Vers des approches hybrides fondées sur l'émergence et l'apprentissage : prise en compte des véhicules autonomes dans le trafic

Joris Dinneweth

université PARIS-SACLAY

Vers des approches hybrides fondées sur l'émergence et l'apprentissage : prise en compte des véhicules autonomes dans le trafic

Towards hybrid approaches based on emergence and learning : considering autonomous vehicles in traffic

École doctorale n° 580,
Sciences et Technologies de l'Information et de la Communication (STIC)
Spécialité de doctorat : Informatique
Graduate School : Informatique et sciences du numérique.
Référent : Faculté des sciences d'Orsay

Thèse préparée dans l'unité de recherche **SATIE**,
(Université Paris-Saclay, ENS ParisSaclay, CNRS)
sous la direction de **Stéphane ESPIÉ**, Directeur de recherche,
le co-encadrement de **René MANDIAU**, Professeur des Universités (UPHF),
le co-encadrement d'**Abderrahmane BOUBEZOUL**, Chargé de recherche, HDR

Thèse soutenue à Paris-Saclay, le 09 décembre 2024, par

Joris DINNEWETH

Composition du jury

Membres du jury avec voix délibérative

Olivier SIMONIN Professeur des Universités, Univ. de Lyon	Président
Marie-Pierre GLEIZES Professeure des Universités, Univ. Paul Sabatier de Toulouse	Rapportrice & Examinatrice
Abderrafiaa KOUKAM Professeur des Universités, Univ. de Technologie de Belfort-Montbéliard	Rapporteur & Examinateur
Zahia GUESSOUM Maîtresse de conférences, HDR, Univ. de Reims Champagne-Ardenne	Examinatrice

THESE DE DOCTORAT
NNT : 2024UPASG099

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Vers des approches hybrides fondées sur l'émergence et l'apprentissage : Prise en compte des véhicules autonomes dans le trafic

Mots clés : Apprentissage par renforcement multiagent, simulation de trafic, comportement humain, véhicule autonome

Résumé : Selon l'Organisation mondiale de la santé, les accidents de la route causent près de 1,2 million de décès et 40 millions de blessés chaque année. Dans les pays riches, des normes de sécurité permettent de prévenir une grande partie des accidents. Les accidents restants trouvent leur cause dans le comportement humain. Ainsi, certains envisagent d'automatiser le trafic, c'est-à-dire de substituer aux humains la conduite de leurs véhicules. Cependant, l'automatisation du trafic routier peut difficilement s'effectuer du jour au lendemain. Ainsi, robots de conduite (RC) et conducteurs humains pourraient cohabiter dans un trafic mixte.

Notre thèse se concentre sur les problèmes sécuritaires qui pourraient émerger en raison des différences comportementales entre les RC et les conducteurs humains. Les RC sont conçus pour respecter les normes formelles, celles du Code de la route. En revanche, les conducteurs humains sont opportunistes et n'hésitent pas à enfreindre les normes formelles et à en adopter de nouvelles, informelles. L'apparition dans le trafic de nouveaux comportements risque de le rendre plus hétérogène et de favoriser la survenue d'accidents ayant pour cause une mauvaise interprétation de ces nouveaux comportements.

Nous pensons que minimiser cette hétérogénéité comportementale permettrait de diminuer les risques susmentionnés. Ainsi, notre thèse propose un modèle décisionnel de RC dont le comportement se veut proche des pratiques humaines non dangereuses, ceci afin de minimiser l'hétérogénéité entre les comportements des RC et des conducteurs humains et dans le but de favoriser leur acceptation par ces derniers. Pour y parvenir, nous adopterons une approche pluridisciplinaire, inspirée par des études de psychologie de la conduite et mêlant simulation de trafic et apprentissage par renforcement multiagent (MARL).

MARL consiste à apprendre un comportement par essai-erreur guidé par une fonction d'utilité. Grâce à sa capacité de généralisation, notamment grâce aux réseaux de neurones, MARL s'adapte à tout type d'environnement, incluant donc le trafic. Nous l'emploierons pour apprendre à notre modèle décisionnel des comportements robustes face à la diversité des situations que comporte le trafic.

Pour éviter les incidents, les constructeurs de RC pourraient concevoir des comportements relativement homogènes et défensifs plutôt qu'opportunistes. Or, cette approche risque de rendre les RC prévisibles et donc vulnérables face à des comportements opportunistes de conducteurs humains. Les conséquences pourraient alors être néfastes tant pour la fluidité que la sécurité du trafic.

Notre première contribution vise à reproduire un trafic hétérogène, c'est-à-dire où chaque véhicule adopte un comportement unique. Nous partons de l'hypothèse qu'en rendant le comportement des RC hétérogène, leur prévisibilité sera réduite et les conducteurs humains opportunistes pourront moins anticiper leurs actions. Ce paradigme considère donc l'hétérogénéité comportementale des RC comme une caractéristique cruciale pour la sécurité et pour la fluidité d'un trafic mixte. Dans une phase expérimentale, nous montrerons la capacité de notre modèle à produire des comportements hétérogènes tout en relevant certains défis MARL.

Notre seconde contribution consistera à intégrer des normes informelles aux processus décisionnels de notre modèle décisionnel de RC. Nous nous concentrerons exclusivement sur l'intégration de la notion de valeur d'orientation sociale qui décrit les comportements sociaux des individus tels que l'altruisme ou l'égoïsme. Partant d'un scénario d'insertion sur autoroute, nous évaluerons l'impact de l'orientation sociale sur la fluidité et la sécurité des véhicules qui s'insèrent. Nous montrerons que l'altruisme permet d'améliorer la sécurité, mais que son impact réel dépend essentiellement de la densité du trafic.

Title : Towards hybrid approaches based on emergence and learning : considering autonomous vehicles in traffic

Keywords : Multi-agent reinforcement learning, traffic simulation, human behavior, autonomous vehicle

Abstract : According to the World Health Organization, road accidents cause almost 1.2 million deaths and 40 million injuries each year. In wealthy countries, safety standards prevent a large proportion of accidents. The remaining accidents are caused by human behavior. For this reason, some are planning to automate road traffic, i.e., to replace humans as drivers of their vehicles. However, automating road traffic can hardly be achieved overnight. Thus, driving robots (DRs) and human drivers could cohabit in mixed traffic.

Our thesis focuses on the safety issues that may arise due to behavioral differences between DRs and human drivers. DRs are designed to respect formal norms, those of the Highway Code. Human drivers, on the other hand, are opportunistic, not hesitating to break formal norms and adopt new, informal ones. The emergence of new behaviors in traffic can make it more heterogeneous and encourage accidents caused by misinterpretation of these new behaviors.

We believe that minimizing this behavioral heterogeneity would reduce the above risks. Therefore, our thesis proposes a decision-making model of DR whose behavior is intended to be close to non-hazardous human practices, in order to minimize the heterogeneity between RC and human driver behavior, and with the aim of promoting their acceptance by the latter. To achieve this, we will adopt a multidisciplinary approach, inspired by studies in driving psychology and combining traffic simulation, multi-agent reinforcement learning (MARL).

MARL consists of learning a behavior by trial and error guided by a utility function. Thanks to its ability to generalize, especially via neural networks, MARL can be adapted to any environment, including traffic. We will use it to teach our decision model robust behavior in the face of the diversity of traffic situations.

To avoid incidents, DR manufacturers could design relatively homogeneous and defensive behaviors rather than opportunistic ones. However, this approach risks making DRs predictable and, therefore, vulnerable to opportunistic behavior by human drivers. The consequences could then be detrimental to both traffic fluidity and safety.

Our first contribution aims at reproducing heterogeneous traffic, i.e., where each vehicle exhibits a unique behavior. We assume that by making the behavior of DRs heterogeneous, their predictability will be reduced and opportunistic human drivers will be less able to anticipate their actions. Therefore, this paradigm considers the behavioral heterogeneity of DRs as a critical feature for the safety and fluidity of mixed traffic. In an experimental phase, we will demonstrate the ability of our model to produce heterogeneous behavior while meeting some of the challenges of MARL.

Our second contribution will be the integration of informal norms into the decision processes of our DR decision model. We will focus exclusively on integrating the notion of social orientation value, which describes individuals' social behaviors such as altruism or selfishness. Starting with a highway merging scenario, we will evaluate the impact of social orientation on the fluidity and safety of merging vehicles. We will show that altruism can improve safety, but that its actual impact is highly dependent on traffic density.

«Ne fais pas attention à moi.
Je viens d'une autre planète.
Je vois toujours des horizons
où tu dessines des frontières.»

Frida Kahlo

Remerciements

Je remercie l'université Gustave Eiffel pour avoir financé mes travaux de thèse.

Je tiens d'abord à remercier Marie-Pierre Gleizes et Abderrafiaa Koukam qui m'ont fait l'honneur de rapporter mon manuscrit. Je remercie également Olivier Simonin et Zahia Guessoum d'avoir accepté d'examiner mes travaux de thèse.

Je remercie mes encadrants Stéphane Espié, René Mandiau et Abderrahmane Boubezoul pour m'avoir accordé une pleine liberté durant ces trois années. Leur grande disponibilité et leurs conseils ont été décisifs. Je remercie Stéphane pour toutes les discussions enrichissantes que nous avons eues. Je remercie René pour m'avoir aiguillé chaque fois qu'il le fallait. Je remercie Abderrahmane pour son optimisme contagieux et ses encouragements permanents.

Plus largement, je remercie l'ensemble des personnes du SATIE pour leur accueil et leur bonne humeur. Particulièrement, je remercie Émi de m'avoir montré les bons coins à champignons de la région ! Je remercie aussi tous les doctorants avec qui j'ai passé d'agréables moments.

Je remercie mes proches pour tout le sens qu'ils donnent au mot *famille*. Je remercie infiniment mes parents de m'avoir offert un trousseau dont les clefs m'ont permis d'ouvrir toutes les portes du bonheur. Leurs encouragements et leurs conseils de *toujours faire ce qui me plaît* ont pleinement porté leurs fruits. Je remercie mes grands-parents tout simplement pour ce qu'ils sont. Plus intimement, je remercie ma grand-mère qui n'a cessé de m'encourager, voyant ma poursuite d'études comme une revanche sur la vie, elle, qui, malgré son amour pour l'école, s'est vue contrainte de travailler à l'usine dès l'âge de quatorze ans. Je remercie la personne avec qui je partage la plus grande complicité : Maxime, mon frère. Nos aventures montagnardes à tutoyer les sommets, parfois pimentées de péripéties, ont été revitalisantes. Enfin, comme je devine tout l'intérêt qu'il portera à la lecture de cette thèse, je remercie mon chien Filou, même si je sais pertinemment qu'il aurait préféré un os à quelconques remerciements. Ses nombreux *ouaf* m'ont beaucoup influencé (et réveillé).

Table des matières

Liste des figures

Liste des Algorithmes

Liste des tableaux

Nomenclature

Notation mathématique

$\mathbb{E}$	Espérance
$\mathcal{N}$	Loi Gaussienne
$\mathbb{N}$	Entiers naturels
$\mathbb{Z}$	Entiers naturels positifs
$\mathbb{N}^+$	Entiers naturels non-nuls
$\mathbb{P}[X]$	Probabilité
$\mathbb{R}$	Réels

Apprentissage par renforcement

α_μ, α_Q	Taux d'apprentissage de l'acteur et du critique
$\mathcal{I}$	Ensemble d'agents
$\mathbb{L}$	Algorithme
$\mathcal{B}, \lvert\mathcal{B}\rvert$	Lot et taille du lot
β	Croyance
d	Donnée
$\mathcal{D}, \lvert\mathcal{D}\rvert$	Jeu de données et sa taille
$\mathcal{L}$	Fonction de perte
α	Taux d'apprentissage
Ω	Fonction d'observation
$\mathcal{D}_p, \lvert\mathcal{D}_p\rvert$	Jeu de données prioritisé et sa taille

u	Retour ou utilité
$\mathcal{R}$	Fonction de récompense
μ_0	Distribution initiale des états
$\mathcal{T}$	Fonction de transition
$a \in \mathcal{A}$	Action et ensemble d'actions
$g \in \mathcal{G}$	Objectif et ensemble d'objectifs
$o \in \mathcal{O}$	Observation et ensemble d'observation
$Q(s, a)$	Fonction de qualité d'action
r	Récompense
s'	État suivant
$s \in \mathcal{S}$	État et ensemble d'états
s_0	État initial
$V(s)$	Fonction de valeur d'état

Spécifique au trafic

$\ddot{x}, \ddot{y}$	Accélérations longitudinale et latérale
$\hat{c}^t$	Contrainte anticipée
$\flat$	Freinage
$t_i^\flat$	Durée de freinage de i
c	Contrainte
k	Densité
$\widehat{\ell}$	Voie cible
$\widehat{x}$	Vitesse désirée
q	Flux
i_{-1}	Véhicule suivant i
$i_{-1}^{\widehat{\ell}_i}$	Véhicule suivant i sur la voie cible
Ψ	Interaction
t_Ψ	Temps d'interaction
$\tilde{j}$	Jerk
ℓ	Voie

$\ell_i \rightarrow \widehat{\ell}_i$	Changement de voie du véhicule i
$\overline{\ell}$	Fin de voie
$\dot{x}_\ell$	Vitesse moyenne de voie
$\dot{x}_\ell^{\max}$	Limite de vitesse de la voie
w_ℓ	Largeur de voie
t_Ψ^+	Interaction longue
$t_i^{\flat,\max}$	Durée maximale de freinage de i
x, y	Positions longitudinale et latérale
i_{+1}	Véhicule précédant i
$i_{+1}^{\widehat{\ell}_i}$	Véhicule précédant i sur la voie cible
p	Pression
Tr	Temps de réaction
RP	Risque perçu
$\dot{x}, \dot{y}$	Vitesses longitudinale et latérale
ϕ	Valeur d'orientation sociale (SVO)
$\hat{\imath}$	Agent cible d'une action empathique
TH	Intertemps
TTC	Temps avant collision
w_i	Longueur du véhicule i

Introduction

Selon l'Organisation mondiale de la santé [1], les accidents de la route causent près de 1,2 million de décès et 40 millions de blessés chaque année. La majorité de ces accidents survient dans les pays pauvres ou en voie de développement. Dans les pays riches, pendant longtemps, les politiques de sécurité routière ont consisté à réglementer les conduites individuelles, à assister les conducteurs dans leurs tâches par des moyens technologiques divers et à équiper les véhicules de systèmes de sécurité comme l'airbag. Avec ces mesures, le nombre d'accidents de la route a dramatiquement chuté et la plupart des accidents restants trouvent leur cause dans le comportement humain. Poursuivant cet objectif d'un trafic dépourvu d'accident, certaines politiques envisagent d'automatiser le trafic, c'est-à-dire de substituer aux humains la conduite de leurs véhicules. Suivant cette logique, puisque l'automatisation du trafic routier peut difficilement s'effectuer du jour au lendemain, véhicules à conduite robotisés – souvent appelés véhicules autonomes – pilotés par des robots de conduite (RC) et véhicules conduits par des humains devront cohabiter au sein d'un trafic mixte.

