d'information mutuelle est centrale à la plupart des travaux utilisant la motivation intrinsèque en AR. Celle-ci, combinée aux algorithmes d'AR profond, permet d'apprendre des comportements plus complexes et plus généralisables que ce que permet l'AR traditionnel.

Mots Clef

Apprentissage par renforcement, motivation intrinsèque, curiosité, acquisition de connaissances, *empowerment*, *options*, génération d'objectifs, méta-récompense.

Abstract

Despite many existing works in reinforcement learning (RL) and the recent successes obtained by combining it with deep learning, RL is facing many challenges. Some of them, like the ability to abstract the action or the difficulty to conceive a reward function without expert knowledge, can be addressed by the use of intrinsic motivation. In this article, we provide a survey on the role of intrinsic motivation in RL and its different usages by detailing interests and limits of existing approaches. Our analysis suggests that mutual information is central to most of the work using intrinsic motivation in RL. The combination of deep RL and intrinsic motivation enables to learn more complicated and more generalisable behaviours than what enables standard RL.

Keywords

Reinforcement learning, intrinsic motivation, curiosity, knowledge acquisition, empowerment, options, generation of objectives, meta-reward.

1 Introduction

En apprentissage par renforcement (AR), un agent apprend par essais-erreurs à maximiser l'espérance des récompenses reçues suite aux actions effectuées dans son environnement [Sutton and Barto, 1998].

Traditionnellement, pour apprendre une tâche, un agent maximise une récompense définie selon la tâche à accomplir : cela peut être un score lorsque l'agent doit apprendre à gagner à un jeu ou une fonction de distance lorsque l'agent apprend à atteindre un objectif. Nous parlons alors de récompense extrinsèque (ou *feedback*) car la fonction de récompense est fournie de manière experte spécifiquement pour la tâche. Avec une récompense extrinsèque, plusieurs résultats spectaculaires ont été obtenus sur les jeux Atari [Bellemare *et al.*, 2015] avec le Deep Q-network (DQN) [Mnih *et al.*, 2015] ou sur le jeu du go avec AlphaGo Zero [Silver *et al.*, 2017]. Cependant, ces approches se révèlent le plus souvent infructueuses lorsque les récompenses sont trop éparses dans l'environnement, et l'agent est alors incapable d'apprendre le comportement recherché pour la tâche [François-Lavet *et al.*, 2018]. D'autre part, les comportements appris par l'agent sont difficilement réutilisables, aussi bien au sein d'une même tâche que pour plusieurs tâches [François-Lavet *et al.*, 2018] : il est difficile pour un agent de généraliser ses compétences de manière à prendre des décisions abstraites dans l'environnement. Par exemple, une décision abstraite (ou de haut-niveau) pourrait être *aller jusqu'à la porte* en utilisant des actions primitives (ou de bas-niveau) consistant à se déplacer dans les quatre directions cardinales ; ou encore de *se déplacer en avant* en contrôlant les différentes articulations d'un robot humanoïde comme dans le simulateur de robotique MuJoCo [Todorov *et al.*, 2012]. Ces actions abstraites sont souvent appelées *options* [Sutton *et al.*, 1999].

Contrairement à l'AR, l'apprentissage développemental [Piaget and Cook, 1952; Cangelosi and Schlesinger, 2018; Oudeyer and Smith, 2016] s'inspire de la tendance qu'ont les bébés, ou plus généralement tout organisme, à explorer spontanément leur environnement [Gopnik *et al.*, 1999; Georgeon *et al.*, 2011] : c'est ce que nous appelons une motivation intrinsèque, laquelle peut être issue d'une récompense intrinsèque. Ce type de motivation permet d'acquérir de manière autonome de nouvelles connaissances ou

compétences, lesquelles facilitent alors l'apprentissage de nouvelles tâches [Baldassarre and Mirolli, 2013]. Ce paradigme offre une plus grande flexibilité d'apprentissage, de part l'utilisation d'une fonction de récompense plus générale, permettant d'adresser les problèmes soulevés précédemment dans le cas d'une récompense extrinsèque. Typiquement, nous verrons que la motivation intrinsèque peut permette d'inciter l'agent à explorer son environnement ou d'apprendre des *options* indépendantes de sa tâche principale.

Dans cet article nous proposons une étude de l'usage de la motivation intrinsèque dans le framework de l'apprentissage profond par renforcement, plus particulièrement nous souhaitons répondre aux questions suivantes :

- Comment caractériser la motivation intrinsèque ?
- Comment la motivation intrinsèque peut-elle s'intégrer au framework d'AR ?
- Quel rôle joue-t-elle vis-à-vis des défis énoncés ci-dessus ?
- Quel lien existe-t-il entre la motivation intrinsèque et la théorie de l'information ?
- Quelles sont les limites actuelles de l'utilisation de récompenses intrinsèques en AR ?

Nous ne prétendons pas faire une étude exhaustive mais plutôt donner les axes de recherches courants et des perspectives à exploiter.

Dans un premier temps, nous définirons les éléments clés de l'article qui sont les processus de décision markovien, les bases de la théorie de l'information et la motivation intrinsèque (Partie 2). Ensuite nous mettrons en avant les problématiques de l'AR et expliquerons comment combiner l'AR et la motivation intrinsèque (Partie 3). Nous aurons alors les éléments pour détailler les trois différents types de travaux intégrant l'AR et la motivation intrinsèque (Partie 4). Puis, nous prendrons du recul et analyserons les points communs entre les différents travaux (Partie 5). Pour terminer, nous mettrons en avant les défis actuels des modèles intégrant la motivation intrinsèque à l'AR (Partie 6).

2 Définitions

2.1 Processus de décision markovien

L'objectif d'un processus de décision markovien (MDP) est de maximiser l'espérance de récompense reçue via une suite d'interaction. Il est défini par :

- S l'ensemble des états possibles du système.
- A l'ensemble des actions possibles.
- P est la fonction de transition $P : S \times A \times S \to \mathbb{R}$.
- R est la fonction de récompense $R : S \times A \to \mathbb{R}$.
- $\gamma \in [0, 1]$ est le facteur d'atténuation.
- $\rho_0 : S \to \mathbb{R}$ est la distribution initiale de l'état du système.

L'agent démarre dans un état s_0 donné par ρ_0 puis effectue une action a_0. Il attend ensuite la réponse de l'environnement qui lui renverra un nouvel état s', donné par la fonction de transition P, et une récompense r évaluée par la fonction de récompense R. L'agent peut répéter la boucle d'interactions jusqu'à la fin d'un épisode.

L'objectif du MDP est de maximiser la récompense sur le long terme :

$$\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]. \tag{1}$$

Un algorithme de renforcement permet d'associer des actions a aux états s via une politique π. L'objectif de l'agent est alors de trouver la politique π^* maximisant la récompense :

$$\pi^* = \arg\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, \pi(s_t))\right]. \tag{2}$$

De manière à trouver l'action maximisant la récompense sur le long-terme dans un état s, il est courant de maximiser l'espérance de gain atténuée depuis cet état, noté $V(s)$, ou l'espérance de gain atténuée depuis le couple (état,action) $Q(s, a)$ (c.f équation 3). Cela permet d'adresser le *credit assignment problem* en mesurant le rôle du couple (état,action) dans l'obtention de la récompense cumulée [Sutton and Barto, 1998].

$$Q_\pi(s, a) = E_{a_t \sim \pi(s_t)}\left(\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)|_{s_0=s, a_0=a}\right). \tag{3}$$

Pour calculer ces valeurs, il est possible d'utiliser l'équation de Bellman [Sutton and Barto, 1998] :

$$Q(s_t, a_t) = R(s_t, a_t) + \gamma Q(P(s_t, a_t), a_{t+1}) \tag{4}$$

Q et/ou π sont souvent approximés via des réseaux de neurones lorsque l'espace d'état est continu ou de très grande taille [Mnih *et al.*, 2016; Lillicrap *et al.*, 2015] .

2.2 Théorie de l'information

L'entropie de Shannon quantifie l'information nécessaire moyenne pour déterminer la valeur d'une variable aléatoire. Soit X une variable aléatoire de probabilité $p(X)$ satisfaisant la condition de normalisation et de positivité, on définit son entropie par :

$$H(X) = -\int_X p(x) \log p(x). \tag{5}$$

L'entropie est maximale lorsque X suit une distribution uniforme, et minimale lorsque $p(X)$ vaut zero partout sauf en une valeur, ce qui est typiquement le cas dans une distribution de Dirac.

On peut aussi définir l'entropie conditionnelle sur la variable aléatoire S. Elle quantifie l'information nécessaire moyenne pour trouver X sachant la valeur d'une autre variable aléatoire S :

$$H(X|S) = -\int_S p(s) \int_X p(x|s) \log p(x|s). \tag{6}$$

L'information mutuelle conditionnelle permet de quantifier l'information que contient une variable aléatoire sur une autre, sachant la valeur d'une troisième variable aléatoire. Elle s'écrit de plusieurs manières :

$$
\begin{aligned}
I(X;Y|S) &= H(X|S) - H(X|Y,S) && (7)\\
&= H(Y|S) - H(Y|A,S) && (8)\\
&= H(X|S) + H(Y|S) - H(X,Y|S)\\
&= \int_{Y,X} p(x,y|s) \log \frac{p(x,y|s)}{p(x|s)p(y|s)} && (9)\\
&= \int_X p(x|s) \int_Y p(y|x,s) \log \frac{p(y|x,s)}{p(y|s)} && (10)\\
&= D_{KL}\Big[p(x,y|s)||p(x|s)p(y|s)\Big] && (11)\\
&= \int_X p(x|s) D_{KL}\Big[p(y|x,s)||p(y|s)\Big] && (12)
\end{aligned}
$$

On voit avec les équations (7) et (8) que l'information mutuelle est symétrique et qu'elle caractérise la baisse d'entropie sur X apportée par Y (ou inversement). L'équation (11) permet de voir l'information mutuelle conditionnelle comme l'écart entre la distribution $P(Y,X|S)$ et cette même distribution si Y et X étaient des variables indépendantes (cas où $H(Y|X,S) = H(Y|S)$). Pour plus de détails sur ces notions, nous renvoyons le lecteur vers [Tishby *et al.*, 2000; Ito, 2016; Cover and Thomas, 2012].

2.3 Motivation intrinsèque

L'idée de la motivation intrinsèque est d'inciter un agent à avoir un certain type de comportement sans que l'environnement intervienne directement. Plus simplement, il s'agit de faire quelque chose pour son inhérente satisfaction plutôt que pour une récompense assignée par l'environnement [Ryan and Deci, 2000]. Ce type de motivation renvoie à l'apprentissage développemental, lequel s'inspire de la tendance des bébés à explorer leur environnement [Gopnik *et al.*, 1999]. Historiquement, la motivation intrinsèque est issue de la tendance des organismes à jouer ou explorer leur environnement sans qu'aucune récompense externe ne leur soit attribuée [White, 1959; Ryan and Deci, 2000].

Plus rigoureusement, Oudeyer [Oudeyer and Kaplan, 2008] explique qu'*une situation est intrinsèquement motivée pour une entité autonome si son intérêt dépend principalement de la collecte ou comparaison d'information depuis différents stimulus indépendamment de leur sémantique.* Le point principal est que l'agent ne doit avoir aucun *a priori* sur la sémantique des observations qu'il reçoit. On remarque que le terme de comparaison d'information renvoie directement à la théorie de l'information ci-dessus. Berlyne [Berlyne, 1965] et Oudeyer [Oudeyer and Kaplan, 2008] proposent plusieurs types de motivation pouvant être caractérisées comme intrinsèques :

TABLE 1 – Types d'apprentissage. Le *feedback* fait ici référence à une supervision experte.

	Avec *feedback*	Sans *feedback*
Actif	Renforcement	Motivation intrinsèque
Passif	Supervisé	Non supervisé

- la nouveauté et la complexité comme étant quelque chose que l'agent ne connaît pas ;
- la surprise et l'incongruité peuvent attirer l'agent car cela remet en question ses précédents apprentissages ;
- l'ambiguïté et l'indistinction renvoient à l'incompréhension de l'agent vis-à-vis des observations.