Si un trafic automatisé offre des garanties sécuritaires plus fortes, il ne faut pas pour autant négliger que, comme tout changement, il provoquera une ou plusieurs crises. Certaines, d'ordre social, ont déjà été constatées avec l'introduction des taxis robotisés faisant concurrence aux conducteurs de taxis humains [2]. Certaines crises peuvent aboutir à une régression, autrement dit à un retour en arrière, à l'adoption de comportements antérieurs, néfastes pour le trafic, sa fluidité et sa sécurité.

Notre thèse se concentre sur une crise d'ordre sécuritaire qui pourrait émerger

1. https://www.who.int/fr/news-room/fact-sheets/detail/road-traffic-injuries

2. https://www.leparisien.fr/video/video-a-san-francisco-des-habitants-sopposent-au-deploiement-de-taxis-sans-chauffeur-avec-des-cones-07-07-2023-XWBP6ATO4JBVHC7BN225IEWRDM.php

en raison des différences comportementales entre les RC et les conducteurs humains (chapitre 1). Les RC sont conçus selon une logique de rationalité absolue qui cherche à optimiser leur comportement tout en respectant scrupuleusement les normes formelles telles qu'édictées par le Code de la route. En revanche, les conducteurs humains agissent selon une logique de rationalité limitée cherchant les opportunités à court terme, quitte à enfreindre les normes formelles et à en inventer de nouvelles, informelles. L'apparition dans le trafic de nouveaux comportements risque de le rendre plus hétérogène et de favoriser la survenue d'accidents ayant pour cause une mauvaise interprétation de ces nouveaux comportements.

Nous pensons que minimiser cette hétérogénéité comportementale permettrait de diminuer les risques susmentionnés. Ainsi, notre thèse propose un modèle décisionnel de RC dont le comportement se veut proche des pratiques humaines non dangereuses, ceci afin de minimiser l'hétérogénéité entre les comportements des RC et des conducteurs humains et dans le but de favoriser leur acceptation par ces derniers. Pour y parvenir, nous adopterons une approche pluridisciplinaire, inspirée par des études de psychologie de la conduite, et mêlant simulation de trafic, système multiagent et apprentissage par renforcement.

L'apprentissage par renforcement appartient au domaine de l'apprentissage machine, son but consiste à apprendre un comportement par essai-erreur guidé par une fonction d'utilité (chapitre 2). Contrairement aux autres formes d'apprentissage, le renforcement génère ses propres données, ce qui élimine les contraintes de disponibilité et de qualité des données d'entraînement. Grâce à sa capacité de généralisation, notamment grâce aux réseaux de neurones, l'apprentissage par renforcement s'adapte à tout type d'environnement, incluant donc le trafic. Nous l'emploierons principalement pour apprendre à notre modèle décisionnel des comportements robustes face à la diversité des situations que comporte le trafic.

Pour éviter les incidents, les constructeurs de RC pourraient concevoir des comportements relativement homogènes et défensifs plutôt qu'opportunistes. Or, cette approche risque de rendre les RC prévisibles et donc vulnérables face à des comportements opportunistes de conducteurs humains. Les conséquences pourraient alors être néfastes tant pour la fluidité que la sécurité du trafic. En d'autres termes, l'introduction de RC serait contre-productive.

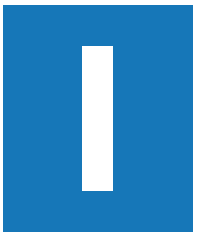
Notre première contribution (chapitre 3) vise à reproduire un trafic hétérogène, c'est-à-dire où chaque véhicule adopte un comportement unique. Nous partons de l'hypothèse qu'en rendant le comportement des RC hétérogène, leur prévisibilité sera réduite et les conducteurs humains opportunistes pourront moins anticiper leurs actions. Ce paradigme considère donc l'hétérogénéité comportementale des RC comme une caractéristique cruciale pour la sécurité et pour la fluidité d'un trafic mixte. Dans une phase expérimentale, nous montrerons la capacité de notre modèle à produire des comportements hétérogènes tout en relevant certains défis de l'apprentissage par renforcement.

Notre seconde contribution (chapitre 4) consistera à intégrer des normes

informelles aux processus décisionnels de notre modèle décisionnel de RC. Nous nous concentrerons exclusivement sur l'intégration de la notion de valeur d'orientation sociale qui décrit les comportements sociaux des individus tels que l'altruisme ou l'égoïsme. Partant d'un scénario d'insertion sur autoroute, nous évaluerons l'impact de l'orientation sociale sur la fluidité et la sécurité des véhicules qui s'insèrent. Nous montrerons que l'altruisme permet d'améliorer la sécurité, mais que son impact réel dépend essentiellement de la densité du trafic.

Ce manuscrit s'achèvera par un bilan général et un aperçu des perspectives de nos travaux.

Partie



Contexte

CHAPITRE

Trafic, simulation et comportements des conducteurs

1

Sommaire du chapitre

Ce premier chapitre contextualise nos travaux de thèse. Nous y décrivons le domaine de la simulation de trafic, la manière dont les comportements humains influencent le trafic et les éventuelles modifications comportementales engendrées par la cohabitation avec les robots de conduite (RC). Considérant les défis et les verrous technologiques, nous concluons ce chapitre en spécifiant la problématique de notre thèse.

Dans la section 1.1, nous commencerons par définir la simulation de trafic routier, et présenterons ses applications. Nous distinguerons trois approches de simulation du trafic routier : (1) la simulation par modèles mathématiques (1.1.1), (2) la simulation par base de règles décisionnelles (1.1.2) et (3) par bases de règles motivationnelles (1.1.3).

Dans la section 1.2 nous analyserons les résultats des études comportementales de conducteurs humains menées par des psychologues. Nous nous focaliserons sur la manière dont les humains gèrent leurs interactions dans le trafic et nous remarquerons que certaines motivations sont partagées par l'ensemble des conducteurs (1.2.1). Puis, nous examinerons comment l'hétérogénéité des conducteurs conditionne les interactions au sein du trafic (1.2.2).

Enfin, la section 1.3 traitera des robots de conduite (RC). Nous y décrirons les différents niveaux d'automatisation du véhicule ainsi que les algorithmes qui ont été proposés (1.3.1). Ensuite, nous listerons les défis qui restent à relever pour parvenir à un trafic mixte, *i.e.* une cohabitation entre conducteurs humains et RC (1.3.2).

Nous conclurons ce chapitre en précisant les problématiques sur lesquelles nous avons souhaité nous concentrer et la manière dont nous souhaitons y répondre.

1.1 Simulation de trafic

Le domaine de la simulation de trafic vise à reproduire et prévoir les phénomènes du trafic à différentes échelles. Commençons par définir quelques notions essentielles, à commencer par celles de *simulation*.

Définition 1: Simulation

Une simulation est un modèle dynamique que l'on perturbe en fonctions d'objectifs [Treuil *et al.*, 2008]. Toute simulation s'effectue dans un *simulateur*, *i.e.*, un programme informatique capable d'interpréter un modèle dynamique et d'y appliquer des perturbations désirées. La simulation vise à faire évoluer les entrées d'un modèle dynamique pour en recueillir les sorties. Les *entrées* d'un modèle dynamique sont des paramètres définissant ce qui peut être perturbé. Les *sorties* d'un modèle dynamique sont les paramètres exprimant ce que l'on cherche à mesurer en fonction des perturbations.

La sortie d'une simulation est donc déterminée par ses entrées. Les perturbations appliquées au modèle dynamique peuvent être déterministes ou stochastiques. Dans le domaine de la simulation de trafic, nous distinguons trois types d'entrées : analytiques (1.1.1), règles décisionnelles (1.1.2) et règles motivationnelles (1.1.3).

1.1.1 Modèles analytiques

Les modèles analytiques reposent sur des entrées purement mathématiques, le plus souvent des données collectées par des capteurs. Ces modèles visent donc à reproduire les phénomènes observés à différentes échelles : macroscopique, mésoscopique ou microscopique.

Les simulations macroscopiques modélisent le trafic sous la forme de flux de véhicules, souvent par des équations différentielles empruntées à la dynamique des fluides. L'échelle macroscopique permet de simuler un trafic à l'échelle d'une ville, voire d'un pays, puisque les véhicules n'y sont pas modélisés, seul le trafic qu'ils produisent l'est. Les approches macroscopiques sont divisées entre les modèles de premier ordre et ceux de second ordre [Hoogendoorn et Knoop, 2013]. Le trafic y est décrit selon trois variables : la *densité* k, la *vitesse* $\dot{x}$, et le *débit* q.

Les *modèles de premier ordre* ont été introduits par Lighthill et Whitham [1955] et supposent que le trafic suit une relation $q_\ell^t = k_\ell^t \cdot \dot{x}_\ell^t$ avec ℓ un segment de route et t le temps. Le flux est supposé homogène selon ℓ et se décrit par une loi de conservation :

$$\partial_t k + \partial_x (k \cdot \dot{x}) = 0 \tag{1.1}$$

Plus tard, Payne [1973] introduit les *modèles de second ordre* afin de modéliser la non-linéarité de la relation entre la densité et la vitesse moyenne du trafic. Son modèle est un système composé de deux équations dont la première est donnée par l'équation 1.1 et la seconde :

$$\partial_t \dot{x} + \dot{x} \cdot \partial_t \dot{x} + \frac{\partial_x p(k)}{k} = \frac{1}{\tau} \cdot (\widehat{\dot{x}}_k - \dot{x}) \tag{1.2}$$

où $p(k)$ est la pression, $\widehat{\dot{x}}_k$ la vitesse désirée et τ le temps de relaxation. Ce modèle a permis notamment de mieux modéliser le phénomène d'*accordéon* (vagues d'arrêts-départs) dans le trafic.

La granularité des simulations macroscopiques empêche l'étude des interactions intervéhiculaires. C'est pourquoi, avec l'augmentation de la puissance de calcul des ordinateurs, des modèles microscopiques ont vu le jour [Kubera *et al.*, 2010]. Les modèles d'ordre microscopiques se divisent en deux catégories selon qu'ils reproduisent des *lois de poursuite* ou des *changements de voie* [Moridpour *et al.*, 2010]. Les lois de poursuite décrivent les manœuvres longitudinales, tandis que les changements de voie décrivent les manœuvres latérales. Ces deux catégories sont donc complémentaires.

Les modèles microscopiques analytiques décrivent les comportements des véhicules par des équations mathématiques. L'un des premiers modèles microscopiques, le *modèle de distance de sécurité*, fut proposé par Pipes [1953]. Son modèle décrit la distance de sécurité qu'un véhicule i, avec une vitesse $\dot{x}_i$, est supposée respecter lorsqu'elle suit un véhicule j. Cette interdistance $\Delta_x(i,j)$ correspond à

celle nécessaire pour répondre à un freinage quasi instantané du véhicule précédent j sans prendre de risque :

$$\Delta_x(i,j) = w_i \cdot \left(1 + \frac{\dot{x}_i}{16{,}1}\right) \tag{1.3}$$

Dans cette équation, w_i représente la longueur du véhicule, tandis que 16,1 km/h correspond approximativement à la conversion de 10 *miles*.

Ce modèle omet cependant le temps de réaction et la décélération maximale supportée par le véhicule, deux paramètres décisifs pour un freinage rapide. Ce manque fut comblé par les modèles stimulus-réponse qui supposent qu'un véhicule réagit à celui qui le précède comme s'il s'agissait d'un stimulus. Ainsi, lorsque le véhicule précédent accélère ou décélère, le véhicule suiveur l'imite avec un certain temps de réaction.

Le *modèle stimulus-réponse* le plus populaire a été proposé par Gazis *et al.* [1961] et est communément appelé *modèle GHR*. Ce modèle décrit la sensibilité d'un conducteur, soit sa propension à accélérer $\ddot{x}_i$, comme proportionnelle à l'interdistance $\Delta_x(i,j)$ et au différentiel de vitesse $\Delta_{\dot{x}}(i,j)$ auxquels on ajoute un *temps de réaction* Tr :

$$\ddot{x}_i(Tr) = \dot{x}_i(Tr) \cdot \frac{\Delta_{\dot{x}}(i,j)}{\Delta_x(i,j)} \tag{1.4}$$

Plus tard, des modèles psychophysiques viennent enrichir la simulation de comportements en s'inspirant des découvertes menées par des psychologues [Schulze et Fliess, 1997]. Leur principale amélioration vient du fait que ces modèles considèrent les conducteurs comme insensibles aux stimuli mineurs tels que les variations légères d'interdistance ou lorsque l'interdistance est très grande. Le *modèle psychophysique* de Wiedemann (1974) propose cinq régimes de conduite, chacun régi par une équation décrivant l'accélération à adopter :

- *Conduite libre* — le véhicule circule librement en l'absence de contraintes.
- *Poursuite* — maintient d'une vitesse identique à celle du véhicule précédent.
- *Approche* — le véhicule se rapproche du véhicule précédent.
- *Freinage* — le différentiel de vitesse nécessite une décélération.
- *Collision* — situations accidentogènes.

Les limites de ces régimes sont définies par des seuils de perception calculés selon l'interdistance et le différentiel de vitesse avec le véhicule précédent (figure 1.1).

Malgré leur simplicité, ces modèles génèrent souvent des accidents que d'autres modèles tels que l'*Optimal Velocity Model (OVM)* se proposent d'éliminer [Bando *et al.*, 1995]. Le modèle OVM considère l'inconstance comportementale des conducteurs comme la cause primaire de la congestion du trafic. Un véhicule réagit aux stimuli selon sa vélocité, et ses préoccupations sécuritaires guident son désir d'accélération :

$$\ddot{x}_i = \ddot{x}_i^{\max}\left[\widehat{\dot{x}}_i(\Delta_x(i,j)) - \dot{x}\right] \tag{1.5}$$

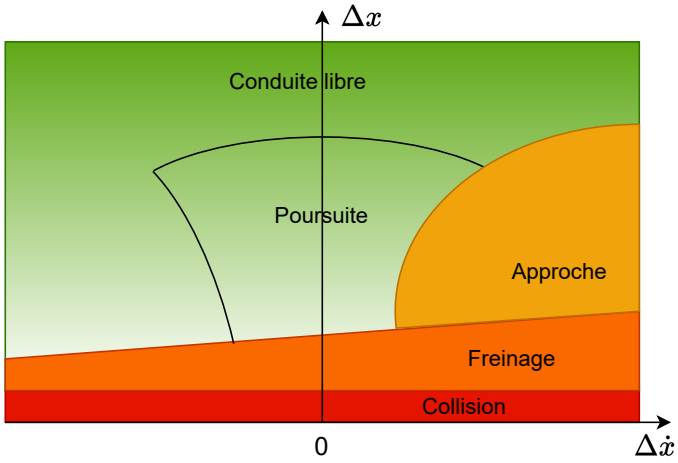


Figure 1.1 – Régimes et seuils du modèle psychophysique de Wiedemann (figure inspirée de [Mitroi *et al.*, 2016])

où $\widehat{\dot{x}}_i(\Delta_x(i,j))$ estime la *vitesse préférentielle* de i selon son interdistance avec j. Ce modèle empêche les véhicules de provoquer des accidents puisqu'ils ne peuvent dépasser un véhicule plus lent, mais produisent en revanche des embouteillages [Ahmed *et al.*, 2021].

Aujourd'hui, la plupart des simulations fondées sur les approches analytiques adoptent le couple de modèles IDM-MOBIL [Kesting *et al.*, 2007]. Le modèle *Intelligent-Driver Model (IDM)* simule la conduite en file, tandis que le modèle *Minimizing Overall Braking Induced by Lane change (MOBIL)* reproduit les changements de voies.

Le modèle IDM décrit l'accélération d'un conducteur selon la vitesse désirée et son interdistance. Les véhicules convergent alors vers leur vitesse préférentielle $\widehat{\dot{x}}_i$ tout en maintenant une distance de sécurité raisonnable [Treiber *et al.*, 2000] :

$$\ddot{x}_i = \ddot{x}_i^{\max}\left[1 - \left(\frac{\dot{x}}{\widehat{\dot{x}}_i}\right)^4 - \left(\frac{s^*(\dot{x}, \Delta_x(i,j))}{\Delta_x(i,j)}\right)^2\right] \tag{1.6}$$

où

$$s^*(\dot{x}, \Delta_x(i,j)) = \Delta_x(i,j)_{min} + \dot{x}_i TH(i,j) + \frac{\dot{x}_i \Delta_{\dot{x}}(i,j)}{2\sqrt{\ddot{x}_i^{\max}\ddot{x}_i^*}} \tag{1.7}$$

avec $\ddot{x}_i^*$ une valeur de décélération confortable. En outre, ce modèle a reçu de nombreuses extensions qui ont affiné les comportements simulés [Kesting *et al.*, 2010].

Le modèle MOBIL, fréquemment couplé au modèle IDM, simule le changement de voie en estimant les risques et les bénéfices associés. Le bénéfice est calculé par la différence d'accélération avant et après (estimé) la manœuvre, tandis que le risque est approximé par la décélération requise pour freiner de manière sécurisée. Un facteur de *politesse* encourage l'altruisme ou l'égoïsme et vient pondérer le

calcul du bénéfice pour simuler l'hétérogénéité naturelle des comportements au sein du trafic. Finalement, le changement de voie n'est effectué que si le bénéfice excède un certain seuil.

Notons qu'entre les modèles macroscopiques et microscopiques, il existe un intermédiaire couvert par les modèles *mésoscopiques* [Chao *et al.*, 2020; Bouha *et al.*, 2015]. Plutôt que de modéliser directement les véhicules, les modèles mésoscopiques proposent de considérer des pelotons homogènes de véhicules partageant globalement les mêmes caractéristiques. L'étude des interactions intervéhiculaires est donc supplantée par celle des interactions interpelotons. Nous ne nous attarderons pas plus sur ces modèles.

Malgré une tendance à incorporer des concepts issus de la psychologie, tels que le temps de réaction, la réaction aux stimuli ou encore le facteur de politesse, les modèles analytiques restent peu flexibles et échouent à capturer et à reproduire la complexité et l'hétérogénéité inhérentes aux décisions et comportements humains.

1.1.2 Modèles à base de règles décisionnelles

Moins rigides, les modèles à base de règles décisionnelles peuvent associer un comportement unique à chaque situation et ainsi simuler des scénarios de trafic plus complexes. Dans ces modèles, ce ne sont plus seulement des véhicules qui sont simulés, mais aussi les agents qui les pilotent. Nous parlons donc de simulation multiagent, où un agent étant défini comme une entité autonome.

Certaines simulations à base d'agents voient parfois émerger des phénomènes à l'échelle macroscopique sans que ces derniers aient été volontairement induits [Picard *et al.*, 2009]. Définissons le concept d'émergence.

Définition 2: Émergence

L'émergence est un concept flou qui décrit une *nouvelle* propriété d'un système résultant des interactions des agents le composant. On distinguera l'émergence forte de l'émergence faible [Drogoul et Ferber, 2018]. Il y *émergence forte* lorsque les nouvelles propriétés du système sont *irréductible*, c'est-à-dire, qu'elles ne peuvent être déduites des propriétés des agents par lesquelles elles émergent. Sinon, l'émergence est dite *faible*.

Dans le trafic, l'émergence s'observe lors de la formation d'embouteillages.

Les *automates cellulaires* font partie des premiers modèles à base de règles adaptés pour le trafic. La simulation prend place sur une grille où chaque cellule est soit vide, soit occupée par un agent [Nagel et Schreckenberg, 1992]. À chaque pas de temps, les agents avancent d'un nombre de cellules équivalent à leur vitesse. Chaque agent accélère jusqu'à atteindre sa vitesse désirée tant qu'il existe un nombre suffisant de cellules libres devant lui. À l'inverse, si un agent poursuit de trop près son prédécesseur, il décélère. Avec une certaine probabilité, les agents freinent afin de reproduire le phénomène naturel et involontaire des variations de

vitesse des conducteurs qui atténuent le déterminisme de la simulation. Certains auteurs, comme Vasirani et Ossowski [2012] utilisent les automates cellulaires pour anticiper et prévenir les interblocages dans les intersections en réservant les cellules correspondant à leur trajectoire future.

En dépit d'un coût computationnel maîtrisé, les automates cellulaires semblent bien incapables de reproduire la variété des situations observées du trafic et les agents sont contraints d'évoluer dans un espace discret.

Face à ces limitations, des modèles plus complexes fondés sur des *raisonnements logiques et temporels* ont vu le jour. Dans les modèles à *logique floue*, les variables ne sont plus représentées quantitativement, mais qualitativement [Mitroi *et al.*, 2016]. Par exemple, les distances peuvent s'exprimer par l'un des termes suivants : très proche, proche, éloigné, très éloigné. Ainsi, chaque agent fait sa propre évaluation de la situation : là où un conducteur considérera un véhicule comme *très proche*, un autre pourra le percevoir comme *proche*. Ces différences de perceptions induisent parfois des attentes contraires et *in fine* des comportements hétérogènes.

Malgré leur plasticité, les modèles à base de règles se heurtent à une limite fondamentale : il leur est impossible de décrire de manière exhaustive l'ensemble des situations qui émergent au sein du trafic [Trannois *et al.*, 1998]. Par exemple, pour une intersection en croix, Onieva *et al.* [2015] définissent un ensemble composé de plus d'une dizaine de règles. Produire une base de règles pour chaque intersection devient vite fastidieux et chaque nouvelle intersection qui diffère des précédentes vient remettre en cause ces modèles.

1.1.3 Modèles à base de règles motivationnelles

En parcourant les études de psychologie de la conduite, on constate que loin des visions robotiques proposées par les modèles à base décisionnelles, les conducteurs sont avant tout guidés par leurs motivations. Ces motivations peuvent être de différentes natures, comme atteindre leur vitesse désirée. En simulant – en mettant en entrée du modèle – les motivations des conducteurs plutôt que des règles décisionnelles, il devient possible de reproduire une plus grande diversité de phénomènes du trafic. Contrairement aux règles décisionnelles, les règles motivationnelles sont en nombre plus limité et offrent une plus grande capacité de généralisation sur les situations rencontrées.

Pour parvenir à extraire des règles motivationnelles, une psychologue de la conduite, F. Saad, a mené des entretiens psychologiques avec des conducteurs en condition réelle, durant une tâche de conduite, leur demandant de verbaliser leurs décisions [Espié et Saad, 2000]. Son étude a permis d'identifier un ensemble de *motivations* commun aux décisions des conducteurs. Globalement, elle révèle que les intentions des conducteurs sont guidées par la minimisation de leurs interactions. Une interaction est un échange ou contact entre une conductrice et un autre véhicule ou avec son environnement. Selon les situations, les *interactions* peuvent

prendre la forme d'une co-action, d'une *coopération* dans l'action, mais également d'une *compétition* [Munduteguy, 2001].

Les règles motivationnelles décrivant les comportements des conducteurs se composent ainsi :

- interaction + longue durée + possible suppression → suppression de l'interaction
- interaction + longue durée + impossibilité de suppression à long terme → adaptation à long terme
- interaction + courte durée → adaptation à court terme

Ici, une interaction est soit immédiate, soit anticipée. Chaque règle décrit des circonstances (à gauche du symbole →) qui induisent un comportement. Par exemple, une conductrice peut anticiper une interaction lorsqu'elle roule plus rapidement que le véhicule qui la précède. Une courte durée réfère à une interaction qui dure moins de deux secondes. L'*adaptation* renvoie à une modulation de la vitesse et au choix d'un temps intervéhiculaire (court ou long), tandis que la *suppression* d'une interaction reflète un changement de voie.

Illustrons ces trois règles motivationnelles par des scénarios réalistes :

1. un conducteur approche d'une zone d'insertion sur autoroute. La différence de vitesse entre les deux flux de véhicules – ceux sur l'autoroute et ceux qui s'y insèrent – va vraisemblablement entraîner des ralentissements sur la voie d'autoroute adjacente à la voie d'insertion. Par anticipation d'une interaction de longue durée, le conducteur change de voie puisqu'il en a l'opportunité.
2. un conducteur se trouve dans une situation identique au cas précédent – interaction de longue durée –, mais sans possiblité de changer de voie. Le conducteur module alors sa vitesse pour éviter un conflit avec le véhicule s'insérant sur sa voie.
3. un conducteur roule derrière un véhicule qui ralentit à l'approche d'une zone de sortie d'autoroute et manisfeste son intention de sortir par l'usage de son clignotant. Le conducteur prédit une interaction courte, module sa vitesse et accepte un temps intervéhiculaire court – sans changer de voie, bien qu'en ayant l'opportunité – jusqu'à ce que le véhicule atteigne la voie de sortie mettant fin à l'interaction.

Contrairement aux bases de règles décisionnelles que nous avons explorées auparavant, ces règles motivationnelles englobent l'ensemble des comportements observés et s'appliquent à tout scénario routier. Reprenons par exemple le dernier scénario que nous venons d'évoquer et changeons le contexte autoroutier pour un contexte urbain. Dans ce nouvel exemple, un conducteur s'approche d'un usager effectuant une manœuvre de stationnement. Anticipant que cet usager achèvera sa manœuvre sous peu – interaction de courte durée –, le conducteur module sa vitesse et accepte un temps intervéhiculaire court.