Typiquement, un étudiant qui fait ses devoirs de mathématique car il les trouve intéressants est intrinsèquement motivé tandis que son camarade qui les fait pour avoir une bonne note est extrinsèquement motivé. De même, jouer avec des jouets pour s'amuser est une motivation intrinsèque tandis que jouer à un jeu télévisé pour gagner de l'argent est une motivation extrinsèque. La notion d'**intrinsèque/extrinsèque** renvoie au *pourquoi de l'action*, à ne pas confondre avec l'internalité/externalité qui renvoie à la localisation de la récompense [Oudeyer and Kaplan, 2008].

La table 1 montre la différence entre l'apprentissage par renforcement et l'usage de motivation intrinsèque. L'apprentissage par renforcement est un apprentissage actif puisque l'agent apprend de ses interactions avec l'environnement, contrairement à des méthodes de classification ou de régression supervisées. L'apprentissage non supervisé est quant à lui un apprentissage passif qui n'utilise pas de labels prédéfinis, donc sans *feedback*. Enfin, le remplacement du *feedback* par une récompense intrinsèque permet de s'affranchir de la supervision experte.

3 Intégrer la motivation intrinsèque à l'AR

Dans cette partie, nous détaillons tout d'abord les deux principaux défis que cherchent à résoudre les travaux combinant récompenses intrinsèques et AR. Ensuite, nous présentons le framework général permettant d'intégrer des récompenses intrinsèques à l'AR.

3.1 Problématiques de l'AR

Les récompenses éparses. Les algorithmes classiques d'AR fonctionnent dans des environnements où les récompenses sont **denses**, *i.e.* que l'agent reçoit une récompense après presque chaque action réalisée. Dans ce type d'environnements, des politiques d'explorations naïves telles que l'exploration ϵ-greedy [Sutton and Barto, 1998] ou l'ajout de bruit gaussien [Lillicrap *et al.*, 2015] sont efficaces. Des méthodes plus élaborées peuvent aussi être utilisées comme l'exploration Boltzmann [Cesa-Bianchi *et al.*, 2017; Mnih *et al.*, 2015], une exploration dans l'espace des

paramètres [Plappert *et al.*, 2017; Rückstiess *et al.*, 2010; Fortunato *et al.*, 2017] ou l'AR bayésien [Ghavamzadeh *et al.*, 2015].

Dans les environnements à récompenses **éparses**, l'agent reçoit un signal de récompense seulement après avoir exécuté une longue séquence spécifique d'actions. Le jeu *Montezuma's revenge* [Bellemare *et al.*, 2015] est un environnement de référence pour illustrer le cas des récompenses éparses. Dans ce jeu, un agent doit se déplacer de salles en salles en y récupérant des objets (clés pour ouvrir les portes, torches, ...). L'agent reçoit une récompense uniquement lorsqu'il trouve des objets ou la sortie d'une salle. Plusieurs actions spécifiques doivent donc être réalisées avant l'obtention d'une récompense. Ce type d'environnements à récompenses éparses est pratiquement impossible à résoudre avec les méthodes d'exploration mentionnées ci-dessus, l'agent parvenant difficilement à apprendre une bonne politique vis-à-vis de la tâche [Mnih *et al.*, 2015].

Plutôt que de travailler sur la politique d'exploration, il est courant de construire une récompense intermédiaire dense qui s'ajoute à celle de la tâche pour faciliter l'apprentissage de l'agent [Su *et al.*, 2015]. Cependant, la construction d'une fonction de récompense fait souvent apparaître des erreurs inattendues [Ng *et al.*, 1999; Amodei *et al.*, 2016] et nécessite le plus souvent des compétences expertes. Par exemple, il est difficile de concevoir une récompense locale pour des tâches de navigation. En effet, il faudrait être capable de calculer le plus court chemin entre l'agent et son objectif, ce qui revient à résoudre le problème de navigation. D'un autre côté l'automatisation de la construction d'une récompense locale (sans faire appel à un expert) demande de trop grandes capacités de calcul [Chiang *et al.*, 2019].

L'abstraction temporelle des actions. L'abstraction temporelle des actions consiste à utiliser des actions de haut niveau, aussi appelées *options*, pouvant avoir des durées d'exécution différentes [Sutton *et al.*, 1999]. A chaque *option* est associée une politique intra-option qui définit les actions (de bas-niveau ou d'autres *options*) à réaliser dans chaque état lorsque l'*option* est exécutée. Abstraire les actions est un élément clé pour accélérer l'apprentissage. En effet, le nombre de choix à réaliser pour atteindre un objectif peut être fortement diminué si des *options* sont utilisées. Cela facilite aussi le renforcement des actions qui sont déterminantes pour l'obtention de la récompense (c'est le *credit assignment problem* [Sutton and Barto, 1998]). Par exemple supposons qu'un robot essaye d'accéder à un gâteau sur une table. Si le robot a une *option* `se rendre à la table` et qu'il la suit, il ne lui restera qu'à prendre le gâteau. Il sera alors facile d'assimiler l'obtention du gâteau à l'option `se rendre à la table`. À l'inverse, si le robot doit apprendre à gérer chacune de ses articulations (actions de bas-niveau), il aura du mal à déterminer quelles actions de bas-niveau lui ont permis d'obtenir le gâteau, parmi toutes celles qu'il a réalisées.

Utiliser des *options* peut par ailleurs faciliter l'exploration lorsque les récompenses sont éparses. Pour illustrer cela, supposons que l'agent ait accès à l'*option* `Aller chercher la clé` dans *Montezuma's revenge*. Le problème devient trivial car une seule action d'exploration peut mener à une récompense là où il faudrait sans *option* une séquence spécifique d'actions de bas-niveau.

Concernant la politique intra-option, elle peut être définie manuellement [Sutton *et al.*, 1999], mais cela nécessite des compétences expertes, ou être apprises avec la fonction de récompenses [Bacon *et al.*, 2017; Riemer *et al.*, 2018], mais les *options* ne seront alors pas réutilisables pour d'autres tâches.

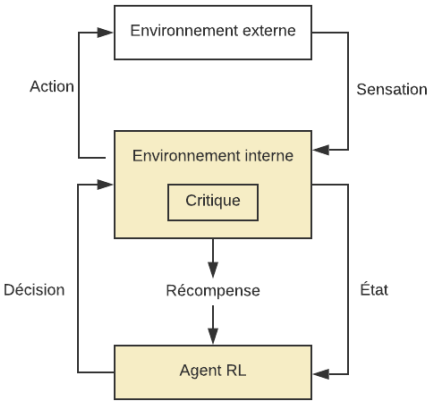

FIGURE 1 – Nouvelle modélisation du MDP. Adaptée de [Singh *et al.*, 2010]

Pour résumer, l'utilisation de la motivation intrinsèque peut aider d'une part, à définir des politiques d'explorations plus élaborées permettant d'améliorer l'AR dans le cas de récompenses éparses ; et d'autre part, à établir une abstraction des décisions ne dépendant pas de l'objectif.

3.2 Nouvelle modélisation de l'AR

L'apprentissage par renforcement est issu du courant behavioriste [Skinner, 1938] et utilise des récompenses extrinsèques [Sutton and Barto, 1998]. Cependant [Singh *et al.*, 2010] et [Barto *et al.*, 2004] ont reformulé le framework de l'AR pour y incorporer la motivation intrinsèque. Plutôt que de considérer l'environnement MDP comme étant l'environnement de l'agent, ils suggèrent que l'environnement MDP peut être constitué d'une partie interne à l'agent et d'une partie externe. La partie externe correspond à l'environnement réel de l'agent et la partie interne est celle qui calcule le signal de récompense total via les observations, actions et récompenses extrinsèques. Dès lors, nous pouvons accepter une récompense intrinsèque comme une récompense de l'environnement MDP. La figure 1 résume le nouveau framework.

L'évolution permet, d'après [Singh *et al.*, 2010], de trouver une fonction de récompense intrinsèque générale qui

maximise une fonction de fitness. Nous pensons que cette motivation intrinsèque peut être une méta-compétence facilitant l'apprentissage d'autres comportements. La curiosité, par exemple, ne génère pas d'avantage sélectif immédiatement mais permet l'acquisition de compétences étant elles-mêmes des avantages sélectifs. Plus largement, l'utilisation de motivation intrinsèque peut permettre d'obtenir des comportements intelligents pouvant servir des objectifs plus efficacement qu'avec du renforcement classique [Lehman and Stanley, 2008] (voir partie 4).

En pratique, la récompense r est souvent calculée comme une somme pondérée de récompense intrinsèque r_{int} et extrinsèque r_{ext} : $r = \alpha r_{int} + \beta r_{ext}$ [Burda *et al.*, 2018; Gregor *et al.*, 2016; Vezhnevets *et al.*, 2017]. Nous pouvons aussi parler de la récompense intrinsèque comme un bonus intrinsèque. Notons que la fonction de récompense évolue au cours du temps, cela ne respecte pas la propriété d'invariance d'un MDP.

4 Principales motivations intrinsèques

Dans cet article, nous proposons de catégoriser les différentes motivations intrinsèques utilisées en AR en trois classes de méta-compétences : l'acquisition de connaissances, l'*empowerment* et la génération d'objectifs. Nous n'adressons pas les motivations à base sociales ou émotionnelles, un état de l'art récent étant déjà disponible [Moerland *et al.*, 2018]. De plus ces motivations font rarement office de méta-compétences mais jouent plutôt le rôle de récompenses auxiliaires [Perolat *et al.*, 2017; Yu *et al.*, 2015; Hughes *et al.*, 2018; Sequeira *et al.*, 2011; 2014; Williams *et al.*, 2015; Moerland *et al.*, 2016] adressant le problème de la fonction de récompense optimale dans l'espace des fonctions de récompenses [Singh *et al.*, 2010].

4.1 Acquisition de connaissances

Les motivations intrinsèques basées sur l'acquisition de connaissances sont les plus utilisées en AR, de part leur capacité à rendre accessible des récompenses éparses. Nous allons étudier les trois principales méthodes existantes pour implémenter une récompense d'acquisition de connaissances. La première utilise le gain d'information, la seconde l'erreur de prédiction des états et la troisième l'évaluation de la nouveauté d'un état. Dans les deux cas, cette récompense intrinsèque permet de compléter la politique d'exploration. Nous focalisons notre étude sur des travaux récents et renvoyons le lecteur vers [Schmidhuber, 2010] pour une revue sur des travaux plus anciens concernant l'acquisition de connaissances.

Le gain d'information. Le gain d'information est une récompense basée sur la réduction de l'incertitude sur la dynamique de l'environnement suite à une action [Oudeyer and Kaplan, 2009; Little and Sommer, 2013], ce qui peut aussi être assimilé au progrès d'apprentissage [Oudeyer and Kaplan, 2009; Schmidhuber, 1991; Frank *et al.*, 2014] ou à la surprise bayésienne [Itti and Baldi, 2006; Schmidhuber, 2008]. Cela permet d'une part de diriger l'agent vers les zones déterministes qu'il ne connaît pas, d'autre part de l'empêcher d'aller dans les zones fortement stochastiques. En effet, si la zone est déterministe, les transitions de l'environnement sont prédictibles et son incertitude à propos des dynamiques de la zone peut baisser. Au contraire, lorsque les transitions sont stochastiques, l'agent se révèle incapable de prédire les transitions et ne réduit pas son incertitude. La stratégie d'exploration VIME [Houthooft *et al.*, 2016] formalise le progrès d'apprentissage de manière bayésienne, l'intérêt de ces approches étant de pouvoir mesurer l'incertitude sur le modèle appris [Blundell *et al.*, 2015]. L'agent apprend donc la dynamique de l'environnement via un réseau de neurone bayésien [Graves, 2011], et utilise la réduction d'incertitude sur la dynamique de l'environnement comme bonus intrinsèque. Autrement dit, l'agent essaye de faire des actions qui sont informatives sur le modèle des dynamiques. Similairement [Achiam and Sastry, 2017] propose de remplacer le modèle bayésien par un réseau de neurones classique suivi d'une distribution de probabilité gaussienne factorisée. Deux approches sont proposées : la première (NLL) utilise comme bonus intrinsèque l'entropie croisée de la prédiction et la seconde (AKL) l'amélioration de prédiction entre un instant t et après k améliorations à $t+k$. Bien que ces deux méthodes soient plus simples que VIME, les bénéfices en terme de performance sont mitigés.