Si l'on peut aisément déterminer les motivations propres au processus d'adaptation, *i.e.* de modulation de vitesse, il n'en est pas de même pour celui de suppression des interactions. En effet, les décisions de changement de voie dépendent de nombreux facteurs qui incluent l'espace routier disponible, les différences de vitesse entre les voies, la stabilité de celles-ci, et de l'anticipation par le conducteur des éléments susmentionnés.

Les travaux de thèse de El Hadouaj [2004] nous éclairent sur le processus de prise de décisions associé au changement de voie. Lorsqu'une conductrice évalue ses opportunités de changement de voie, elle associe une *utilité* à chacune des voies, y compris la sienne. Ce gain dépend de la différence entre la vitesse moyenne d'une voie et sa vitesse désirée, mais également de la *stabilité* du trafic proche comme lointain. Si une voie obtient un gain plus élevé que celle sur laquelle il se trouve, la conductrice y cherche alors un créneau sécurisé pour effectuer sa manœuvre. Si aucun créneau ne semble disponible ou que le gain estimé d'un changement de voie s'avère désavantageux, la conductrice se maintient sur sa voie.

Concernant les intersections, Mandiau *et al.* [2008] ont conçu un mécanisme de coordination fondé sur la *théorie des jeux* pour la gestion des interactions dans les intersections en croix. Les agents évaluent les situations rencontrées et ne s'engagent sur l'intersection que s'ils pensent éviter de provoquer un interblocage. Avec leur approche, la coordination entre deux agents, voire plus, peut être résolue de manière efficace.

D'autres auteurs ont remarqué que certaines *normes formelles* semblaient également sujettes à des évaluations subjectives, du moins dans la culture étudiée. Champion *et al.* [2002] montrent ainsi que certaines pratiques réelles des conducteurs dans les intersections ne peuvent être reproduites en simulation que si l'on considère la priorité comme une notion subjective. Considérant ce dernier point, Doniec *et al.* [2008b] ont modélisé les conflits survenant aux intersections en croix ou ronds-points. L'intersection est vue comme un problème de coordination où chacun souhaite éviter les collisions et les interblocages en anticipant les comportements d'autrui. Leur modèle de *résolution de conflit* se fonde sur un modèle de *raisonnement distribué par contraintes* (figure 1.2). Les décisions des conducteurs dépendent donc de la vitesse et de la position des véhicules, mais également de caractéristiques psychologiques comme l'impatience.

Dans leur modèle, les *normes formelles*, telles que la priorité, ne prévalent pas sur les *normes informelles*, comme l'impatience. Ainsi, un agent prioritaire ralentira certainement, voire cédera la priorité, s'il estime qu'une collision paraît possible avec un véhicule arrivant à grande vitesse. À l'inverse, un agent roulant à vive allure enfreindra le Code de la route s'il juge avoir le temps nécessaire pour passer. L'impatience d'un agent se traduit par une plus grande propension à transgresser les normes formelles, et par conséquent, à s'engager dans l'intersection en dépit de

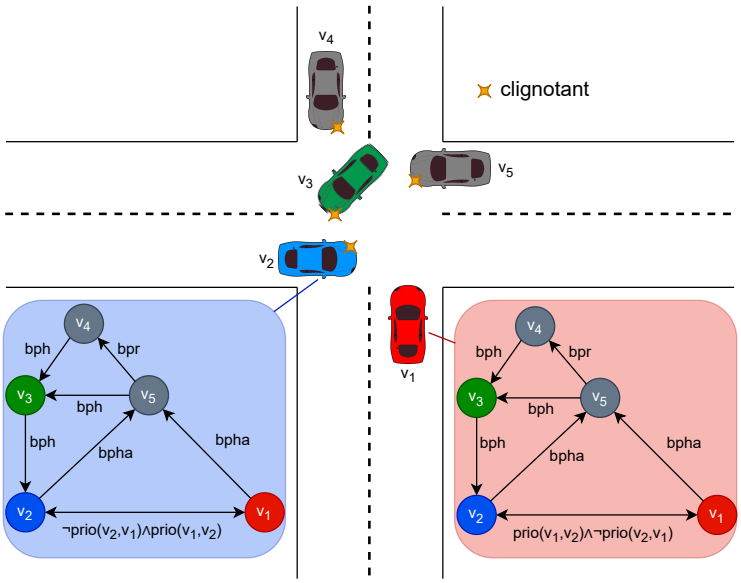


Figure 1.2 – Représentation mentale de deux agents à une intersection en croix (figure inspirée de Doniec *et al.* [2008a])

toute priorité. Un tel scénario se traduit par une logique de prédicats :

$$\begin{aligned}&(\neg \mathit{prioCode}(i,j) \wedge \mathit{prioCode}(j,i)) \wedge \mathit{impatience}(i) \wedge \\ &(\mathit{distC}(i,j) < \mathit{distC}(j,i)) \rightarrow (\mathit{prio}(x,y) \wedge \neg \mathit{prio}(y,x))\end{aligned} \tag{1.8}$$

où $distC$ représente la distance à couvrir par l'agent i avant d'entrer en conflit avec le véhicule j.

En conclusion de cette première section, nous pouvons affirmer que les approches à base de règles motivationnelles conviennent mieux à notre problématique que les approches analytiques ou à base de règles décisionnelles. Les modèles analytiques semblent incapables de capturer les nuances de comportements entre les conducteurs, tandis que les modèles à base de règles décisionnelles souffrent d'un manque d'exhaustivité qui pourrait amener leurs concepteurs à définir toujours plus de règles.

Afin de mieux comprendre les motivations des conducteurs, penchons-nous à présent sur leur comportement.

1.2 Comportement des conducteurs

L'étude des comportements nous offre une meilleure compréhension de la manière dont les conducteurs interagissent entre eux et avec leur environnement. Elle nous permet notamment de comprendre quels facteurs influencent la prise de

décision des conducteurs. Toutefois, il est à noter qu'il existe très peu d'études comportementales décrivant les comportements individuels et les *motifs décisionnels* des conducteurs. Commençons par étudier les motivations des conducteurs (1.2.1), puis observons comment l'hétérogénéité des comportement façonne le trafic (1.2.2).

1.2.1 Facteurs motivationnels

Dans la section précédente, nous avons décrit une partie des motivations des conducteurs. Étoffons cette analyse en commençant par décrire l'activité de conduite.

Pour un individu, l'activité de conduite se décompose en trois niveaux Michon [1985] :

- **Stratégique** — la planification du trajet, la préparation et le suivi ou la modification de l'itinéraire
- **Tactique** — la gestion des situations en cours
- **Opérationnel** — le contrôle du véhicule

À tous ces niveaux, la conduite englobe quatre activités :

1. L'*exploration perceptive*, majoritairement visuelle, lors de laquelle un conducteur cherche des indices nécessaires à l'activité de conduite ;
2. L'*identification* consiste à reconnaître un indice et à le rattacher à une classe d'événements ;
3. La *prévision*, l'anticipation d'événements futurs et des actions potentielles à partir des indices perçus ;
4. Le *processus de décision* qui englobe les processus de sélection et de choix.

Kondoh *et al.* [2008] complètent notre compréhension de la gestion des interactions. Les auteurs ont étudié le risque subjectif des conducteurs dans un contexte de conduite en file. Leur étude a permis d'établir une corrélation entre le moment où les conducteurs freinent et le niveau d'interaction dans lequel ils se trouvent. L'équation résultante, ce qu'ils nomment le *risque perçu* RP, correspond à une somme pondérée de l'inverse de l'intertemps TH — le temps nécessaire pour qu'un véhicule rejoigne la position du véhicule précédent — et du *temps avant collision* $TTC =$ entre les deux véhicules :

$$RP = \frac{1}{TH} + \frac{4}{TTC} \tag{1.9}$$

où l'inverse de TTC équivaut au taux d'expansion de l'angle visuel d'un véhicule qui se rapproche. Bien qu'imparfaite, car trop simpliste, cette métrique nous permet néanmoins de mesurer la subjectivité inhérente à la perception du risque. Malheureusement, aucune étude équivalente ne traite de la perception du risque associée au changement de voie.

Un conducteur humain est par définition opportuniste. Pour étudier l'opportunisme dans le trafic, Ksontini *et al.* [2015] se sont appuyés sur la *théorie de l'affordance*. Cette théorie stipule que le comportement humain résulte de la perception des possibilités offertes par son environnement. Les auteurs ont conçus un modèle de conducteur disposant d'un module de perception fondé sur les affordances pour l'occupation de l'espace routier. Dans leur modèle, les agents se représentent l'environnement routier non plus par des marquages au sol, mais par des opportunités d'occupation de l'espace. Cette modélisation permet de simuler plus finement les comportements des deux roues motorisées dont les trajectoires suivent en réalité des *files virtuelles*.

La gestion des interactions dépend également de la *valeur d'orientation sociale (SVO [1])* des individus. La SVO est une notion de psychologie évaluant les préférences sociales d'un individu et son niveau de coopération [Murphy *et al.*, 2011]. Ce modèle quantifie l'importance accordée à l'utilité d'autrui par rapport à celle accordée à soi-même. Un individu est considéré comme *altruiste* s'il néglige sa propre utilité au profit de celle d'autres personnes, comme *prosocial* s'il attribue une utilité équivalente aux deux, et comme *égoïste* s'il considère son utilité comme supérieure à celle des autres.

Schwarting *et al.* [2019] ont formalisé la SVO pour le domaine du trafic afin de prédire les comportements en temps réel selon leur utilité u. Les auteurs quantifient la SVO d'un individu i par un angle $\phi_i \in \left[-\frac{\pi}{2}; \frac{\pi}{2}\right]$ d'un cercle (figure 1.3) et décrivent les interactions binaires avec les autres conducteurs $j \in \bar{\imath}$ comme un jeu où son utilité est donnée par :

$$u_i = \cos(\phi_i) \cdot \mathcal{R}_i + \sin(\phi_i) \cdot \mathcal{R}_j \tag{1.10}$$

où $\mathcal{R}_i$ et $\mathcal{R}_j$ dénotent des fonctions générales quantifiant respectivement l'utilité de i et j. Cette formalisation retranscrit le fait qu'assister un individu j augmente le gain personnel $\mathcal{R}_i$ d'un conducteur proportionnellement à son niveau d'altruisme ϕ_i. Si i s'avère *égoïste* ($\phi_i = 0$), il ne tirera aucun bénéfice de ses actions empathiques (car $\sin(0) = 0$). Avec un comportement *altruiste* ($\phi_i = \frac{\pi}{2}$), son gain diminuera chaque fois que l'utilité de j baissera.

Comme l'on pouvait s'y attendre, beaucoup de facteurs motivationnels reposent sur des notions subjectives, propres à chaque conducteur. L'évaluation d'une situation et la réponse apportée dépendent de la personnalité, des connaissances, et des compétences de l'individu qui les expérimente. Compte tenu de la diversité comportementale, les situations qui émergent du trafic se caractérisent essentiellement de par leur hétérogénéité.

1. Social Value Orientation

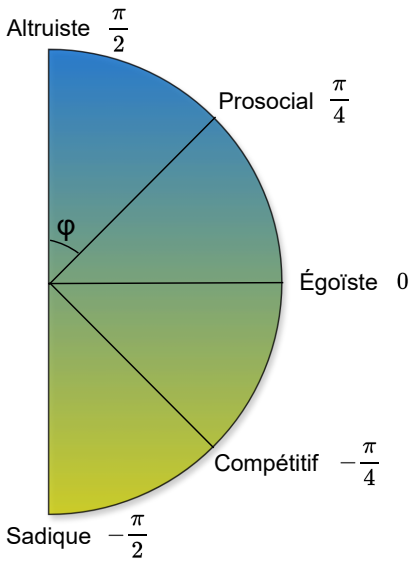


Figure 1.3 – Demi-cercle SVO décrivant les comportements sociaux

1.2.2 Hétérogénéité comportementale

Le système routier est caractérisé par la grande diversité de ses éléments [Munduteguy et Darses, 2007]. Chaque conducteur ou conductrice a une pratique de conduite différente (expérience de conduite, niveau d'attention, rapport au risque), chaque véhicule est différent (type, puissance, poids), chaque situation est différente (infrastructure, caractéristique du trafic, condition météorologique). Du fait de cette hétérogénéité, le trafic apparaît comme un environnement hautement *incertain* où les conducteurs doivent constamment s'adapter à la *dynamique* des situations dans lesquelles ils sont impliqués et aux *contraintes* qui leur sont imposées. L'*hétérogénéité comportementale* conditionne les interactions interindividuelles, elle résulte entre autres des différences de personnalité, de connaissances, et de compétences.

La personnalité d'un individu est en partie conditionnée par sa *culture* [Nordfjærn *et al.*, 2011; Warner *et al.*, 2009; Özkan *et al.*, 2006]. Les *traits* de personnalité influencent la manière dont un individu évalue une situation [Näätänen et Summala, 1974]. Par exemple, l'évaluation du risque inhérent à une situation dépend, en partie du moins, d'une perception et d'une évaluation subjective de la situation. La plupart du temps, le *risque subjectif* perçu par un conducteur ou une conductrice est nul [Näätänen et Summala, 1974], bien qu'il puisse être évalué différemment par ses pairs. Le risque est *multifactoriel*, il dépend à la fois de la situation, du contexte et du niveau de vigilance du conducteur.

Si la personnalité d'un individu caractérise sa subjectivité, ses *connaissances*

lui permettent de diminuer l'incertitude des situations auxquelles il est confronté. Parmi les *connaissances permanentes*, on retrouve les normes qui régissent les interactions et le trafic. Deux types de normes existent selon qu'elles soient formelles ou informelles [Munduteguy, 2001].

Les règles formelles sont établies par la législation, elles visent à uniformiser et à codifier les comportements. Les normes formelles sont inscrites dans le Code de la route, elles incluent entre autres : les priorités, les limites de vitesse et les interdictions. Aucune de ces règles n'est conçue pour être sujette à des interprétations subjectives, et de ce fait, elles servent à diminuer l'incertitude. Leur transgression par les conducteurs découle le plus souvent de leur méconnaissance ou de décisions opportunistes.

Le trafic apparaît comme un *environnement social*, donc aussi régulé par des règles informelles. Les règles informelles peuvent prendre de multiples formes ; les connaître permet une meilleure anticipation des comportements. On devine par exemple qu'un véhicule léger ne reste jamais très longtemps derrière un poids lourd [Munduteguy et Darses, 2007]. Sachant cela, une conductrice peut donc attribuer une *intention* de changer de voie au conducteur du véhicule, parfois avant même que celui-ci ait pris sa décision.

Les connaissances des conducteurs deviennent des *indices* qui enrichissent leur anticipation face aux situations qu'ils rencontrent. La collecte d'indices dépend aussi de l'expérience de conduite qui s'acquiert avec le temps par la pratique. Avec l'*expérience*, les *compétences* des conducteurs s'améliorent ainsi que leurs prédictions et l'efficacité du recueil d'indices sur lesquels elles se fondent. Ainsi, on observe que les conducteurs novices délaissent fréquemment les sous-tâches de niveau intermédiaire/haut pour se focaliser sur le contrôle de leur véhicule. L'expérience joue aussi sur l'anticipation des comportements d'autrui et des risques situationnels, puisqu'un conducteur qui arrive à prédire correctement une situation, prédira aussi plus finement les risques associés.

Ce recueil d'indices dépend principalement de l'*(in)attention* des conducteurs. L'inattention chez le conducteur est définie par Regan *et al.* [2011] comme une « insuffisance ou manque total d'attention pour les activités critiques d'une conduite sûre ». On estime que l'inattention due aux distracteurs externes est la cause principale d'accidents de la route dans 10% des cas, et une cause non négligeable dans 78% des cas [Young *et al.*, 2009].

Le rôle prépondérant de l'inattention dans les accidents de la route s'explique en partie par le fait que l'attention agit tel un *réservoir* dans lequel un individu peut puiser lorsqu'il effectue des tâches cognitives complexes. Les ressources attentionnelles sont limitées et différentes selon les individus. Pour continuer avec cette métaphore, la taille du *réservoir* attentionnel se voit affectée par l'âge, la fatigue, l'état neurologique, le stress et la pratique. Plus un individu dispose de compétences en rapport avec une tâche spécifique, moins la réalisation de cette dernière requiert de ressources. Ainsi, une conductrice expérimentée oriente plus

efficacement son attention et épuise moins ses ressources attentionnelles qu'une novice.

On distingue quatre types d'attention : l'attention sélective, l'attention divisée, l'attention soutenue et la vigilance [McDowd, 2007].

L'*attention sélective* est l'habilité à se concentrer sur des stimuli spécifiques tout en ignorant les informations impertinentes. Lorsque cette attention lui fait défaut, un individu peut passer à côté d'indices sensoriels saillants ou être submergé par un nombre trop important d'indices saillants. Dans le premier cas, c'est l'exemple typique du conducteur qui ne remarque pas que le feu vient de passer au vert.

L'*attention divisée* est la capacité à répartir son attention entre plusieurs tâches ou sources d'informations. Lorsque cette attention lui fait défaut, un individu éprouve des difficultés à exécuter plusieurs tâches simultanément. On l'observe quand un conducteur interrompt momentanément la conversation qu'elle entretenait avec ses passagers pour se focaliser sur sa conduite.

L'*attention soutenue* est la capacité à maintenir sa concentration sur une période prolongée tout en résistant aux distractions, à la fatigue et à l'ennui. Elle ne doit pas être confondue avec la *vigilance* qui permet de maintenir son attention sur de longues durées pour répondre aux stimuli peu fréquents. Comme rapporté par Näätänen et Summala [1974], le risque subjectif influence le niveau de vigilance. Dans la tâche de conduite, ces deux types d'attention sont affectés par la *fatigue*, les *distractions* internes, *e.g.*, des pensées sur d'autres problèmes de la vie quotidienne, et externes, *e.g.*, appels téléphoniques.

Pour résumer, l'attention est une capacité qui filtre les informations et les stimuli impertinents afin de concentrer efficacement les ressources cognitives d'un individu. Les mécanismes qui la sous-tendent restent cependant assez flous et entravent donc sa modélisation. Selon leur niveau d'attention et leur expérience, deux conducteurs peuvent orienter différemment leur recueil d'indices, ce qui les conduit parfois à des attentes différentes quant au résultat d'une même situation [Munduteguy et Darses, 2007].

Lorsque les conducteurs tentent de prédire les comportements d'autrui, ils tentent de déterminer leurs intentions. Cette capacité à déterminer les intentions d'autrui est décrite par la *théorie de l'esprit* [Duval *et al.*, 2011]. Elle joue un rôle fondamental dans les interactions sociales et donc dans les décisions empathiques. Comme le note néanmoins Munduteguy [2001], la théorie de l'esprit ne peut être appliquée systématiquement dans le trafic, car la charge cognitive qu'elle produit outrepasserait les capacités cognitives des individus. L'appel à la théorie de l'esprit est donc certainement réservé aux situations fort incertaines.

Dans la majorité des cas, les conducteurs font référence à des *stéréotypes* afin de déterminer les comportements d'autrui. Selon la *théorie de l'attribution sociale* [Deschamps, 1974], un individu attribue des caractéristiques à un autre suivant le *groupe* auquel il appartient. Un stéréotype est une catégorie dans laquelle on catalogue un individu et qui suscite des *attentes*. Une catégorie peut, par exemple,

dépendre du type de véhicule et de sa puissance, de l'âge du conducteur, de sa manière de conduire. Le stéréotypage tend à atténuer l'incertitude d'un conducteur. Par exemple, certains conducteurs considèrent les taxis comme « imprévisibles » ce qui les poussera à accroître leur niveau de vigilance dans les situations où ces derniers sont présents [Munduteguy et Darses, 2007].

Chaque individu possède ses propres catégories de stéréotypage. De ce fait, les prédictions faites sur le comportement d'un conducteur divergent selon les observateurs.

En résumé, le trafic consiste en un environnement social hétérogène. Chaque conducteur raisonne différemment de ses pairs selon sa personnalité, ses connaissances et son expérience. Le contexte joue également un rôle majeur dans ce processus de raisonnement, même si nous avons préféré nous focaliser sur les comportements individuels. L'attention joue un rôle prépondérant dans la prise de décision, mais ses mécanismes semblent trop complexes pour être *a priori* modélisés. Enfin, l'évaluation subjective des situations par les individus conduit parfois à des attentes différentes, se traduisant par l'observation de comportements hétérogènes.

Dans cette section, nous avons souligné que les causes derrière les comportements observés sont multifactorielles. Globalement, les conducteurs cherchent à minimiser leurs interactions et, le cas échéant, à les supprimer. Cette suppression des interactions est guidée par des processus de pondération des coûts et des bénéfices associés aux changements de voie. On constate qu'à travers les motivations évoquées, les conducteurs recherchent avant tout la stabilité.

Cette stabilité est cependant altérée par l'hétérogénéité comportementale, laquelle résulte entre autres des différences de personnalité, de compétences et de connaissances. L'évaluation subjective des situations effectuée par les conducteurs induit des divergences de comportements. Certains évalueront une situation comme dénuée de risque, tandis que d'autres, au contraire, la jugeront dangereuse.

En conclusion, nous constatons que le trafic se caractérise par l'homogénéité des motivations interindividuelles et l'hétérogénéité des comportements intra-individuels. Deux caractéristiques fondamentales et pourtant absentes de la plupart des modèles informatiques de conducteur humain.

1.3 Robots de conduite

Ces dernières années, l'*automatisation* de haut niveau des modes de transport a gagné en popularité, que ce soit dans l'aviation, le transport maritime ou ferroviaire. Cette évolution semble néanmoins avoir épargné le transport routier. Le principal frein à cette évolution a été discuté dans la section précédente : naviguer dans un trafic routier requiert une fine gestion d'interactions de nature et de complexité diverses. Si l'automatisation du transport routier n'est certes pas globale, nous

verrons dans un premier temps qu'elle n'est pas pour autant complètement absente des véhicules et que plusieurs niveaux d'automatisation existent (1.3.1).

Nombreux sont les défis à relever pour automatiser le transport routier, et de ce fait, plusieurs scénarios sont envisagés quant au futur de la conduite robotisée. D'aucuns envisagent un trafic mixte où conducteurs humains et robots de conduite (RC) cohabiteraient, tandis que d'autres soulignent l'irréalisme d'une telle hypothèse. Nous verrons donc, dans un second temps, les défis posés par le trafic mixte ainsi que les éventuelles solutions pour les surmonter (1.3.2).

1.3.1 Automatisation des véhicules

L'automatisation des véhicules répond à la volonté politique d'accroître la sûreté du trafic par des dispositifs d'*assistance à la conduite* ou de *contrôle du véhicule*. La finalité de l'automatisation, le RC, promet de délester complètement le conducteur de toute charge cognitive liée à la conduite, rendant par la même occasion sa propre présence superflue. L'automatisation des véhicules comporte ainsi six niveaux [Trommer *et al.*, 2016] :

Aucune automatisation — le conducteur humain conduit sans aucune assistance.

1 **Assistance** — le véhicule peut assister le conducteur par un contrôle continu du véhicule longitudinalement **ou** latéralement, mais pas les deux simultanément. Le conducteur doit se tenir prêt à reprendre le contrôle à tout moment. Ce niveau inclut le régulateur de vitesse et l'assistance au maintien dans la voie.

2 **Automatisation partielle** — le véhicule peut assister la conductrice par un contrôle continu du véhicule longitudinalement **et** latéralement. Cette dernière doit se tenir prêt à reprendre le contrôle à tout moment. Ce niveau inclut les systèmes de conduite semi-autonome sur autoroute, comme le Tesla Autopilot [2] ou le General Motors Super Cruise [3].

3 **Automatisation conditionnelle** — le véhicule gère la conduite dans certaines conditions prédéfinies, et demande au conducteur de prendre la relève dans les autres situations. À ce niveau, le système peut piloter le véhicule sur autoroute et dans les embouteillages.

4 **Autonomisation élevée** — le véhicule effectue des tâches de conduite dans certaines conditions et environnements prédéfinis. En dehors de ces conditions, le véhicule demande au conducteur de prendre la relève ou se met en sécurité si ce dernier ne répond pas. Ce niveau englobe les taxis autonomes opérant dans des zones géographiques limitées.

2. https://fr.wikipedia.org/wiki/Tesla_Autopilot
3. https://www.gmc.com/connectivity-technology/super-cruise

5. **Autonomisation complète** — le véhicule gère l'ensemble des tâches de la conduite sans aucune intervention humaine requise. À ce niveau, le système est totalement autonome.

Les cas prédéfinis correspondent à certains types de routes, à certaines échelles de vitesse ou encore à des conditions environnementales spécifiques.

La mise en place de nouveaux systèmes d'assistance ou d'automatisation engendre systématiquement des *adaptations comportementales* directes pour les conducteurs des véhicules équipés ou indirectes pour les autres usagers avec qui ces conducteurs interagissent. Prenons l'exemple de l'*ACC*[4] qui vise à améliorer le confort des conducteurs en les délestant du besoin de constamment adapter leur vitesse au véhicule qui les précède. En simulation, Vander Werf *et al.* [2002] montrent qu'une adoption globale permettrait de doubler le volume du trafic sur autoroute, mais qu'en deçà de 40% de *taux de pénétration*, l'impact de cette technologie s'avère négligeable.

L'évaluation de l'impact de ces systèmes pose néanmoins plusieurs problèmes énumérés par Saad [2006] :

1. Le premier problème concerne la diversité des changements comportementaux qu'ils entraînent. Par exemple, des recherches ont conduit à des résultats contradictoires lorsqu'elles ont mesuré l'impact de l'ACC sur les comportements des conducteurs. Certains concluaient que la vitesse des conducteurs augmentait lorsqu'ils utilisaient l'ACC, d'autres ne remarquaient aucun changement sur ce point.
2. Le second problème concerne la diversité des situations routières. Certains systèmes sont conçus pour répondre à des situations particulières et s'avèrent inefficaces dans d'autres contextes. Les comportements des conducteurs changent selon les conditions environnementales et l'acceptation de ces systèmes dépend également de leur capacité à considérer et à réagir adéquatement aux changements circonstanciels.
3. Le troisième problème concerne le différentiel d'efficacité selon les conducteurs. Plus précisément, l'efficacité de ces systèmes diffère selon les styles de conduite et la personnalité des conducteurs. Par exemple, l'ACC aurait un impact négatif sur les conducteurs en recherche de sensations.
4. Le quatrième problème concerne l'apprentissage de ces nouveaux systèmes. Les conducteurs doivent apprendre à interagir avec ces systèmes et de la qualité de ces interactions naît une confiance en eux ou un sentiment de défiance.