L'erreur de prédiction. L'idée est ici de diriger l'agent vers des zones pour lesquelles la prédiction de l'état suivant est difficile. Plutôt que de considérer la réduction d'erreur dans le modèle des dynamiques, l'agent utilise l'erreur directement comme récompense intrinsèque. Ainsi, au lieu d'utiliser des modèles probabilistes comme précédemment, Dynamic AE [Stadie *et al.*, 2015] calcule la distance entre l'état prédit et l'état réel dans un espace compressé via un auto-encodeur [Hinton and Salakhutdinov, 2006]. Cette distance fait ensuite office de récompense intrinsèque. Cependant cette approche est incapable de gérer la stochasticité locale de l'environnement [Burda *et al.*, 2019]. Par exemple, il s'avère qu'ajouter une télévision affichant aléatoirement des images dans un environnement 3D attire l'agent ; il va alors regarder passivement la télé puisqu'il sera incapable de prédire la prochaine observation. Ce problème est aussi appelé le *problème du bruit blanc* [Pathak *et al.*, 2017; Schmidhuber, 2010]. Une solution serait de s'assurer que les transitions entre états soient apprenables, *i.e.* que la transition n'est pas trop stochastique, mais cette problématique est difficile à résoudre en pratique [Lopes *et al.*, 2012].

Le module de curiosité intrinsèque (ICM) [Pathak *et al.*, 2017] apprend la dynamique de l'environnement dans un espace de caractéristiques. Il construit d'abord une représentation des états en apprenant à prédire les actions réalisées dans un état à partir de cet état et de l'état d'ar-

rivée. Cela restreint la représentation à ce qui peut être contrôlé par l'agent. Il prédit ensuite dans cet espace l'état suivant. L'erreur de prédiction est alors utilisée comme récompense intrinsèque, ainsi l'erreur n'incorpore pas le bruit blanc puisque celui-ci ne dépend pas des actions. ICM permet notamment à un agent d'explorer son environnement dans les jeux *VizDoom* et *Super Mario Bros*. Dans *Super Mario Bros*, l'agent franchit 30% du premier niveau sans récompense extrinsèque. Finalement, en considérant toujours l'erreur de prédiction comme bonus intrinsèque, [Burda *et al.*, 2019] propose un résumé des différentes manières de définir un espace de caractéristiques et montre, d'une part, qu'utiliser des caractéristiques aléatoires peut être performant mais peu généralisable lorsque l'environnement change, d'autre part qu'utiliser l'espace brut d'états (e.g. les pixels) est inefficace. AR4E [Oh and Cavallaro,] réutilise le modèle ICM, mais encode l'action dans un grand espace avant de l'ajouter à l'état courant lors de la prédiction de l'état suivant. Cette astuce améliore ICM, mais il manque une analyse expliquant leurs résultats.

Nouveauté d'un état. Il existe une large littérature sur la mesure de la nouveauté d'un état comme motivation intrinsèque. Une première idée a été d'ajouter un bonus intrinsèque lorsque l'agent se dirige dans un état dans lequel il ne va jamais [Brafman and Tennenholtz, 2002; Kearns and Singh, 2002] ; ces méthodes sont dites basées sur le comptage. Au fur et à mesure qu'il visite un état, la récompense intrinsèque liée à cet état baisse. Bien que cette méthode soit efficace dans un environnement tabulaire (avec un espace d'états discretisé), elle est difficilement applicable lorsque les états sont très nombreux ou continus puisqu'on ne retourne alors jamais dans un même état.

Une première solution proposée par [Tang *et al.*, 2017], nommée TRPO-AE-hash, est de faire un hashage de l'espace d'états lorsqu'il est continu ou très grand. Cependant, les résultats ne sont que légèrement meilleurs que ceux d'une politique d'exploration classique. D'autres tentatives d'adaptation à un très grand espace d'états ont été proposées, comme DDQN-PC [Bellemare *et al.*, 2016], A3C+ [Bellemare *et al.*, 2016] ou DQN-PixelCNN [Ostrovski *et al.*, 2017], qui reposent sur des modèles de densité [Van den Oord *et al.*, 2016; Bellemare *et al.*, 2014]. Ces modèles permettent de calculer le *pseudo-count* [Bellemare *et al.*, 2016], adaptation du comptage permettant la généralisation du décompte d'un état auprès des états avoisinants. Bien que ces algorithmes fonctionnent sur des environnements avec des récompenses éparses, les modèles de densité rajoutent une importante couche de complexité calculatoire [Ostrovski *et al.*, 2017]. Par ailleurs, même si ces modèles gèrent des espaces d'états de très haute dimension (e.g. pixels), ils ne peuvent pas être utilisés avec des espaces d'états continus.

Afin de diminuer la complexité calculatoire induite par les modèles de densité, ϕ-EB [Martin *et al.*, 2017] propose de ne pas modéliser la densité sur l'espace d'états brut, mais sur un espace de caractéristiques dénombrable induit par le calcul de $V(s)$. Plus indirectement, DQN+SR [Machado *et al.*, 2018] utilise la norme de la représentation successeure [Kulkarni *et al.*, 2016b] comme récompense intrinsèque. Pour justifier ce choix, les auteurs expliquent que ce bonus est corrélé au décompte.

Astucieusement, DORA l'exploratrice [Fox *et al.*, 2018] utilise un autre MDP ne contenant aucune récompenses. La valeur d'un état dans ce MDP est biaisée de manière optimiste, de sorte qu'elle baisse au fur et à mesure que l'agent la met à jour. La valeur calculée est utilisée comme approximateur du décompte d'état. Bien que l'approche s'utilise naturellement dans un espace d'états continus, il manque des expérimentations pour la comparer aux approches existantes [Bellemare *et al.*, 2016].

RND [Burda *et al.*, 2018] mesure la nouveauté d'un état en distillant un réseau de neurones aléatoire (dont les poids sont figés) dans un autre réseau de neurones apprenant. Pour chaque état, le réseau de neurones aléatoire génère des caractéristiques aléatoires. Le second réseau apprend à reproduire la sortie du réseau aléatoire pour chaque état. L'erreur de prédiction fait office de récompense. Cela revient à récompenser la nouveauté d'un état puisque l'erreur sera importante tant que le second réseau aura peu visité l'état en question, et sera faible lorsqu'il aura beaucoup appris dessus. L'agent n'arrive cependant pas à apprendre une exploration à long terme. Par exemple dans *Montezuma's revenge*, l'agent utilise ses clés pour ouvrir les premières portes qu'il voit mais n'arrive pas à accéder aux deux dernières portes. De plus, les caractéristiques aléatoires peuvent être insuffisantes pour représenter la richesse d'un environnement.

Nouveauté comme écart aux autres états. Une autre manière d'évaluer la nouveauté d'un état est de l'estimer comme la distance aux états habituellement parcourus. L'exploration informée de [Oh *et al.*, 2015] utilise un modèle de l'environnement pour prédire quelle action lui permettra d'aller dans un état différent des d derniers états visités. Les auteurs utilisent pour cela un noyau gaussien. Cependant ils n'utilisent pas cette distance comme récompense intrinsèque mais comme moyen de choisir l'action en lieu et place de l'aléatoire dans la stratégie ϵ-greedy. Il serait intéressant de l'évaluer comme récompense intrinsèque. EX^2 [Fu *et al.*, 2017] apprend un discriminateur pour différencier les états entre eux : lorsque le discriminateur n'arrive pas à différencier l'état courant de ceux d'un buffer, cela veut dire qu'il est peu allé dans cet état et l'agent recevra un bonus intrinsèque, et inversement lorsqu'il arrive à différencier l'état.

Le module de curiosité épisodique ECO [Savinov *et al.*, 2018] approfondit cette idée en s'inspirant de la mémoire épisodique. Le modèle proposé contient un module de comparaison capable de renvoyer un bonus si l'agent est proche ou loin des états contenus dans un buffer. Ainsi, il calcule la probabilité que le nombre d'actions nécessaire pour aller de l'état sélectionné dans le buffer à l'état

courant soit inférieure à un seuil. En stockant des états éparses dans le buffer, l'agent pose des points de repères dans l'environnement et essaye de s'éloigner de ceux-ci, cela revient à partitionner l'environnement. Le probabilité que l'agent soit écarté de chaque état du buffer est utilisée pour calculer la récompense intrinsèque. Ce modèle a été appliqué sur des environnements 3D comme *DMLab* [Beattie *et al.*, 2016] ou *VizDoom* [Kempka *et al.*, 2016] et permet à l'agent d'explorer l'ensemble de son environnement. Cependant pour calculer la récompense intrinsèque, l'agent doit comparer son observation courante à tous ses états en mémoire. Un passage à l'échelle de cette méthode risque d'être difficile lorsque la richesse de l'espace d'état deviendra plus grande, il serait en effet plus compliqué de partitionner efficacement l'espace d'état. D'un autre côté, l'avantage de cette méthode est que l'agent ne subit pas l'effet du bruit blanc (cf. §4.1).

Dans ces méthodes basées sur le calcul de la nouveauté d'un état, [Stanton and Clune, 2018] distingue la nouveauté inter-épisodes, utilisée par A3C+ [Bellemare *et al.*, 2016] et EX^2 [Fu *et al.*, 2017], et la nouveauté intra-épisodes, que l'on retrouve dans le module de curiosité épisodique ECO [Savinov *et al.*, 2018] et dans l'exploration informée [Oh *et al.*, 2015]. Typiquement, une nouveauté intra-épisodes remettra à zéro les décomptes d'états au début de chaque épisode. Cela pourrait être une piste pour pallier à la difficulté de RND [Burda *et al.*, 2018] concernant l'exploration à long terme.

4.2 *Empowerment*

L'*empowerment* a été développé pour répondre à la question suivante : existe-t-il une fonction d'utilité locale qui rende compte de la survie d'un organisme [Klyubin *et al.*, 2005; Salge *et al.*, 2014b] ? Cette hypothétique fonction devrait être locale dans le sens où elle n'affecte pas le comportement de l'organisme sur le très long terme (la mort en elle-même n'affecte pas cette fonction par exemple) et les comportements induits doivent favoriser la survie de l'espèce. Typiquement, cette fonction peut expliquer la tendance d'un animal à vouloir dominer sa meute, et plus généralement l'envie d'un humain d'acquérir un statut social, d'avoir plus d'argent ou d'être plus fort, le besoin d'avoir un important taux de sucre ou la peur d'être blessé [Klyubin *et al.*, 2005; Salge *et al.*, 2014a]. Chacune de ces motivations permet d'élargir les possibilités d'action de l'agent, et par là son influence : une personne riche pourra faire plus de choses qu'une personne pauvre. Ces motivations sont par ailleurs locales, dans le sens où la récompense est presque immédiate. [Klyubin *et al.*, 2005] nomme cette capacité de contrôle de l'environnement l'*empowerment* d'un agent.