Ces quatre problèmes montrent la complexité inhérente à l'évaluation de tout système d'automatisation. Les changements comportementaux qu'ils induisent se révèlent difficiles à prévoir et varient selon les individus, les situations et l'aisance

4. Adaptative Cruise Control

avec laquelle les conducteurs interagissent avec ces systèmes. De plus, leur efficacité varie selon leur taux de pénétration.

1.3.2 Vers un trafic mixte ?

En matière de sécurité routière, les politiques mondiales [5] comme nationales [6] envisagent l'automatisation complète des véhicules à l'horizon 2050. Leurs motivations reposent sur l'hypothèse selon laquelle un trafic automatisé renforcerait la sécurité des usagers et diminuerait les congestions. En supposant la confirmation de cette hypothèse, il semble nécessaire d'étudier la phase transitoire, transformant graduellement le trafic actuel au trafic automatisé.

Cette transition, appelée *trafic mixte*, requiert une bonne *cohabitation* entre RC et conducteurs humains [Dinneweth *et al.*, 2022]. Cette cohabitation ne présente néanmoins rien d'évident. Les RC sont conçus pour se conformer aux normes formelles, tandis que les conducteurs humains raisonnent également selon des *normes sociales* et informelles pouvant parfois se révéler contraires à la réglementation du Code de la route. L'application de systèmes de règles contradictoires produit des interactions propices aux incidents [Saad et Mundutéguy, 2002]. En conditions réelles, des chercheurs ont constaté une augmentation de la fréquence de ce type de collision [Petrović *et al.*, 2020].

Par exemple, lorsqu'un RC arrive au niveau d'un panneau Stop, il marque un temps d'arrêt de trois secondes conformément à la législation en vigueur. Dans la même situation, les pratiques des certains conducteurs consistent à décélérer, puis à marquer un arrêt uniquement si un véhicule de la voie principale approche, voire à repartir sans marquer d'arrêt. Lorsqu'à cet arrêt, un conducteur se trouve derrière un RC, il s'attend à le voir repartir aussitôt que l'absence de véhicule gênant sur la voie principale est observée. Si le RC marque un arrêt réglementaire en l'absence de véhicule sur la voie principale, il contredit alors les attentes du conducteur qui, dans certains cas, l'emboutit. En conditions réelles, on constate que les RC se font emboutir par l'arrière plus fréquemment que les conducteurs humains [Petrović *et al.*, 2020].

Du fait de ces *dissonances cognitives*, d'aucuns jugeraient opportun de restreindre la durée de cette phase de cohabitation. Dans les faits, le trafic se compose d'usagers dont les modes de transports ne font l'objet d'aucune automatisation, tels que des vélos et des trottinettes. On peut donc considérer que le trafic mixte persistera malgré l'automatisation des véhicules et que son étude semble, par conséquent, cruciale afin d'assurer une bonne cohabitation entre les divers usagers.

En théorie, l'introduction de RC améliorerait la sécurité et l'efficacité du trafic, mais cette tendance s'avère toutefois non linéaire, ce qui signifie qu'elle est corrélée

5. https://www.who.int/fr/news-room/fact-sheets/detail/road-traffic-injuries

6. https://www.strategie.gouv.fr/sites/strategie.gouv.fr/files/atoms/files/note_danalyse-na47-voiture-sans-chauffeur.pdf

au taux de pénétration [Guériau et Dusparic, 2020]. Plusieurs auteurs affirment qu'un faible taux de pénétration ($\leq$ 20-40%) provoquerait des conséquences néfastes comme une augmentation de la congestion du trafic et des incidents [Arvin *et al.*, 2018]. Le contexte expérimental qui sous-tend ces résultats paraît toutefois discutable, car les comportements humains simulés dépendent des modèles de type Wiedmann ou IDM-MOBIL qui diffèrent très largement des pratiques humaines observées [Dinneweth *et al.*, 2022].

La plupart des systèmes autonomes de navigation visent à rester en dehors des trajectoires supposées des autres usagers [Sadigh *et al.*, 2018]. Dans cette optique, le RC se voit contraint d'adopter un comportement défensif et d'éviter toute initiative, rendant en définitive ses décisions totalement prévisibles. Dès lors, cette *prévisibilité* devient une *vulnérabilité* que les usagers *opportunistes* tenteront certainement d'exploiter. Nous avons d'ailleurs déjà étudié les comportements opportunistes dans la première section, où l'impatience des conducteurs conduit à une transgression des règles de priorité sur un carrefour en croix [Doniec *et al.*, 2008a]. D'après les auteurs, un comportement opportuniste résulte (1) d'une *précondition*, une propension à violer le Code de la route et (2) d'une *postcondition*, une condition nécessaire à l'exécution d'une politique opportuniste. Dans un carrefour en croix, la précondition correspond à la disposition de l'agent à violer une norme formelle (la priorité) nourrit par un sentiment d'impatience, et la postcondition dépend du différentiel de dynamique du véhicule à qui l'agent souhaite couper la priorité.

On peut soupçonner que des RC trop prévisibles, dont l'unique objectif consisterait à éviter toute perturbation sur le trafic, produisent en définitive le résultat contraire. Ce comportement risquerait en effet d'être interprété comme une vulnérabilité que des usagers opportunistes exploiteraient, puisqu'elle signifierait l'activation d'une postcondition. Il suffit alors qu'un nombre conséquent d'usagers opportunistes prennent conscience de l'influence qu'ils peuvent exercer sur ces RC pour qu'émerge une perturbation locale du trafic. Selon les situations, les conséquences peuvent se répercuter à des échelles bien plus larges et conduire à des incidents.

Sans le formuler ainsi, certains auteurs ont souhaité pallier ce problème en jouant sur l'influence que les RC pourraient exercer sur les conducteurs humains [Sadigh *et al.*, 2018]. Leurs travaux ont consisté à modéliser les interactions binaires entre conducteurs humains et RC. Du fait de l'absence de données sur ce type d'interactions, les auteurs ont modélisé les prises de décisions des RC avec de l'*apprentissage par renforcement*, une approche que nous développerons au prochain chapitre. Leurs résultats montrent que les sujets humains, au volant du second véhicule, se voient effectivement influencés par les décisions du RC.

Cette preuve de concept a permis d'ouvrir la voie à la conception de RC *socialement désirables*, notamment par l'ajout de valeur d'orientation sociale (SVO) aux systèmes décisionnels [Toghi *et al.*, 2021a,b, 2022]. Confirmant nos doutes, les conclusions de ces auteurs révèlent que la conception de RC trop

altruistes s'avère néfaste tant pour la fluidité du trafic que pour sa sécurité, du moins dans le contexte étudié d'insertion sur autoroute. Par ailleurs, une attitude profondément égoïste conduit à des résultats similaires.

Également menée en simulation, une étude montre qu'en trafic mixte, les conducteurs humains se montrent plus altruistes envers leurs pairs qu'envers des RC [Sun *et al.*, 2024]. Selon les auteurs, les RC semblent mieux acceptés par les conducteurs humains lorsqu'ils se comportent de manière ni trop conservatrice ni trop agressive.

En définitive, la faisabilité du trafic mixte semble compromise, du moins si les RC se cantonnent à appliquer des stratégies défensives, trop influençables, qui tendent à produire l'exact inverse de l'effet escompté. L'opportunisme naturel des conducteurs conduira probablement ces derniers à exploiter de telles vulnérabilités pour se dérober aux situations contraignantes. Dans ce cas, l'effet sur le trafic pourrait se révéler désastreux tant pour la vitesse du flux que pour la sécurité.

Il existe toutefois une alternative qui consiste à atténuer les différences comportementales qui distinguent les conducteurs humains des RC. Ce paradigme vise à doter les RC de caractéristiques humaines et sociales et à mimer leurs comportements. En plus d'accroître l'acceptabilité des RC tant par leurs passagers que par les autres usagers, nous pensons que cela permettrait de contenir l'émergence de nouvelles interactions. De fait, la cohabitation et la sécurité, conditions capitales à la concrétisation du trafic mixte, s'en trouveraient probablement améliorées.

Évidemment, tant la diversité des contextes que l'hétérogénéité des comportements représentent un défi colossal pour les systèmes décisionnels autonomes de par l'incertitude qu'elles génèrent. Pour lever cette incertitude, les humains émettent des hypothèses fondées sur leur expérience de conduite, leurs connaissances et des indices situationnels et comportementaux parfois biaisés ou stéréotypés. Leurs prédictions s'avèrent justes dans plus de 90% des cas [Saad et Mundutéguy, 2002]. Le processus cognitif humain semble cependant trop complexe pour être répliqué dans sa totalité.

Conclusion

Dans ce chapitre, nous avons définit le concept de simulation de trafic. Nous avons distingué trois approches concernant sa modélisation : analytiques, règles décisionnelles et motivationnelles. Les approches analytiques modélisent le trafic à l'échelle macroscopique ou microscopique, mais se fondent sur des lois mathématiques de comportements obtenus par régression et de ce fait, prennent mal en compte le contexte des situations de conduite. Les approches par base de règles décisionnelles modélisent les comportements des agents qui prennent des décisions en fonction du contexte environnemental. Enfin, les approches motivationnelles comprennent un nombre limités de motivations – génériques – que poursuivent les

agents. La grande majorité des modèles étudiés négligent l'aspect psychologique inhérent aux prises de décision des conducteurs.

Les études psychologiques constatent que le trafic se caractérise par une homogénéité des motivations interindividuelles ainsi qu'une hétérogénéité des comportements intra-individuels. Globalement, les conducteurs recherchent la stabilité et tendent à minimiser leurs interactions tout en optimisant certains critères tels que minimiser leur temps de parcours et maintenir le risque perçu sous un seuil acceptable pour eux. L'évaluation des situations par les conducteurs s'avère subjective et multifactorielle, chacun perçoit le risque différemment selon sa personnalité, sa culture, son niveau d'attention, ses compétences, son expérience, ses émotions, etc.

Ces caractéristiques rendent le trafic difficilement appréhendable par les RC. Les systèmes décisionnels qui adoptent une conduite respectant strictement le Code de la route et des règles de prudence élevées deviennent prévisibles, ce qui les rend vulnérables du point de vue des conducteurs opportunistes. L'exploitation du comportement prévisible des RC par des conducteurs opportunistes – distances intervéhiculaires importantes, respect strict de la priorité – peut engendrer des perturbations sur le trafic, voire des accidents dans le pire des cas. Ces nouveaux types d'interaction émergent du fait de différences motivationnelles entre RC et conducteurs humains. Nous pensons que la sécurité du trafic mixte pourrait être améliorée si les RC imitaient les pratiques humaines lorsqu'elles sont sans danger.

Ainsi, notre thèse ambitionne de réduire les écarts qui existent entre ces deux catégories d'usagers. Pour y arriver, nous souhaitons concevoir un modèle décisionnel pour les RC qui imite les pratiques réelles des conducteurs tout en préservant un certain niveau de sécurité. Ce premier chapitre nous a permis d'identifier les éléments clés qui caractérisent les conducteurs et leurs interactions. Certains d'entre-eux seront exclus de notre modèle du fait de l'absence de connaissances suffisantes à leur modélisation. Nous pensons, par exemple, aux éléments tels que les mécanismes d'attention ou à l'influence des émotions sur le comportement.

Notre modèle tentera cependant de reproduire certaines caractéristiques fondamentales du trafic telles que l'hétérogénéité des profils de conducteurs et la subjectivité inhérente à leur évaluation des situations. Nous pensons que l'hétérogénéité diminuera la prévisibilité des RC et ainsi leur vulnérabilité face aux conducteurs opportunistes. Ce modèle pourra également être paramétré afin de s'adapter aux spécificités comportementales locales et culturelles. Globalement, nous supposons que la sécurité et l'efficacité du trafic mixte seraient moins perturbées si l'on pouvait atténuer les différences entre les usagers qui le composent, ce qui semble crucial au commencement de la phase de transition où le taux de pénétration des RC sera faible.

Nous pourrions concevoir notre modèle sur la base de données récoltées en condition réelle, mais ce serait omettre que les comportements dépendent du contexte et de la psyché des conducteurs, deux éléments que les capteurs ne

peuvent retranscrire. Du fait des inconvénients et des limites des modèles mathématiques comme des modèles à base de règles décisionnelles, nous optons pour une approche hybride alliant apprentissage et émergence. Plus précisément, nous envisageons l'apprentissage par renforcement pour deux raisons. La première est que l'apprentissage par renforcement permet de résoudre des problèmes séquentiels de prise de décisions dont les objectifs sont renseignés sous la forme de motivations, ce qui s'approche du mode de fonctionnement humain. La seconde est que l'apprentissage par renforcement s'avère fertile à l'émergence.

Nos travaux se focaliseront sur un scénario d'insertion sur autoroute du fait des défis de cette situation tant pour les humains que pour les RC. Dans ces situations, (1) la fusion de flux de véhicules provoque un nombre plus élevé d'interactions, (2) cet effet se trouve exacerbé par le différentiel de vitesse entre les flux, et (3) l'anticipation des comportements des conducteurs influence considérablement le déroulement de ces interactions.

Ainsi, nous pourrions simuler de nouveaux types d'interactions spécifiques au trafic mixte. De plus, l'approche hybride permettrait d'expliquer les décisions prises et potentiellement d'accroître l'acceptabilité d'un tel système [Miller, 2019]. Le prochain chapitre apportera les clés de compréhension quant au fonctionnement et aux défis posés par l'apprentissage par renforcement dans un contexte multiagent [Bazzan et Klügl, 2009; El Fallah-Seghrouchni *et al.*, 1999].