Définition. L'*empowerment* est défini avec la théorie de l'information. Il interprète la boucle d'interaction comme un envoi d'information dans l'environnement : une action est un signal envoyé et l'observation est le signal reçu. Plus l'action est informative sur les observations suivantes, plus l'*empowerment* est élevé. L'*empowerment* est mesuré comme la capacité d'un canal reliant les actions et les observations de l'agent. Soit $a_t^n = (a_t, a_{t+1}, ..., a_{t+n})$ les actions réalisées par l'agent de l'instant t à $t+n$ et s_{t+n} l'état de l'environnement à l'instant $t+n$. L'*empowerment* de l'état s_t, noté $\Sigma(s_t)$, est alors défini par :

$$\Sigma(s_t) = \max_{p(a_t^n)} I(a_t^n; s_{t+n}|s_t) \tag{13}$$

$$= \max_{p(a_t^n)} H(a_t^n|s_t) - H(a_t^n|s_{t+n}, s_t). \tag{14}$$

Maximiser l'*empowerment* revient à rechercher l'état dans lequel l'agent a le plus de contrôle sur l'environnement. Typiquement, le second terme de l'équation 14 permet à l'agent d'être sûr de là où il va, tandis que le premier terme insiste sur la diversité des états accessibles. Pour une vue d'ensemble sur les manières de calculer l'*empowerment*, le lecteur peut se référer à [Salge *et al.*, 2014b]. Nous nous focalisons dans la suite sur l'utilisation de l'*empowerment* dans le cadre de l'AR. Les travaux utilisant l'*empowerment* en dehors de ce contexte, e.g. [Karl *et al.*, 2017; Guckelsberger *et al.*, 2016; Capdepuy *et al.*, 2007; Salge *et al.*, 2014b], ne sont pas détaillés dans cet article.

L'*empowerment* en tant que récompense intrinsèque. En AR, l'agent maximisant l'*empowerment* est donc récompensé s'il se dirige dans des zones où il contrôle son environnement. Comme l'objectif de l'agent est de maximiser la fonction de récompense intrinsèque, celle-ci est définie par :

$$\begin{aligned} R_{int}(s, a, s') &= \Sigma(s') \\ &\approx -\mathbb{E}_{\omega(a|s)} \log \omega(a|s) \\ &\quad + \mathbb{E}_{p(s'|a,s)\omega(a|s)} \log p(a|s, s'). \end{aligned} \tag{15}$$

où $\omega(a|s)$ est la distribution choisissant les actions a_t^n. Dans l'idéal, $\omega(a|s)$ est la distribution maximisant l'équation 15 conformément à l'équation 14.
Le problème est que $p(a|s, s')$ est difficile à obtenir car il nécessite $p(s'|a, s)$ ce qui implique l'utilisation d'un modèle de densité.

[Mohamed and Rezende, 2015] propose de calculer l'*empowerment* en approximant l'équation 15. Pour cela, il calcule une borne inférieure à cette information mutuelle, cette borne sera ensuite reprise par plusieurs travaux :

$$I(a; s'|s) \geq H(a|s) + \mathbb{E}_{p(s'|a,s)\omega(a|s)} \log q_\xi(a|s, s'). \tag{16}$$

Son idée est de faire apprendre l'approximateur q_ξ de la distribution de probabilité $p(a|s, s')$ de manière supervisée sur les données que reçoit l'agent de l'environnement en utilisant la méthode du maximum de vraisemblance. Son approche permet de généraliser le calcul de l'*empowerment* à des observations continues. Dans ces travaux, les expérimentations montrent que la

maximisation de l'*empowerment* est notamment utile dans des environnements dynamiques, c'est-à-dire des environnements qui modifient l'état de l'agent même s'il fait une action stationnaire. Il prend l'exemple de l'environnement classique proie-prédateur : la proie est l'apprenant et a intérêt à éviter de se faire attraper car si elle meurt, elle n'aura plus aucun contrôle sur ses états suivants. Implicitement, la proie évite donc de mourir en maximisant son *empowerment*. Au contraire d'un environnement dynamique, un environnement statique admet une politique optimale statique (l'agent ne bouge plus lorsqu'il a trouvé le meilleur état) rendant l'*empowerment* comme récompense intrinsèque moins intéressant vis-à-vis d'une tâche. Les expérimentations proposées dans [Mohamed and Rezende, 2015] utilisent cependant la planification pour estimer l'*empowerment* et non des interactions avec l'environnement pour récupérer les données, ce qui implique l'utilisation d'un modèle de l'environnement.

[Gregor *et al.*, 2016] essaie de maximiser l'*empowerment* via des interactions avec l'environnement en utilisant $\omega(a|s) = \pi(a|s)$. La récompense devient alors :

$$R(a, h) = -\log \pi(a|h) + \log \pi(a|s', h) \qquad (17)$$

où h est l'historique d'observation (dont l'observation courante) et d'actions. Ses expérimentations sur divers environnements montrent que les trajectoires apprises mènent à des zones diverses et qu'un pré-entrainement via l'*empowerment* aide à apprendre une tâche. Les tâches apprises restent cependant relativement simples.

L'*empowerment* peut aussi se révéler intéressant en AR multi-agents. L'AR multi-agents fonctionne similairement à l'AR mono-agent, sauf que plusieurs agents apprennent simultanément à résoudre une tâche et doivent se coordonner. [Jaques *et al.*, 2018] a montré que dans un jeu non coopératif de type dilemme social [Leibo *et al.*, 2017], la récompense intrinsèque d'un agent pouvait stabiliser l'apprentissage en compensant la baisse de récompense individuelle due à une politique maximisant la récompense long-terme de l'ensemble des agents.

Conclusion. La principale difficulté de l'utilisation de l'*empowerment* en AR est donc son calcul. La plupart des approches utilisent un modèle de l'environnement pour calculer la récompense intrinsèque basée sur l'*empowerment* [Mohamed and Rezende, 2015; de Abril and Kanai, 2018]. Cependant le coeur même de l'AR est que l'agent ne connaît pas *a priori* la dynamique de l'environnement ou la fonction de récompense. Les travaux existants dans ce contexte restent donc limités et ne suffisent pas à montrer le potentiel de l'*empowerment* pour aider l'apprentissage.
Il est intéressant de noter que l'*empowerment* peut pousser un agent à apprendre des comportements même dans un environnement *a priori* statique. En effet, supposons que l'agent choisisse non pas des actions primitives directement, mais des *options*. S'il n'a pas encore appris les *options*, il est incapable de les discerner, c'est donc comme si l'agent n'avait aucun contrôle sur son environnement. Si au contraire, ses *options* sont parfaitement distinguées dans l'espace d'état, l'agent a le contrôle de son environnement. Or il s'agit là non pas de choisir les états maximisant l'*empowerment* mais de définir des *options* qui augmentent l'*empowerment*. Nous revenons sur ce point dans la partie 4.3.

4.3 Génération d'objectifs

La génération d'objectifs est la capacité d'un agent à apprendre des compétences diverses de manière non supervisée. Les compétences ou objectifs générés par l'agent sont des *options* (cf. §3.1). Comparé à l'AR multi-objectifs [Liu *et al.*, 2015], les compétences sont ici générées de manière non supervisée. Dans les travaux utilisant la génération d'objectifs, l'agent apprend généralement sur deux échelles de temps : il génère des *options* et apprend les politiques intra-option associées en utilisant une récompense intrinsèque ; si un objectif global existe, il apprend à utiliser ces compétences pour réaliser cet objectif global, appelé tâche, en utilisant la récompense extrinsèque associée à la tâche. L'intérêt est d'une part d'apprendre des compétences intéressantes pouvant servir à plusieurs tâches, or elles le seront d'autant plus qu'elles sont décorrélées de la tâche apprise [Heess *et al.*, 2016]. D'autre part, l'abstraction temporelle des actions réalisée via les compétences facilite l'apprentissage. Prenons l'exemple de MuJoCo [Todorov *et al.*, 2012] qui est un environnement souvent utilisé dans les travaux sur la génération d'objectifs. Dans cet environnement, les articulations d'un robot peuvent être contrôlées par un agent pour accomplir par exemple des tâches de locomotion. L'idée de certains travaux est donc de générer des compétences de type *avancer* ou *reculer* avec une récompense intrinsèque. Ces compétences peuvent alors servir à une tâche de navigation.

Classiquement, l'AR a un seul objectif et n'apprend pas à réaliser plusieurs objectifs. Une manière de généraliser l'AR profond à l'apprentissage de plusieurs objectifs, voir à tous les objectifs possibles dans l'espace d'état, est d'utiliser l'approximateur de fonction de valeur universelle (UVFA) [Schaul *et al.*, 2015]. UVFA intègre la représentation de l'état objectif dans l'observation de l'agent. La politique trouvée est alors conditionnée sur l'objectif : $\pi(s)$ devient $\pi(s, g)$ où g est un objectif. La même idée peut être retrouvée avec la recherche de politiques contextuelles [Fabisch and Metzen, 2014]. Ainsi, les travaux apprenant les *options* avec une motivation intrinsèque apprennent des politiques $\pi(s, g)$. Bien que l'espace d'exploration augmente alors, [Andrychowicz *et al.*, 2017] améliore l'efficacité des données en apprenant sur plusieurs objectifs à la fois via une seule interaction.

En effet, via une interaction (s, s', r_g, a, g), il est possible de créer une nouvelle interaction avec un nouvel objectif (lequel serait réussi) $(s, s', r_{g'}, a, g')$ tant qu'une fonction de récompense $R(s, a, s', g)$ est accessible, ce qui est généralement le cas lorsque la récompense est intrinsèque.

Dans la suite, nous allons présenter plusieurs travaux incorporant des récompenses expertes dans un algorithme hiérarchique. Ensuite nous étudierons les deux principaux ensembles de travaux portant sur l'auto-génération d'objectifs. La première approche utilise l'espace d'états pour générer les objectifs et calculer la récompense intrinsèque ; la seconde utilise la théorie de l'information.

Entre récompenses expertes et récompenses intrinsèques. Il existe quelque travaux précurseurs montrant l'intérêt de la décomposition hiérarchique des actions. Parmi ceux-ci, [Kulkarni *et al.*, 2016a] présente le modèle *hierarchical-DQN* dans lequel la représentation des objectifs est définie de manière experte via des tuples $(entite, relation, entite2)$. Une entité peut être un objet à l'écran ou l'agent, et la relation renvoie notamment à une distance. Ainsi l'objectif peut être que l'agent atteigne un objet. La récompense experte vaut simplement un si l'objectif est atteint, zéro sinon. Il montre que cela peut aider l'apprentissage notamment lorsque les récompenses sont éparses comme dans *Montezuma's revenge*. Cependant, s'affranchir de l'apprentissage de la représentation des compétences revient à esquiver le problème principal : il est en effet difficile de choisir quelles caractéristiques sont assez intéressantes pour être des objectifs dans un grand espace d'état. D'autres travaux démontrent le potentiel de l'approche en utilisant des objectifs auxiliaires spécifiques à la tâche [Riedmiller *et al.*, 2018] ou plus abstraits [Dilokthanakul *et al.*, 2019; Rafati and Noelle, 2019]. Plus particulièrement, une heuristique régulièrement utilisée pour générer une compétence est la recherche d'états faisant office de goulots d'étranglements [McGovern and Barto, 2001; Menache *et al.*, 2002]. Il s'agit d'identifier des états charnière quant aux prochains états visités (par exemple, une porte). De récents travaux [Zhang *et al.*, 2019; Tomar *et al.*, 2018] utilisent la représentation successeure [Kulkarni *et al.*, 2016b] pour généraliser l'approche à des espaces d'états continus. Le désavantage de ce type de travaux est que les récompenses ne sont pas suffisamment générales pour s'appliquer à tous les environnements et nécessitent une expertise.

Distance entre objectifs. Certains travaux utilisent l'espace d'états pour créer un espace d'objectif, l'intérêt est de pouvoir se servir de chaque état comme objectif. Ainsi, *Hierarchical Actor-Critic* (HAC) [Levy *et al.*, 2019] se sert directement de l'espace d'états comme espace d'objectif pour apprendre trois niveaux d'options (les *options* du second niveau sont choisies de manière à répondre à l'*option* du troisième niveau). Une distance entre l'objectif et l'état final fait donc office de récompense intrinsèque. Au contraire, HIRO [Nachum *et al.*, 2018] utilise comme objectif la différence entre l'état initial et l'état à la fin de l'objectif ; la récompense intrinsèque est alors une distance entre la direction prise et l'objectif. Les objectifs permettent ainsi d'orienter les compétences vers certaines zones spatiales. Cependant, il y a deux problèmes dans l'utilisation de l'espace d'états comme espace d'objectifs. D'une part une distance (comme L2) a peu de sens dans un espace très grand comme une image composée de pixels, d'autre part il est difficile de faire fonctionner un algorithme d'AR sur un espace d'action trop grand. Concrètement, un algorithme ayant comme espace d'objectif des images peut impliquer pour la politique d'*options* un espace d'action de 84x84 dimensions. Un espace d'action aussi large est actuellement inconcevable, aussi, ces algorithmes ne fonctionnent que sur des espaces d'états de faible dimension. Pour pallier ce problème, FuN [Vezhnevets *et al.*, 2017] utilise comme récompense intrinsèque la direction prise dans un espace de caractéristiques d'états. Les caractéristiques utilisées sont celles qui sont utiles à la tâche, elles sont donc construites par la rétro-propagation de la récompense extrinsèque. L'agent apprend donc uniquement des compétences liées à la tâche et a besoin d'un accès à la récompense extrinsèque.

Le problème est finalement de construire une bonne représentation de l'espace d'états [Schwenker and Palm, 2019] faisant office d'espace d'objectifs, i.e. choisir les informations à ne pas perdre lors de la compression des états en une nouvelle représentation. Ainsi, la distance entre deux objectifs aurait un sens et serait une bonne récompense intrinsèque. Pour construire un espace de caractéristiques, RIG [Nair *et al.*, 2018] se sert d'un auto-encodeur variationel (VAE) [Kingma and Welling, 2013], mais ce type d'approche peut-être très sensible à des distracteurs (i.e. des caractéristiques inutiles à la tâche ou l'objectif, présentes dans les états) et ne permet donc pas de pondérer les caractéristiques correctement. [Zhou *et al.*, 2019] utilise quant à lui des méthodes non supervisées comme l'algorithme de *slow features analysis* [Wiskott and Sejnowski, 2002] et le réseau *growing when required* [Marsland *et al.*, 2002] pour construire une carte topologique. Un agent hiérarchique se sert ensuite des noeuds de la carte comme espace d'objectif. Ils ne se comparent cependant pas aux autres approches. Sub-optimal representation learning [Nachum *et al.*, 2019] essaye de borner la sous-optimalité de la représentation des objectifs, offrant des garanties théoriques. L'agent s'avère capable d'apprendre à aller partout en sélectionnant les caractéristiques importantes pour la tâche. Cependant, comme FuN [Vezhnevets *et al.*, 2017], l'agent a besoin d'une récompense dense. [Florensa *et al.*, 2019] reformule l'équation de Bellman et présente des perspectives intéressantes en apprenant une représentation d'états pour laquelle la distance L2 entre deux états correspond au nombre d'actions à effectuer pour aller d'un état à l'autre. Il manque cependant des expérimentations montrant son intérêt.

Information mutuelle entre objectif et trajectoire. Une deuxième approche ne nécessite pas de fonction de distance mais consiste à maximiser l'information mutuelle entre un objectif et sa trajectoire associée. Informellement, il s'agit d'apprendre des compétences selon la capacité de l'agent à les distinguer entre elles à partir de la trajectoire (i.e. les états parcourus) de la politique de la compétence choisie. Dans cette section, nous appelons $I(g;c)$ l'information mutuelle entre g, un objectif, et c, une partie (changeante selon les travaux) de la trajectoire.

SNN4HRL [Florensa *et al.*, 2017] apprend des compétences en maximisant l'information mutuelle $I(g;c)$ où c est un agrégé de la trajectoire. Chaque objectif est généré de manière uniforme donc maximiser l'information mutuelle revient à minimiser $H(g|c)$ (cf. équation 7). Or, c'est équivalent de maximiser la récompense intrinsèque $\log q(g|c)$ (équation 16), où q est la probabilité prédite par un modèle. Pour calculer cet élément, il discrétise l'espace d'états et calcule la probabilité d'accomplir son objectif courant dans son état courant. Pour calculer cette probabilité, l'agent compte le nombre de fois où il a parcouru cette partition pour chaque objectif. Ensuite, l'agent ayant appris les compétences est intégré dans une structure hiérarchique dans laquelle un manager choisit les objectifs à accomplir. Notons que l'espace d'objectifs est ici discret.

VALOR [Achiam *et al.*, 2018] et DIAYN [Eysenbach *et al.*, 2018] reprennent la même idée, mais se distinguent des travaux précédents en utilisant un réseau de neurone plutôt qu'une discrétisation pour calculer $\log q(g|c)$ et en choisissant c comme une partie de la trajectoire de la compétence dans l'environnement. Ils réussissent à faire apprendre à un agent des tâches de locomotion dans des espaces d'état ayant plus de 100 degrés de liberté. De plus, ils montrent l'intérêt de cette méthode utilisée comme pré-entrainement pour de l'apprentissage hiérarchique et comme initialisation pour l'apprentissage d'une tâche. DIAYN choisit c comme un état de la trajectoire et calcule la récompense intrinsèque à chaque itération de la trajectoire. VALOR se distingue en considérant c comme un agrégé de la trajectoire d'états et en assignant la récompense à la fin de la trajectoire. Avec VALOR, l'agent parvient à apprendre jusqu'à 10 compétences différentes et jusqu'à 100 en augmentant peu à peu le nombre d'objectifs via un curriculum [Achiam *et al.*, 2018]. VIC [Gregor *et al.*, 2016] avait déjà fait quelques expériences avec la même approche, mais sur des environnements plus simples et sans exhiber la même diversité de comportements.

Deux principales limites à cette approche peuvent être distinguées :

1. l'agent est incapable d'apprendre à générer des objectifs sans désapprendre ses compétences. Ainsi, la distribution d'objectifs générée par l'agent doit rester uniforme [Gregor *et al.*, 2016; Eysenbach *et al.*, 2018].
2. Le calcul de $\log q(g|c)$ implique par ailleurs que l'espace d'objectifs soit discrétisé. Il est donc impossible d'utiliser des *options* continues. DISCERN [Warde-Farley *et al.*, 2018] s'attaque à ce problème en considérant l'espace d'objectifs comme l'espace d'états. Ensuite il fait une approximation de $\log q(g|c)$ en essayant de classifier l'état final de la trajectoire auprès du bon objectif parmi d'autres objectifs tirés de la même distribution que le vrai objectif. Cela revient à apprendre à trouver l'objectif le plus proche de l'état final depuis une liste d'objectifs.

Nous avons énoncé au §4.2 que l'*empowerment* d'un agent s'améliorait au fur à et mesure que ses compétences se distinguaient. Les travaux présentés ici augmentent implicitement l'*empowerment* d'un agent, du point de vue de la politique d'options. En effet ils maintiennent une entropie sur les objectifs élevées et associent une direction dans l'espace d'état à un objectif. Ainsi, $H(a|s)$ est maximale, puisque la distribution de probabilité est uniforme, et $H(a|s,s')$ se réduit au fur et à mesure que l'agent apprend à distinguer les *options*.

Choisir la compétence à apprendre en estimant le progrès d'apprentissage. D'autres travaux sont à l'intersection entre l'apprentissage de compétences et le progrès d'apprentissage, l'idée est de choisir l'objectif qui ne soit ni trop difficile, ni trop facile pour faciliter l'apprentissage de l'agent. Goal GAN [Florensa *et al.*, 2018] apprend à générer des objectifs de plus en plus complexes (via un GAN [Goodfellow *et al.*, 2014]) de manière à ce que l'agent sache aller partout par la suite, mais utilise une fonction de récompense manuelle : il suppose l'accès aux coordonnées de l'agent pour calculer une récompense binaire. Nous ne détaillerons pas plus ces travaux qui supposent pour la plupart que la représentation de l'objectif et la fonction de récompense associée sont donnés [Fabisch and Metzen, 2014; Deisenroth *et al.*, 2013].

Conclusion. Pour résumer, il existe deux principaux ensembles de travaux portant sur l'auto-génération d'objectifs. Le premier ensemble apprend ses objectifs à l'aide de l'espace d'états, l'avantage est alors d'avoir un espace continu et lisse d'objectifs permettant l'interpolation. Le désavantage est qu'il faut trouver la bonne métrique de comparaison et la juste manière de compresser l'espace d'états. Le second ensemble de travaux utilise la théorie de l'information, cela force l'utilisation d'un espace discret d'objectifs mais évite les complications associées à un espace continu qui est originellement de haute dimension.

4.4 Conclusion

D'autres approches incorporant la motivation intrinsèque dans l'AR ne sont pas directement liées à ces trois grandes catégories de travaux et ne sont donc pas détaillées dans cet article [Hester and Stone, 2017; Machado *et al.*, 2017a; Lakshminarayanan *et al.*, 2016; Machado *et al.*, 2017b; Stanton and Clune, 2018; Still and Precup, 2012; Little and Sommer, 2013; Frank *et al.*, 2014; Montúfar *et al.*, 2016].

De même que les méthodes essayant de combiner plusieurs motivations intrinsèques [Hester and Stone, 2017; de Abril and Kanai, 2018], qui nécessiteraient cependant des expérimentations sur des environnements plus compliqués.

5 Analyse des travaux

Dans cette section, nous allons étudier les points communs entre les travaux présentés afin de mettre en avant des perspectives de recherche.

5.1 Information mutuelle comme point commun

Une redondance semble apparaître tout au long de notre étude des différents travaux, que ce soit via l'acquisition de connaissance, l'*empowerment* ou l'apprentissage d'objectifs. L'information mutuelle apparait comme centrale pour élargir les capacités d'un agent.

Utilisation directe de l'information mutuelle. Nous avons d'abord vu que l'*empowerment* est entièrement défini via l'information mutuelle (cf. §4.2). VIME [Houthooft *et al.*, 2016] et AKL [Achiam and Sastry, 2017] maximisent le gain d'information, c'est-à-dire l'information que contient le prochain état sur le modèle de l'environnement. Plusieurs travaux sur la génération d'objectifs [Eysenbach *et al.*, 2018; Achiam *et al.*, 2018; Florensa *et al.*, 2017] utilisent directement l'information mutuelle entre la trajectoire issue d'un objectif et l'objectif en lui-même pour récompenser l'agent. [Still and Precup, 2012] suggère que l'agent doit maximiser l'information mutuelle entre son action et les état suivants pour améliorer la politique d'exploration.

Fonctions équivalentes à l'information mutuelle. L'erreur de prédiction [Nachum *et al.*, 2019; 2018; Pathak *et al.*, 2017] est aussi relative à l'information mutuelle [de Abril and Kanai, 2018], puisque la prédiction de l'état suivant un couple (état,action) peut être correcte seulement si l'action contient de l'information sur l'état suivant. Notons toutefois que les états suivants possibles doivent être suffisamment divers. Par ailleurs, [Nachum *et al.*, 2019] explique que sa méthode apprend une représentation d'états maximisant l'information mutuelle entre l'état en question et les états suivants. Finalement, [Bellemare *et al.*, 2016] a montré que la récompense issue du *pseudo-count* [Bellemare *et al.*, 2016; Ostrovski *et al.*, 2017] est proche de celle du gain d'information.

Généralisation des modèles utilisés. Globalement, les modèles étudiés ont souvent comme point commun de contenir deux modules :

1. Le premier est un module cherchant une fonction d'évaluation entre les actions et états parcourus par l'agent (sa dernière trajectoire [Eysenbach *et al.*, 2018; Achiam *et al.*, 2018], sa dernière action [Stadie *et al.*, 2015; Pathak *et al.*, 2017; Burda *et al.*, 2019], le nombre de fois où chaque état a été parcouru [Bellemare *et al.*, 2016; Ostrovski *et al.*, 2017] ou les dernières trajectoires [Savinov *et al.*, 2018; Fu *et al.*, 2017; Oh *et al.*, 2015] ...) et une autre source de données (un objectif [Eysenbach *et al.*, 2018; Achiam *et al.*, 2018], les états suivants [Pathak *et al.*, 2017; Mohamed and Rezende, 2015] ...). Cette fonction est souvent une fonction de causalité.
2. Le deuxième est une politique maximisant une récompense intrinsèque issue de la fonction d'évaluation.