CHAPITRE

Apprentissage par renforcement multiagent

2

Sommaire du chapitre

L'apprentissage par renforcement appartient au domaine de l'intelligence artificielle et de la science des données. Ce dernier regroupe trois familles d'algorithmes : (1) l'apprentissage supervisé traite les problèmes de classification et de régression en entraînant un modèle avec des données étiquetées [Cunningham *et al.*, 2008] ; (2) l'apprentissage non supervisé discrimine des motifs et organise des structures à partir de données non étiquetées pour les rendre plus compréhensibles [Barlow, 1989] ; (3) l'apprentissage par renforcement (RL) répond aux problèmes de prise de décision séquentielle [Sutton et Barto, 2018].

Contrairement aux autres formes d'apprentissage, le RL autogénère ses don-

nées [Lapan, 2020]. Le RL jouit d'une capacité de généralisation des situations lui permettant de s'adapter à un environnement labile comme le trafic. Cette propriété répond aux exigences de notre thèse, où, rappelons-le, nous souhaitons étudier des adaptations comportementales.

Ce second chapitre de contexte comporte trois sections. La section 2.1 introduira les différentes composantes d'une approche par apprentissage par renforcement multiagent (MARL). Nous y définirons les problèmes MARL selon un couple formé d'un modèle de jeu (2.1.1) et d'un concept de solution (2.1.2). Le premier établit les règles du problème, tandis que le second précise l'ensemble des solutions acceptables.

Une fois le problème défini, la section 2.2 listera les éléments nécessaires à sa résolution. Cette liste comprend une fonction d'approximation, un algorithme et un mode d'entraînement. Dans les approches MARL, la fonction d'approximation consiste souvent en un réseau de neurones. Nous examinerons les propriétés qui rendent les réseaux de neurones incontournables et dépeindrons les diverses architectures existantes (2.2.1). Puis, considérant les caractéristiques de notre problématique, nous dresserons l'ensemble des critères que devra respecter l'algorithme retenu (2.2.2). Nous terminerons cette section par la sélection d'un mode d'entraînement et d'exécution (2.2.3).

Enfin, la section 2.3 identifiera les défis posés par la résolution de problèmes MARL. Notre étude se focalisera sur les deux principaux écueils, à savoir : la non-stationnarité (2.3.1) et la dimensionnalité (2.3.2). Le premier prévient toute garantie de convergence et le second exacerbe les problèmes du premier.

2.1 Définir le problème

Un système multiagent est un système composé de plusieurs entités autonomes (*agents*) ayant des informations différentes ou des objectifs différents, ou les deux [Shoham et Leyton-Brown, 2008]. Commençons par décrire ces composants.

L'environnement est un monde réel ou virtuel dont l'*état* évolue au gré du temps et des *actions* effectuées par les agents. Un environnement spécifie les *observations* envoyées aux agents et les actions dont ils disposent. Les états comme les actions se composent de variables discrètes ou continues. Souvent, seule une partie de l'état est observable par les agents, compliquant ainsi leur prise de décision.

Les agents correspondent à des entités autonomes qui entreprennent des actions selon leur perception de l'environnement. Ils disposent d'objectifs à atteindre, dont la nature dépend du problème. Par exemple, ces objectifs peuvent être d'accumuler une quantité (maximiser un revenu) ou d'atteindre un état particulier de l'environnement (franchir une ligne d'arrivée).

L'*apprentissage par renforcement (RL)* s'emploie pour résoudre le problème d'un système multiagent. On parle alors d'*apprentissage par renforcement multiagent*

(MARL) [Sutton, 1990; Sutton et Barto, 2018]. Dans cette configuration, les agents disposent de *fonctions de récompense* en guise d'objectif. Ces fonctions envoient des signaux scalaires, des récompenses, après chaque interaction d'un agent avec son environnement (figure 2.1). Les récompenses guident les agents dans leur apprentissage d'une stratégie d'action appelée *politique*.

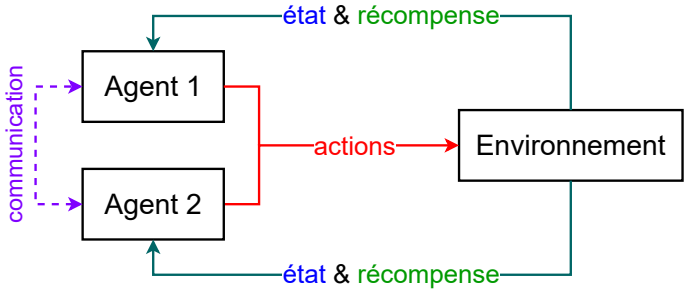


Figure 2.1 – MARL avec deux agents

L'apprentissage par renforcement s'effectue par essai-erreur. Les agents doivent explorer l'environnement pour découvrir les séquences d'actions les plus fructueuses. Un agent converge vers une *politique optimale* s'il accumule le maximum de récompenses possibles.

L'approche MARL a été envisagée pour la résolution de nombreux problèmes de natures diverses : jeux de plateaux [Silver *et al.*, 2017], jeux vidéos [Baker *et al.*, 2019; Vinyals *et al.*, 2019], robotique [Perrusquía *et al.*, 2021], conduite autonome [Dinneweth *et al.*, 2022], finance [Karpe *et al.*, 2021], gestion de réseaux électriques [Chen *et al.*, 2021], *etc.* Toutefois, la majorité des applications à succès considèrent un faible nombre d'agents, souvent deux.

Un problème MARL se définit selon un couple composé d'un *modèle de jeu* (2.1.1) et d'un *concept de solution* (2.1.2). En théorie des jeux, un modèle de jeu décrit formellement les interactions entre les agents, tandis qu'un concept de solution explicite les solutions du jeu en question [Fudenberg et Tirole, 1991].

2.1.1 Modèle de jeu

Plusieurs modèles de jeu existent. Le plus simple, le *jeu de forme normale*, définit une interaction non répétée entre plusieurs agents. Ce modèle décrit, par exemple, les problèmes dits de type *bandits*, où un agent doit déterminer parmi un ensemble de machines à sous, celle avec le meilleur rendement [Bergemann et Valimaki, 2018].

Définition 3: Jeu de forme normale

Un jeu de forme normale consiste en un tuple $\langle \mathcal{I}, \{\mathcal{A}_i\}_{i\in\mathcal{I}}, \{\mathcal{R}_i\}_{i\in\mathcal{I}} \rangle$ avec :

- un ensemble fini d'agents $\mathcal{I} = \{1, \ldots, |\mathcal{I}|\}$

- un ensemble fini d'actions $\mathcal{A}_i$ pour chaque agent i
- une fonction de récompense $\mathcal{R}_i : \mathcal{A} \to \mathbb{R}$ pour chaque agent, où $\mathcal{A} = \mathcal{A}_1 \times \ldots \times \mathcal{A}_{|\mathcal{I}|}$ représente l'espace des actions jointes

Dans un jeu de *forme normale*, chaque agent adopte une politique $\pi_i : \mathcal{A}_i \to [0;1]$ qui assigne des probabilités aux actions disponibles, telles que $\sum_{a_i \in \mathcal{A}_i} \pi_i(a_i) = 1$. Chaque agent tire une action $a_i \in \mathcal{A}_i$ selon une probabilité $\pi_i(a_i)$ renseignée par sa politique. L'action jointe $a = \{a_1, \ldots, a_{|\mathcal{I}|}\}$ résultante correspond à une interaction des agents avec l'environnement qui renvoie leurs récompenses respectives $r_i = \mathcal{R}_i(a)$. Les jeux de forme normale se distinguent selon leur mode de répartition des récompenses :

- Dans les *jeux à somme nulle*, la somme des récompenses vaut toujours zéro, $\forall a \in \mathcal{A} : \sum_{i \in \mathcal{I}} \mathcal{R}_i(a) = 0$. La récompense d'un agent i équivaut toujours à l'opposé de celle reçue par ses opposants $\mathcal{R}_i = \mathcal{R}_{\bar{\imath}}$, où $\bar{\imath} = \mathcal{I} \setminus \{i\}$ dénote l'ensemble des opposants de i. Cette formalisation décrit les jeux *compétitifs* où chaque gain d'un agent engendre une perte chez ses adversaires.
- Dans les *jeux à récompense commune*, tous les agents reçoivent une récompense identique, $\mathcal{R}_i = \mathcal{R}_{\bar{\imath}}$. Cette formalisation décrit les jeux *coopératifs* où l'ensemble des agents partagent un objectif commun.
- Similairement au précédent point, les *jeux à moyenne d'équipe* encouragent la coopération, mais d'une manière moins restrictive [Gronauer et Diepold, 2022]. Ici, les agents peuvent disposer de fonctions de récompense différentes, bien que toutes favorisent la coopération $\bar{\mathcal{R}} = \frac{1}{|\mathcal{I}|} \sum_{i \in |\mathcal{I}|} \mathcal{R}_i$. Cette formalisation convient, par exemple, aux jeux par équipes, où chacun individu endosse un rôle particulier tout en cherchant à gagner collectivement.
- Dans les *jeux à somme générale*, aucune restriction ne s'applique aux fonctions de récompense. Les agents peuvent coopérer ou rivaliser selon les circonstances. Ces jeux conviennent à la simulation de trafic où les usagers adoptent tantôt un comportement coopératif, tantôt compétitif.

Certains problèmes peuvent se définir à l'aide d'une combinaison de jeux. Par exemple, un match de football opposant deux équipes correspond à un jeu à somme nulle, car une seule équipe l'emportera, et à un jeu à moyenne d'équipe, car les joueurs d'une même équipe coopèrent pour gagner.

Les *jeux répétés de forme normale* introduisent de la séquentialité dans les interactions. Ces derniers répètent T fois des jeux de forme normale, où T est fini ou infini. À chaque pas de temps $t = 0, \ldots, T-1$, chaque agent échantillonne une action a_i^t avec une probabilité $\pi_i\left(a_i^t \mid h^t\right)$. Ainsi, la politique est conditionnée par l'*historique des actions jointes* $h^t = \left(a^0, \ldots, a^{t-1}\right)$ entreprises avant l'instant t. Cet historique permet un raisonnement statistique. Par exemple, dans un problème

de type *bandit manchot*, un agent peut maintenir un historique des gains pour chaque machine à sous et ainsi accumuler plus de bénéfices.

À la différence des jeux infinis, les *jeux finis* $T \in [0;+\infty[$ comportent une probabilité $1-\gamma$ de se terminer à chaque instant, avec $\gamma \in [0;1]$ le *facteur de réduction*. Le facteur de réduction γ reflète l'incertitude sur le long terme et influence les comportements des agents. Les comportements des agents peuvent donc différer selon l'imminence de la fin du jeu. Par conséquent, modifier γ revient à changer la nature même du problème à résoudre Naik *et al.* [2019]. Pour les jeux infinis, le facteur de réduction γ n'a pas de sens concret.

Jusqu'ici, les jeux étudiés dépeignent une unique interaction, restreignant le nombre de problèmes pouvant être modélisés. Les *jeux stochastiques* lèvent cette limitation avec un état qui évolue selon les interactions des agents et les probabilités de transition des états [Shapley, 1953].

Définition 4: Jeu stochastique

Un jeu stochastique se définit par un tuple $\langle \mathcal{I}, \mathcal{S}, \{\mathcal{A}_i\}_{i\in\mathcal{I}}, \{\mathcal{R}_i\}_{i\in\mathcal{I}}, \mathcal{T}, \mu_0 \rangle$ avec :

- un ensemble fini d'agents $\mathcal{I} = \{1, \ldots, |\mathcal{I}|\}$
- un ensemble fini d'états $\mathcal{S}$
- un ensemble fini d'actions $\mathcal{A}_i$ pour chaque agent i
- une fonction de récompense $\mathcal{R}_i : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ pour chaque agent où $\mathcal{A} = \mathcal{A}_1 \times \ldots \times \mathcal{A}_{|\mathcal{I}|}$
- une fonction de probabilité de transition entre états $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0;1]$ tel que $\forall s \in \mathcal{S}, a \in \mathcal{A} : \sum_{s'\in\mathcal{S}} \mathcal{T}(s,a,s') = 1$
- la distribution initiale d'états $\mu_0 : \mathcal{S} \to [0;1]$ tel que $\sum_{s\in\mathcal{S}} \mu_0(s) = 1$

Un jeu stochastique commence dans l'*état initial* $s^0 \in \mathcal{S}$ échantillonné depuis μ_0 [Littman, 1994]. À chaque instant t, chaque agent i reçoit l'état courant $s^t \in \mathcal{S}$ de l'environnement, prend une action $a^t \in \mathcal{A}_i$ selon une probabilité $\pi\left(a_i^t \,\middle|\, h^t\right)$. Leur politique produit l'*action jointe* a^t et repose sur l'*historique d'état-action* $h^t = \left(s^0, a^0, s^1, a^1, \ldots, s^t\right)$. L'historique est dit *entièrement observable*, car il contient les informations de l'ensemble des agents et tous y ont accès. Après avoir exécuté l'action jointe, le jeu bascule dans un état suivant $s^{t+1} \in \mathcal{S}$ avec une probabilité $\mathcal{T}\left(s^t, a^t, s^{t+1}\right)$ et chaque agent reçoit une récompense $r_i^t = \mathcal{R}_i\left(s^t, a^t, s^{t+1}\right)$. Ce processus se répète jusqu'à ce que l'environnement atteigne un *état terminal* ou indéfiniment pour les jeux infinis.

Tout comme les *processus de décision Markoviens*, les jeux stochastiques respectent la *propriété de Markov* [Puterman, 1994]. En cela, on les nomme parfois *jeux de Markov*.