5.2 Motivation intrinsèque comme compression de l'information

Compression de l'information. Schmidhuber propose que l'organisme est guidé par le désir de compresser l'information qu'il reçoit [Schmidhuber, 2008]. Ainsi, plus nous arrivons à compresser les données reçues de l'environnement, plus la récompense intrinsèque reçue est élevée. Il note toutefois que c'est l'amélioration de la compression qui est importante et non le degré de compression en lui même, sous peine qu'un agent soit inactif devant du bruit ou devant une obscurité uniforme. Or la compression de données est fortement liée à l'observation de régularités dans ces mêmes données. Par exemple ce que nous appelons visage est, dans notre environnement, un ensemble, apparaissant de manière récurrente, composé d'une forme ovale contenant deux yeux, un nez et une bouche. De même, un état de l'environnement peut être décrit avec quelques-unes des caractéristiques les plus pertinentes. Cela implique que la motivation intrinsèque se traduit par une recherche de nouvelles régularités dans l'environnement.

Lien avec les travaux. Il a été montré dans [Schmidhuber, 2008; Houthooft *et al.*, 2016] que les travaux sur le gain d'information sont directement liés au progrès de la compression d'information ; le module de curiosité épisodique [Savinov *et al.*, 2018] essaye d'encoder l'environnement en sauvegardant les états les plus diversifiés possibles ; et les modèles prédictifs [Burda *et al.*, 2019] encodent les dynamiques de l'environnement dans un modèle paramétré (souvent un réseau de neurone). L'*empowerment* est similaire, rappelons qu'il s'agit de diriger l'agent vers une zone dans laquelle il a du contrôle, i.e. que les états sont déterminés par les actions de l'agent. Il est possible de reformuler l'*empowerment* comme l'intérêt d'un agent pour des zones où ses actions sont une compression des états suivants. En effet, l'*empowerment* est maximal si chaque trajectoire mène à ses propres états (toujours les mêmes dans le même ordre) distincts de ceux des autres trajectoires ; tandis qu'il est minimal si toutes les trajectoires mènent à un même état. Les travaux sur la génération d'objectifs cherchent explicitement à compresser des trajectoires dans un espace d'objectifs. Enfin, une partie des travaux [Vezhnevets *et al.*, 2017; Nachum *et al.*, 2019; Pathak *et al.*, 2017] repose sur la qualité d'une compression

de l'espace d'état.

5.3 Conclusion

Pour résumer, l'ensemble des travaux s'attache à compresser les nouvelles régularités détectées dans les trajectoires de l'agent. Pour cela, la théorie de l'information est un puissant outil de mesure.

6 Limites et challenges

Plusieurs travaux sont limités par des problématiques sortant du cadre de l'AR, telles que les performances des modèles de densité [Bellemare *et al.*, 2016; Ostrovski *et al.*, 2017], la difficulté d'approximation de l'information mutuelle entre deux variables aléatoires continues [Gregor *et al.*, 2016] ou les performances des modèles prédictifs [Nachum *et al.*, 2018; Nair *et al.*, 2018]. Ces limites dépassent le cadre de cet article. Aussi, malgré l'hétérogénéité des travaux sur la motivation intrinsèque en AR, et les limitations propres à chacune de ses méthodes, nous avons identifié et présentons dans cette section quatre problématiques majeures qui sont communes à l'ensemble des approches.

6.1 Stochasticité de l'environnement

Nous avons vu dans la partie précédente qu'il était intéressant de maximiser le progrès de compression et que la plupart des travaux étaient relatifs à la compression d'information, et non au progrès de la compression. Cet écart explique la difficulté de plusieurs travaux [Burda *et al.*, 2019] à gérer l'effet du bruit blanc [Schmidhuber, 2010] ou plus généralement la stochasticité de l'environnement. Certains travaux de l'état de l'art gèrent cette problématique [Savinov *et al.*, 2018; Pathak *et al.*, 2017; Burda *et al.*, 2018; Houthooft *et al.*, 2016], mais chacun avec ses défauts.

6.2 Acquisition de connaissance sur le long-terme

A notre connaissance, aucune approche existante n'est capable de gérer la recherche d'information long-terme [Burda *et al.*, 2018]. Dans *Montezuma's revenge*, il s'agit d'éviter d'utiliser une clé trop rapidement pour pouvoir l'utiliser plus tard. Dans la vie de tous les jours, il peut s'agir d'éviter de dépenser son argent trop vite. Cette difficulté pourrait être résolue avec une approche utilisant la planification [Hafner *et al.*, 2018]. L'apprentissage hiérarchique de compétences pourrait apporter une solution, en transformant le long terme en court terme via une hiérarchie de compétences multi-niveaux [Riemer *et al.*, 2018]. Finalement, il manque d'autres environnements que *Montezuma's revenge* mettant vraiment en avant cette problématique.

6.3 Construire une représentation des états

Nous avons vu que construire de bonnes caractéristiques d'état est important dans la découverte d'objectifs afin de travailler sur un espace d'objectifs réduit. C'est aussi primordial dans les travaux sur l'acquisition de connaissances pour avoir une erreur de prédiction significative. Le module de curiosité intrinsèque ICM [Pathak *et al.*, 2017] propose une représentation des états intéressante, restreinte à ce qui peut être contrôlé par l'agent, mais sa limite est que le module apprend une partie suffisante des caractéristiques permettant de déterminer l'action, et non l'ensemble des caractéristiques déterminées par l'action. Il manque encore un moyen de compresser parfaitement l'espace d'états dans l'espace des caractéristiques contrôlées par l'agent. [Florensa *et al.*, 2019] est allé dans ce sens en apprenant une représentation d'états pour laquelle la distance L2 entre deux états correspond au nombre d'actions à effectuer pour aller d'un état à l'autre, mais il manque des expérimentations. Par ailleurs, [Lesort *et al.*, 2018] présente plusieurs pistes pouvant améliorer la représentation des états d'un agent.

6.4 Décorréler les objectifs de la tâche

L'avantage de décorréler l'apprentissage des objectifs de l'apprentissage d'une tâche est de favoriser l'exploration et le transfert d'apprentissage. On parle alors d'apprentissage *bottom-up* car on apprend les compétences avant la tâche. Si cet apprentissage a fait des progrès significatifs, il est encore impossible d'apprendre des tâches spécifiques en même temps que les compétences permettant de les réaliser sans subir d'oubli catastrophique [McCloskey and Cohen, 1989; Florensa *et al.*, 2018]. En effet, lorsque l'agent apprend séquentiellement des tâches, il oublie les premières tâches en apprenant les suivantes. Des travaux adressant le problème d'oubli catastrophique existent déjà [Kirkpatrick *et al.*, 2017; Parisi *et al.*, 2019] mais ils n'ont, à notre connaissance, pas été évalués avec la motivation intrinsèque et un large nombre de tâches.

7 Conclusion

L'AR fait face à plusieurs défis, comme l'apprentissage avec des récompenses éparses ou l'abstraction des actions de l'agent en décisions de plus haut niveau. Nous avons vu que la motivation intrinsèque pouvait être utilisée en AR et que ses nombreuses applications pouvait résoudre partiellement ces problématiques. Plusieurs types de motivations intrinsèques existent comme méta-compétences, chacune avec leur littérature. Parmi celles-ci, l'acquisition de connaissances est effectuée via des modèles prédictifs, des modèles bayésiens ou des modèles de densité pour inciter l'agent à explorer l'environnement. L'*empowerment* est une motivation universelle poussant l'agent à avoir des comportements divers tels que la survie. La génération d'objectifs est directement liée à l'*empowerment* et permet d'abstraire les actions de l'agent, aidant à résoudre le *credit assignment problem*. Lorsque l'apprentissage est de type *bottom-up*, l'abstraction des décisions facilite l'exploration et le transfert d'apprentissage. Plusieurs défis restent cependant à adresser : les méchanismes d'acquisition de connaissance gèrent difficilement la stochasticité de l'environnement et ont encore des difficultés à générer des trajectoires utiles à l'exploration sur le long terme ; avoir des

représentations d'état plus significatives pourrait ouvrir de nouvelles perspectives pour la génération d'objectifs ; les travaux apprenant des compétences multi-tâches souffrent encore de l'oubli catastrophique. Notre analyse suggère que la théorie de l'information, en permettant de compresser l'information contenue dans les séquences d'interactions de l'agent, devrait jouer un rôle proéminent dans la résolution des défis mentionnés.

Références

[Achiam and Sastry, 2017] Joshua Achiam and Shankar Sastry. Surprise-based intrinsic motivation for deep reinforcement learning. *arXiv preprint arXiv :1703.01732*, 2017.

[Achiam *et al.*, 2018] Joshua Achiam, Harrison Edwards, Dario Amodei, and Pieter Abbeel. Variational option discovery algorithms. *arXiv preprint arXiv :1807.10299*, 2018.

[Amodei *et al.*, 2016] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv :1606.06565*, 2016.

[Andrychowicz *et al.*, 2017] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. In *Advances in Neural Information Processing Systems*, pages 5048–5058, 2017.

[Bacon *et al.*, 2017] Pierre-Luc Bacon, Jean Harb, and Doina Precup. The option-critic architecture. In *AAAI*, pages 1726–1734, 2017.

[Baldassarre and Mirolli, 2013] Gianluca Baldassarre and Marco Mirolli. Intrinsically motivated learning systems : an overview. In *Intrinsically motivated learning in natural and artificial systems*, pages 1–14. Springer, 2013.

[Barto *et al.*, 2004] Andrew G Barto, Satinder Singh, and Nuttapong Chentanez. Intrinsically motivated learning of hierarchical collections of skills. In *Proceedings of the 3rd International Conference on Development and Learning*, pages 112–19, 2004.

[Beattie *et al.*, 2016] Charles Beattie, Joel Z Leibo, Denis Teplyashin, Tom Ward, Marcus Wainwright, Heinrich Küttler, Andrew Lefrancq, Simon Green, Víctor Valdés, Amir Sadik, et al. Deepmind lab. *arXiv preprint arXiv :1612.03801*, 2016.

[Bellemare *et al.*, 2014] Marc Bellemare, Joel Veness, and Erik Talvitie. Skip context tree switching. In *International Conference on Machine Learning*, pages 1458–1466, 2014.

[Bellemare *et al.*, 2015] Marc G. Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment : An evaluation platform for general agents (extended abstract). In *IJCAI*, pages 4148–4152. AAAI Press, 2015.

[Bellemare *et al.*, 2016] Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems*, pages 1471–1479, 2016.

[Berlyne, 1965] Daniel E Berlyne. Structure and direction in thinking. 1965.

[Blundell *et al.*, 2015] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. *arXiv preprint arXiv :1505.05424*, 2015.

[Brafman and Tennenholtz, 2002] Ronen I Brafman and Moshe Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct) :213–231, 2002.

[Burda *et al.*, 2018] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. *arXiv preprint arXiv :1810.12894*, 2018.

[Burda *et al.*, 2019] Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A. Efros. Large-scale study of curiosity-driven learning. In *International Conference on Learning Representations*, 2019.

[Cangelosi and Schlesinger, 2018] Angelo Cangelosi and Matthew Schlesinger. From babies to robots : the contribution of developmental robotics to developmental psychology. *Child Development Perspectives*, 12(3) :183–188, 2018.

[Capdepuy *et al.*, 2007] Philippe Capdepuy, Daniel Polani, and Chrystopher L Nehaniv. Maximization of potential information flow as a universal utility for collective behaviour. In *2007 IEEE Symposium on Artificial Life*, pages 207–213. Ieee, 2007.

[Cesa-Bianchi *et al.*, 2017] Nicolò Cesa-Bianchi, Claudio Gentile, Gábor Lugosi, and Gergely Neu. Boltzmann exploration done right. In *Advances in Neural Information Processing Systems*, pages 6284–6293, 2017.

[Chiang *et al.*, 2019] Hao-Tien Lewis Chiang, Aleksandra Faust, Marek Fiser, and Anthony Francis. Learning navigation behaviors end-to-end with autorl. *IEEE Robotics and Automation Letters*, 4(2) :2007–2014, 2019.

[Cover and Thomas, 2012] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[de Abril and Kanai, 2018] Ildefons Magrans de Abril and Ryota Kanai. A unified strategy for implementing curiosity and empowerment driven reinforcement learning. *arXiv preprint arXiv :1806.06505*, 2018.

[Deisenroth *et al.*, 2013] Marc Peter Deisenroth, Gerhard Neumann, Jan Peters, et al. A survey on policy search for robotics. *Foundations and Trends® in Robotics*, 2(1–2) :1–142, 2013.

[Dilokthanakul *et al.*, 2019] Nat Dilokthanakul, Christos Kaplanis, Nick Pawlowski, and Murray Shanahan. Feature control as intrinsic motivation for hierarchical reinforcement learning. *IEEE transactions on neural networks and learning systems*, 2019.

[Eysenbach *et al.*, 2018] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need : Learning skills without a reward function. *CoRR*, abs/1802.06070, 2018.

[Fabisch and Metzen, 2014] Alexander Fabisch and Jan Hendrik Metzen. Active contextual policy search. *The Journal of Machine Learning Research*, 15(1) :3371–3399, 2014.

[Florensa *et al.*, 2017] Carlos Florensa, Yan Duan, and Pieter Abbeel. Stochastic neural networks for hierarchical reinforcement learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

[Florensa *et al.*, 2018] Carlos Florensa, David Held, Xinyang Geng, and Pieter Abbeel. Automatic goal generation for reinforcement learning agents. In *International Conference on Machine Learning*, pages 1514–1523, 2018.

[Florensa *et al.*, 2019] Carlos Florensa, Jonas Degrave, Nicolas Heess, Jost Tobias Springenberg, and Martin Riedmiller. Self-supervised learning of image embedding for continuous control. *arXiv preprint arXiv :1901.00943*, 2019.

[Fortunato *et al.*, 2017] Meire Fortunato, Mohammad Gheshlaghi Azar, Bilal Piot, Jacob Menick, Ian Osband, Alex Graves, Vlad Mnih, Remi Munos, Demis Hassabis, Olivier Pietquin, et al. Noisy networks for exploration. *arXiv preprint arXiv :1706.10295*, 2017.

[Fox *et al.*, 2018] Lior Fox, Leshem Choshen, and Yonatan Loewenstein. DORA the explorer : Directed outreaching reinforcement action-selection. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

[François-Lavet *et al.*, 2018] Vincent François-Lavet, Peter Henderson, Riashat Islam, Marc G Bellemare, Joelle Pineau, et al. An introduction to deep reinforcement learning. *Foundations and Trends® in Machine Learning*, 11(3-4) :219–354, 2018.

[Frank *et al.*, 2014] Mikhail Frank, Jürgen Leitner, Marijn Stollenga, Alexander Förster, and Jürgen Schmidhuber. Curiosity driven reinforcement learning for motion planning on humanoids. *Frontiers in neurorobotics*, 7 :25, 2014.

[Fu *et al.*, 2017] Justin Fu, John Co-Reyes, and Sergey Levine. Ex2 : Exploration with exemplar models for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2577–2587, 2017.

[Georgeon *et al.*, 2011] Olivier L Georgeon, James B Marshall, and Pierre-Yves R Ronot. Early-stage vision of composite scenes for spatial learning and navigation. In *2011 IEEE International Conference on Development and Learning (ICDL)*, volume 2, pages 1–6. IEEE, 2011.

[Ghavamzadeh *et al.*, 2015] Mohammad Ghavamzadeh, Shie Mannor, Joelle Pineau, Aviv Tamar, et al. Bayesian reinforcement learning : A survey. *Foundations and Trends® in Machine Learning*, 8(5-6) :359–483, 2015.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[Gopnik *et al.*, 1999] Alison Gopnik, Andrew N Meltzoff, and Patricia K Kuhl. *The scientist in the crib : Minds, brains, and how children learn.* William Morrow & Co, 1999.

[Graves, 2011] Alex Graves. Practical variational inference for neural networks. In *Advances in neural information processing systems*, pages 2348–2356, 2011.

[Gregor *et al.*, 2016] Karol Gregor, Danilo Jimenez Rezende, and Daan Wierstra. Variational intrinsic control. *arXiv preprint arXiv :1611.07507*, 2016.

[Guckelsberger *et al.*, 2016] Christian Guckelsberger, Christoph Salge, and Simon Colton. Intrinsically motivated general companion npcs via coupled empowerment maximisation. In *2016 IEEE Conference on Computational Intelligence and Games (CIG)*, pages 1–8. IEEE, 2016.

[Hafner *et al.*, 2018] Danijar Hafner, Timothy P. Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. *CoRR*, abs/1811.04551, 2018.

[Heess *et al.*, 2016] Nicolas Heess, Greg Wayne, Yuval Tassa, Timothy Lillicrap, Martin Riedmiller, and David Silver. Learning and transfer of modulated locomotor controllers. *arXiv preprint arXiv :1610.05182*, 2016.

[Hester and Stone, 2017] Todd Hester and Peter Stone. Intrinsically motivated model learning for developing curious robots. *Artificial Intelligence*, 247 :170–186, 2017.

[Hinton and Salakhutdinov, 2006] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786) :504–507, 2006.

[Houthooft *et al.*, 2016] Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. Vime : Variational information maximizing exploration. In *Advances in Neural Information Processing Systems*, pages 1109–1117, 2016.

[Hughes *et al.*, 2018] Edward Hughes, Joel Z Leibo, Matthew G Philips, Karl Tuyls, Edgar A Duéñez-Guzmán,

Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin R McKee, Raphael Koster, et al. Inequity aversion resolves intertemporal social dilemmas. *arXiv preprint arXiv :1803.08884*, 2018.

[Ito, 2016] Sosuke Ito. *Information thermodynamics on causal networks and its application to biochemical signal transduction*. Springer, 2016.

[Itti and Baldi, 2006] Laurent Itti and Pierre F Baldi. Bayesian surprise attracts human attention. In *Advances in neural information processing systems*, pages 547–554, 2006.

[Jaques *et al.*, 2018] Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro A Ortega, DJ Strouse, Joel Z Leibo, and Nando de Freitas. Intrinsic social motivation via causal influence in multi-agent rl. *arXiv preprint arXiv :1810.08647*, 2018.

[Karl *et al.*, 2017] Maximilian Karl, Maximilian Soelch, Philip Becker-Ehmck, Djalel Benbouzid, Patrick van der Smagt, and Justin Bayer. Unsupervised real-time control through variational empowerment. *arXiv preprint arXiv :1710.05101*, 2017.

[Kearns and Singh, 2002] Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2-3) :209–232, 2002.

[Kempka *et al.*, 2016] Michał Kempka, Marek Wydmuch, Grzegorz Runc, Jakub Toczek, and Wojciech Jaśkowski. Vizdoom : A doom-based ai research platform for visual reinforcement learning. In *2016 IEEE Conference on Computational Intelligence and Games (CIG)*, pages 1–8. IEEE, 2016.

[Kingma and Welling, 2013] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv :1312.6114*, 2013.

[Kirkpatrick *et al.*, 2017] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13) :3521–3526, 2017.

[Klyubin *et al.*, 2005] Alexander S Klyubin, Daniel Polani, and Chrystopher L Nehaniv. Empowerment : A universal agent-centric measure of control. In *Evolutionary Computation, 2005. The 2005 IEEE Congress on*, volume 1, pages 128–135. IEEE, 2005.

[Kulkarni *et al.*, 2016a] Tejas D Kulkarni, Karthik Narasimhan, Ardavan Saeedi, and Josh Tenenbaum. Hierarchical deep reinforcement learning : Integrating temporal abstraction and intrinsic motivation. In *Advances in neural information processing systems*, pages 3675–3683, 2016.

[Kulkarni *et al.*, 2016b] Tejas D Kulkarni, Ardavan Saeedi, Simanta Gautam, and Samuel J Gershman. Deep successor reinforcement learning. *arXiv preprint arXiv :1606.02396*, 2016.

[Lakshminarayanan *et al.*, 2016] Aravind S Lakshminarayanan, Ramnandan Krishnamurthy, Peeyush Kumar, and Balaraman Ravindran. Option discovery in hierarchical reinforcement learning using spatio-temporal clustering. *arXiv preprint arXiv :1605.05359*, 2016.

[Lehman and Stanley, 2008] Joel Lehman and Kenneth O Stanley. Exploiting open-endedness to solve problems through the search for novelty. In *ALIFE*, pages 329–336, 2008.

[Leibo *et al.*, 2017] Joel Z Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 464–473. International Foundation for Autonomous Agents and Multiagent Systems, 2017.

[Lesort *et al.*, 2018] Timothée Lesort, Natalia Díaz-Rodríguez, Jean-Franois Goudou, and David Filliat. State representation learning for control : An overview. *Neural Networks*, 2018.

[Levy *et al.*, 2019] Andrew Levy, Robert Platt, and Kate Saenko. Hierarchical reinforcement learning with hindsight. In *International Conference on Learning Representations*, 2019.

[Lillicrap *et al.*, 2015] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv :1509.02971*, 2015.

[Little and Sommer, 2013] Daniel Ying-Jeh Little and Friedrich Tobias Sommer. Learning and exploration in action-perception loops. *Frontiers in neural circuits*, 7 :37, 2013.

[Liu *et al.*, 2015] Chunming Liu, Xin Xu, and Dewen Hu. Multiobjective reinforcement learning : A comprehensive overview. *IEEE Transactions on Systems, Man, and Cybernetics : Systems*, 45(3) :385–398, 2015.

[Lopes *et al.*, 2012] Manuel Lopes, Tobias Lang, Marc Toussaint, and Pierre-Yves Oudeyer. Exploration in model-based reinforcement learning by empirically estimating learning progress. In *Advances in Neural Information Processing Systems*, pages 206–214, 2012.

[Machado *et al.*, 2017a] Marios C Machado, Marc G Bellemare, and Michael Bowling. A laplacian framework for option discovery in reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2295–2304. JMLR. org, 2017.

[Machado *et al.*, 2017b] Marlos C Machado, Clemens Rosenbaum, Xiaoxiao Guo, Miao Liu, Gerald Tesauro, and Murray Campbell. Eigenoption discovery through

the deep successor representation. *arXiv preprint arXiv :1710.11089*, 2017.

[Machado *et al.*, 2018] Marlos C Machado, Marc G Bellemare, and Michael Bowling. Count-based exploration with the successor representation. *arXiv preprint arXiv :1807.11622*, 2018.

[Marsland *et al.*, 2002] Stephen Marsland, Jonathan Shapiro, and Ulrich Nehmzow. A self-organising network that grows when required. *Neural networks*, 15(8-9) :1041–1058, 2002.

[Martin *et al.*, 2017] Jarryd Martin, Suraj Narayanan Sasikumar, Tom Everitt, and Marcus Hutter. Count-based exploration in feature space for reinforcement learning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 2471–2478, 2017.

[McCloskey and Cohen, 1989] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks : The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989.

[McGovern and Barto, 2001] Amy McGovern and Andrew G Barto. Automatic discovery of subgoals in reinforcement learning using diverse density. 2001.

[Menache *et al.*, 2002] Ishai Menache, Shie Mannor, and Nahum Shimkin. Q-cut—dynamic discovery of subgoals in reinforcement learning. In *European Conference on Machine Learning*, pages 295–306. Springer, 2002.

[Mnih *et al.*, 2015] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540) :529, 2015.

[Mnih *et al.*, 2016] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937, 2016.

[Moerland *et al.*, 2016] Thomas M Moerland, Joost Broekens, and Catholijn M Jonker. Fear and hope emerge from anticipation in model-based reinforcement learning. In *IJCAI*, pages 848–854, 2016.

[Moerland *et al.*, 2018] Thomas M Moerland, Joost Broekens, and Catholijn M Jonker. Emotion in reinforcement learning agents and robots : a survey. *Machine Learning*, 107(2) :443–480, 2018.

[Mohamed and Rezende, 2015] Shakir Mohamed and Danilo Jimenez Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. In *Advances in neural information processing systems*, pages 2125–2133, 2015.

[Montúfar *et al.*, 2016] Guido Montúfar, Keyan Ghazi-Zahedi, and Nihat Ay. Information theoretically aided reinforcement learning for embodied agents. *arXiv preprint arXiv :1605.09735*, 2016.

[Nachum *et al.*, 2018] Ofir Nachum, Shixiang (Shane) Gu, Honglak Lee, and Sergey Levine. Data-efficient hierarchical reinforcement learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 3303–3313. 2018.

[Nachum *et al.*, 2019] Ofir Nachum, Shixiang Gu, Honglak Lee, and Sergey Levine. Near-optimal representation learning for hierarchical reinforcement learning. In *International Conference on Learning Representations*, 2019.

[Nair *et al.*, 2018] Ashvin V Nair, Vitchyr Pong, Murtaza Dalal, Shikhar Bahl, Steven Lin, and Sergey Levine. Visual reinforcement learning with imagined goals. In *Advances in Neural Information Processing Systems*, pages 9209–9220, 2018.

[Ng *et al.*, 1999] Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations : Theory and application to reward shaping. In *ICML*, volume 99, pages 278–287, 1999.

[Oh and Cavallaro,] Changjae Oh and Andrea Cavallaro. Learning action representations for self-supervised visual exploration.

[Oh *et al.*, 2015] Junhyuk Oh, Xiaoxiao Guo, Honglak Lee, Richard L Lewis, and Satinder Singh. Action-conditional video prediction using deep networks in atari games. In *Advances in neural information processing systems*, pages 2863–2871, 2015.

[Ostrovski *et al.*, 2017] Georg Ostrovski, Marc G Bellemare, Aaron van den Oord, and Rémi Munos. Count-based exploration with neural density models. *arXiv preprint arXiv :1703.01310*, 2017.

[Oudeyer and Kaplan, 2008] Pierre-Yves Oudeyer and Frederic Kaplan. How can we define intrinsic motivation ? In *Proceedings of the 8th International Conference on Epigenetic Robotics : Modeling Cognitive Development in Robotic Systems, Lund University Cognitive Studies, Lund : LUCS, Brighton*. Lund University Cognitive Studies, Lund : LUCS, Brighton, 2008.

[Oudeyer and Kaplan, 2009] Pierre-Yves Oudeyer and Frederic Kaplan. What is intrinsic motivation ? a typology of computational approaches. *Frontiers in neurorobotics*, 1 :6, 2009.

[Oudeyer and Smith, 2016] Pierre-Yves Oudeyer and Linda B Smith. How evolution may work through curiosity-driven developmental process. *Topics in Cognitive Science*, 8(2) :492–502, 2016.

[Parisi *et al.*, 2019] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter.

Continual lifelong learning with neural networks : A review. *Neural Networks*, 2019.

[Pathak *et al.*, 2017] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning (ICML)*, volume 2017, 2017.

[Perolat *et al.*, 2017] Julien Perolat, Joel Z Leibo, Vinicius Zambaldi, Charles Beattie, Karl Tuyls, and Thore Graepel. A multi-agent reinforcement learning model of common-pool resource appropriation. In *Advances in Neural Information Processing Systems*, pages 3643–3652, 2017.

[Piaget and Cook, 1952] Jean Piaget and Margaret Cook. *The origins of intelligence in children*, volume 8. International Universities Press New York, 1952.

[Plappert *et al.*, 2017] Matthias Plappert, Rein Houthooft, Prafulla Dhariwal, Szymon Sidor, Richard Y Chen, Xi Chen, Tamim Asfour, Pieter Abbeel, and Marcin Andrychowicz. Parameter space noise for exploration. *arXiv preprint arXiv :1706.01905*, 2017.

[Rafati and Noelle, 2019] Jacob Rafati and David C Noelle. Unsupervised methods for subgoal discovery during intrinsic motivation in model-free hierarchical reinforcement learning. 2019.

[Riedmiller *et al.*, 2018] Martin A. Riedmiller, Roland Hafner, Thomas Lampe, Michael Neunert, Jonas Degrave, Tom Van de Wiele, Vlad Mnih, Nicolas Heess, and Jost Tobias Springenberg. Learning by playing solving sparse reward tasks from scratch. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 4341–4350, 2018.

[Riemer *et al.*, 2018] Matthew Riemer, Miao Liu, and Gerald Tesauro. Learning abstract options. In *Advances in Neural Information Processing Systems*, pages 10445–10455, 2018.

[Rückstiess *et al.*, 2010] Thomas Rückstiess, Frank Sehnke, Tom Schaul, Daan Wierstra, Yi Sun, and Jürgen Schmidhuber. Exploring parameter space in reinforcement learning. *Paladyn, Journal of Behavioral Robotics*, 1(1) :14–24, 2010.

[Ryan and Deci, 2000] Richard M Ryan and Edward L Deci. Intrinsic and extrinsic motivations : Classic definitions and new directions. *Contemporary educational psychology*, 25(1) :54–67, 2000.

[Salge *et al.*, 2014a] Christoph Salge, Cornelius Glackin, and Daniel Polani. Changing the environment based on empowerment as intrinsic motivation. *Entropy*, 16(5) :2789–2819, 2014.

[Salge *et al.*, 2014b] Christoph Salge, Cornelius Glackin, and Daniel Polani. Empowerment–an introduction. In *Guided Self-Organization : Inception*, pages 67–114. Springer, 2014.

[Savinov *et al.*, 2018] Nikolay Savinov, Anton Raichuk, Raphaël Marinier, Damien Vincent, Marc Pollefeys, Timothy Lillicrap, and Sylvain Gelly. Episodic curiosity through reachability. *arXiv preprint arXiv :1810.02274*, 2018.

[Schaul *et al.*, 2015] Tom Schaul, Daniel Horgan, Karol Gregor, and David Silver. Universal value function approximators. In *International Conference on Machine Learning*, pages 1312–1320, 2015.

[Schmidhuber, 1991] Jürgen Schmidhuber. Curious model-building control systems. In *[Proceedings] 1991 IEEE International Joint Conference on Neural Networks*, pages 1458–1463. IEEE, 1991.

[Schmidhuber, 2008] Jürgen Schmidhuber. Driven by compression progress : A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. In *Workshop on Anticipatory Behavior in Adaptive Learning Systems*, pages 48–76. Springer, 2008.

[Schmidhuber, 2010] Jürgen Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, 2(3) :230–247, 2010.

[Schwenker and Palm, 2019] Friedhelm Schwenker and Guenther Palm. Artificial development by reinforcement learning can benefit from multiple motivations. *Frontiers in Robotics and AI*, 6 :6, 2019.

[Sequeira *et al.*, 2011] Pedro Sequeira, Francisco S Melo, Rui Prada, and Ana Paiva. Emerging social awareness : Exploring intrinsic motivation in multiagent learning. In *2011 IEEE International Conference on Development and Learning (ICDL)*, volume 2, pages 1–6. IEEE, 2011.

[Sequeira *et al.*, 2014] Pedro Sequeira, Francisco S Melo, and Ana Paiva. Learning by appraising : an emotion-based approach to intrinsic reward design. *Adaptive Behavior*, 22(5) :330–349, 2014.

[Silver *et al.*, 2017] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676) :354, 2017.

[Singh *et al.*, 2010] Satinder Singh, Richard L Lewis, Andrew G Barto, and Jonathan Sorg. Intrinsically motivated reinforcement learning : An evolutionary perspective. *IEEE Transactions on Autonomous Mental Development*, 2(2) :70–82, 2010.

[Skinner, 1938] B. F. Skinner. The behavior of organisms. In *New York : Appleton*, 1938.

[Stadie *et al.*, 2015] Bradly C Stadie, Sergey Levine, and Pieter Abbeel. Incentivizing exploration in reinforcement learning with deep predictive models. *arXiv preprint arXiv :1507.00814*, 2015.

[Stanton and Clune, 2018] Christopher Stanton and Jeff Clune. Deep curiosity search : Intra-life exploration can improve performance on challenging deep reinforcement learning problems. 2018.

[Still and Precup, 2012] Susanne Still and Doina Precup. An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences*, 131(3) :139–148, 2012.

[Su *et al.*, 2015] Pei-Hao Su, David Vandyke, Milica Gasic, Nikola Mrksic, Tsung-Hsien Wen, and Steve Young. Reward shaping with recurrent neural networks for speeding up on-line policy learning in spoken dialogue systems. *arXiv preprint arXiv :1508.03391*, 2015.

[Sutton and Barto, 1998] Richard S Sutton and Andrew G Barto. *Reinforcement learning : An introduction*, volume 1. MIT press Cambridge, 1998.

[Sutton *et al.*, 1999] Richard S Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps : A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2) :181–211, 1999.

[Tang *et al.*, 2017] Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, OpenAI Xi Chen, Yan Duan, John Schulman, Filip DeTurck, and Pieter Abbeel. # exploration : A study of count-based exploration for deep reinforcement learning. In *Advances in neural information processing systems*, pages 2753–2762, 2017.

[Tishby *et al.*, 2000] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[Todorov *et al.*, 2012] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco : A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012.

[Tomar *et al.*, 2018] Manan Tomar, Rahul Ramesh, and Balaraman Ravindran. Successor options : An option discovery algorithm for reinforcement learning. 2018.

[Van den Oord *et al.*, 2016] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems*, pages 4790–4798, 2016.

[Vezhnevets *et al.*, 2017] Alexander Sasha Vezhnevets, Simon Osindero, Tom Schaul, Nicolas Heess, Max Jaderberg, David Silver, and Koray Kavukcuoglu. Feudal networks for hierarchical reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 3540–3549, 2017.

[Warde-Farley *et al.*, 2018] David Warde-Farley, Tom Van de Wiele, Tejas Kulkarni, Catalin Ionescu, Steven Hansen, and Volodymyr Mnih. Unsupervised control through non-parametric discriminative rewards. *arXiv preprint arXiv :1811.11359*, 2018.

[White, 1959] Robert W White. Motivation reconsidered : The concept of competence. *Psychological review*, 66(5) :297, 1959.

[Williams *et al.*, 2015] Henry Williams, Christopher Lee-Johnson, Will N Browne, and Dale A Carnegie. Emotion inspired adaptive robotic path planning. In *2015 IEEE congress on evolutionary computation (CEC)*, pages 3004–3011. IEEE, 2015.

[Wiskott and Sejnowski, 2002] Laurenz Wiskott and Terrence J Sejnowski. Slow feature analysis : Unsupervised learning of invariances. *Neural computation*, 14(4) :715–770, 2002.

[Yu *et al.*, 2015] Chao Yu, Minjie Zhang, Fenghui Ren, and Guozhen Tan. Emotional multiagent reinforcement learning in spatial social dilemmas. *IEEE transactions on neural networks and learning systems*, 26(12) :3083–3096, 2015.

[Zhang *et al.*, 2019] Jingwei Zhang, Niklas Wetzel, Nicolai Dorka, Joschka Boedecker, and Wolfram Burgard. Scheduled intrinsic drive : A hierarchical take on intrinsically motivated exploration. *arXiv preprint arXiv :1903.07400*, 2019.

[Zhou *et al.*, 2019] Xiaomao Zhou, Tao Bai, Yanbin Gao, and Yuntao Han. Vision-based robot navigation through combining unsupervised learning and hierarchical reinforcement learning. *Sensors*, 19(7) :1576, 2019.