Propriété 1: Propriété de Markov

La distribution de probabilité conditionnelle des états et des récompenses futures dépend uniquement de l'état courant.

$$\mathbb{P}\left[s^{t+1}, r^t \,\middle|\, s^t, a^t, \ldots, s^0, a^0\right] = \mathbb{P}\left[s^{t+1}, r^t \,\middle|\, s^t, a^t\right] \tag{2.1}$$

avec $r^t = \left(r_1^t, \ldots, r_{|\mathcal{I}|}^t\right)$ la récompense jointe à l'instant t.

Autrement dit, l'état courant s^t doit contenir l'ensemble des informations requises à la modélisation de la fonction de transition T.

Parfois, les agents ne peuvent observer que partiellement l'état de l'environnement. Ce cas est décrit par les *jeux stochastiques partiellement observables (POSG)* [Hamila, 2012]

Définition 5: Jeu stochastique partiellement observable

Un jeu stochastique partiellement observable englobe les éléments du jeu stochastique, auxquels se greffent deux éléments, $\left\langle \{\mathcal{O}_i\}_{i\in\mathcal{I}}, \{\Omega_i\}_{i\in\mathcal{I}} \right\rangle$ avec :

- un ensemble fini d'observations $\mathcal{O}_i$
- une fonction d'observation $\Omega_i : \mathcal{A} \times \mathcal{S} \times \mathcal{O}_i \to [0;1]$ tel que $\forall a \in \mathcal{A}, s \in \mathcal{S} : \sum_{o_i \in \mathcal{O}_i} \Omega_i(a, s, o_i) = 1$

Un POSG définit les probabilités d'observation d'états $\mathbb{P}\left[s^t, o^t \,\middle|\, s^{t-1}, a^{t-1}\right]$, où $o^t = \left(o_1^t, \ldots, o_{|\mathcal{I}|}^t\right)$ est l'*observation jointe* à un instant t. À tout instant t, chaque agent reçoit une observation $o_i^t \in \mathcal{O}_i$ avec une probabilité $\Omega_i\left(o_i^t \,\middle|\, a^{t-1}, s_t\right)$. Chaque agent tire ensuite une action a_i^t selon une probabilité $\pi_i\left(a_i^t \,\middle|\, h_i^t\right)$ où l'*historique d'observation* $h_i^t = \left(o_i^0, \ldots, o_i^t\right)$ ne contient plus que ses propres observations.

Du fait de l'observabilité partielle, l'agent se trouve dépourvu de moyens directs fiables pour prédire les politiques des autres agents $\bar{\imath} = \mathcal{I} \setminus \{i\}$ et, au travers de leur politique jointe, la fonction de transition. Cette incertitude conduit l'agent à estimer la fonction de transition au travers d'un *état de croyance* β_i^t, soit la probabilité de distribution sur les états $s \in \mathcal{S}$, avec $\beta_i^0 = \mu_0$ l'*état de croyance initial*. L'état de croyance s'actualise à chaque instant t, tel que :

$$\beta_i^{t+1}(s') \propto \sum_{s\in\mathcal{S}} \beta_i^t(s) \mathcal{T}\left(s' \,\middle|\, s, a_i^t\right) \Omega\left(o_i^{t+1} \,\middle|\, a_i^t, s'\right) \tag{2.2}$$

En théorie, l'espace mémoire requis pour sauvegarder l'état de croyance et le temps nécessaire à sa mise à jour sont chacun exponentiel par rapport au nombre de variables qui définissent un état. En pratique, seule une approximation de l'état de croyance est raisonnablement envisageable pour les environnements complexes [Albrecht *et al.*, 2023]. Nous détaillerons ultérieurement des méthodes d'approximation.

D'autres formes de jeu existent, comme les jeux à formes extensives, formalisant les jeux où les agents prennent des actions tour à tour, mais nous ne les développerons pas. Pour l'heure, bornons-nous au second élément qui constitue un problème MARL, à savoir le concept de solution.

2.1.2 Concept de solution

Si le modèle de jeu définit les processus d'interactions, le concept précise ses solutions acceptables. Le concept de solution apparaît évident pour un jeu à récompense commune : maximiser l'espérance du *retour* u (éq. 2.3) ; mais il l'est moins pour les autres modèles.

$$u^t = r^{t+1} + \gamma r^{t+2} + \ldots = \sum_{k=0}^{\infty} \gamma^k r^{t+k+1} \tag{2.3}$$

Dans ce qui suit, nous supposons que le jeu est fini, *i.e.*, les espaces d'agent, d'état, d'action et d'observation sont finis et discrets. Une *solution* consiste en une *politique jointe* $\pi = (\pi_1, \ldots, \pi_{|\mathcal{I}|})$ satisfaisant des exigences propres au concept qui la sous-tend en termes de *retour espéré* $u_i(\pi)$ pour chaque agent i sous la politique jointe π :

$$u_i^t(\pi) = \sum_{t=0}^{t-1} \gamma^t \mathcal{R}_i \left(s^t, a^t, s^{t+1}\right) \tag{2.4}$$

Ce retour peut s'écrire par les *fonctions de valeur d'état* $V : \mathcal{S} \to \mathbb{R}$ ou de *qualité d'action* $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$:

$$V_i^{\pi}(s) = \sum_{a \in \mathcal{A}} Q^{\pi}(s, a) \pi\left(a \,|\, s\right) \tag{2.5}$$

$$Q_i^{\pi}(s, a) = \sum_{s' \in \mathcal{S}} \mathcal{T}\left(s' \,|\, s, a\right) \left[\mathcal{R}_i\left(s, a, s'\right) + \gamma \sum_{o' \in \mathcal{O}} \Omega\left(o' \,|\, a, s'\right) V_i^{\pi}\left(s'\right)\right] \tag{2.6}$$

La figure 2.2 souligne les différences entre ces deux fonctions. La fonction de valeur d'état V estime les *récompenses futures* qu'un agent peut espérer obtenir à partir d'un état s. La fonction de qualité d'action Q estime les récompenses futures qu'un agent peut espérer obtenir à partir d'un état s en prenant une action a. Cette dernière fonction estime donc plus finement les récompenses futures, mais nécessite en contrepartie plus d'interactions pour être correctement estimée. Dans la figure, la fonction de valeur V estime que l'état s_0 vaut 0 en moyenne si l'agent adopte une politique stochastique. Cette valeur correspond à la moyenne des récompense (1 et -1). Dans cet exemple, la fonction de valeur V n'apporte aucune information utile, contrairement à la fonction de qualité d'action Q qui renseigne l'action la plus avantageuse.

En *théorie des jeux*, la *meilleure réponse* sert de fondement aux autres concepts de solution.

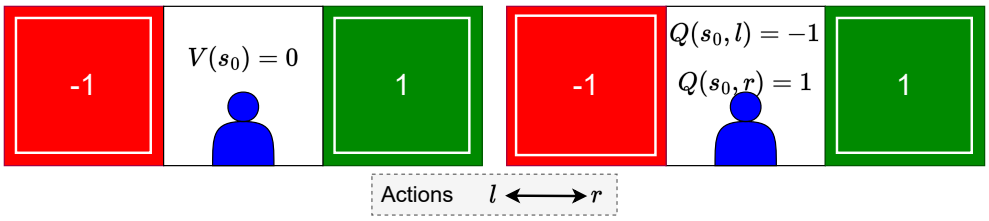


Figure 2.2 – Fonctions de valeur d'état $V(s)$ (gauche) et de qualité d'action $Q(s,a)$ (droite). L'agent (personnage bleu) peut se déplacer selon deux axes jusqu'à atteindre l'un des deux états finaux (cellules colorées) de l'environnement et remporter leur récompense respective

Définition 6: Meilleure réponse

La meilleure réponse d'un agent BR_i face à la politique jointe des autres agents $\pi_{\bar{\imath}}$ consiste à maximiser son retour u_i espéré tel que :

$$BR_i\left(\pi_{\bar{\imath}}\right) = \underset{\pi_i}{\operatorname{argmax}}\, u_i\left(\langle \pi_i, \pi_{\bar{\imath}} \rangle\right) \tag{2.7}$$

Plusieurs meilleures réponses peuvent exister pour un même jeu.

Pour tout jeu de forme normale (ou stochastique) à somme nulle composé de deux agents, une *solution minimax* existe. Cette dernière consiste en une *meilleure réponse mutuelle* $\left(\pi_i, \pi_j\right)$ telle que $\pi_i \in BR_i\left(\pi_j\right)$ et $\pi_j \in BR_j\left(\pi_i\right)$.

Définition 7: Solution minimax

Dans un jeu de forme normale à somme nulle avec deux agents i et j, une politique jointe π correspond à une solution minimax si :

$$u_i(\pi) = \max_{\pi_i'} \min_{\pi_j'} u_i\left(\pi_i', \pi_j'\right) = \min_{\pi_j'} \max_{\pi_i'} u_i\left(\pi_i', \pi_j'\right) = -u_j(\pi) \tag{2.8}$$

Plusieurs solutions minimax peuvent exister, mais par définition, toutes rapportent les mêmes retours $u_i(\pi)$ et $u_j(\pi)$. Pour les jeux à somme nulle, des valeurs minimax peuvent également convenir lorsque le nombre d'agents excède deux. Abordons maintenant les concepts impliquant $n > 2$ agents.

Le concept d'*équilibre de Nash* repose sur l'idée d'une *meilleure réponse mutuelle* $\pi_i \in BR_i\left(\pi_{\bar{\imath}}\right)\ \forall i \in \mathcal{I}$ pour un jeu à somme générale, non répété et de forme normale [Kreps, 1989].

Définition 8: Équilibre de Nash

Dans un jeu à somme générale, une politique jointe π correspond à un

équilibre de Nash si :

$$\forall i, \pi_i' : u_i\left(\left\langle \pi_i', \pi_{\bar{\imath}} \right\rangle\right) \leq u_i(\pi) \tag{2.9}$$

Un équilibre ne sous-entend aucunement que les agents impliqués reçoivent la même quantité de récompenses. Tout jeu fini de forme normale comporte au minimum un équilibre de Nash [Nash Jr, 1950].

L'équilibre est dit *pur* si les agents agissent selon une politique *déterministe* $\exists a_i \in \mathcal{A}_i : \pi_i(a_i) = 1$. À l'inverse, l'équilibre est dit *mixte* si les agents agissent selon une politique *stochastique* $\exists a_i \in \mathcal{A}_i : \pi_i(a_i) < 1$. Pour les jeux stochastiques, l'existence d'un équilibre de Nash s'avère également possible sous certaines conditions [Takahashi, 1964]. Normalement, un équilibre peut être atteint si γ est proche de 1.

Le concept d'équilibre de Nash est un concept strict où aucun agent ne peut accroître son retour par une déviation unilatérale de l'équilibre. Il existe toutefois une version moins restrictive nommée *équilibre ϵ-Nash*.

Définition 9: Équilibre ϵ-Nash

Dans un jeu à somme générale composé de $|\mathcal{I}|$ agents, une politique jointe π est un équilibre de ϵ-Nash pour $\epsilon > 0$ si :

$$\forall i, \pi_i' : u_i\left(\left\langle \pi_i', \pi_{\bar{\imath}} \right\rangle\right) - \epsilon \leq u_i(\pi) \tag{2.10}$$

L'équilibre de Nash correspond au cas particulier $\epsilon = 0$.

Jusqu'ici, les concepts de solution supposent que les politiques des agents se caractérisent par leur *indépendance probabiliste*. Cette hypothèse peut alors contraindre le retour espéré des agents. L'*équilibre corrélé* se libère de cette contrainte en autorisant des corrélations entre les politiques [Aumann, 1987].

Définition 10: Équilibre corrélé

Dans un jeu à somme générale de forme normale composé de $|\mathcal{I}|$ agents, une politique jointe $\pi_c(a)$ assigne des probabilités aux actions $a \in \mathcal{A}$. La politique jointe est un équilibre corrélé si :

$$\forall i, \xi_i : \sum_{a \in \mathcal{A}} \pi_c(a)\mathcal{R}(\langle \xi_i(a_i), a_{\bar{\imath}} \rangle) \leq \sum_{a \in \mathcal{A}} \pi_c(a)\mathcal{R}(a) \tag{2.11}$$

où $\xi_i : \mathcal{A}_i \to \mathcal{A}_i$ est un modificateur d'actions.

Cet équilibre suppose que chaque agent connaisse la distribution de probabilités $\pi_c(a)$ ainsi que sa propre action recommandée a_i. Les solutions d'équilibre comportent néanmoins certaines limitations :

- Un équilibre ne garantit pas une maximisation du retour espéré, mais seulement que chaque agent répond de la meilleure manière, sachant la politique conjointe.
- Plusieurs équilibres sont parfois possibles, chacun pouvant comporter des retours espérés différents. Mais, dans ce cas, comment les agents s'accordent-ils sur l'équilibre à adopter ?
- Un équilibre est un concept incomplet dans le sens où il ne spécifie pas les comportements à adopter en dehors de cet équilibre.

Afin de réduire la taille de l'espace des équilibres possibles, des critères peuvent être appliqués. L'*optimum de Pareto* est le premier critère que nous abordons [Stiglitz, 1981].

Définition 11: Optimum de Pareto

Une politique jointe π est Pareto dominée par une autre π' si :

$$\forall i : u_i\left(\pi'\right) \geq u_i(\pi) \text{ et } \exists i : u_i\left(\pi'\right) > u_i(\pi) \tag{2.12}$$

Une politique jointe π est *Pareto optimale* si elle n'est pas Pareto dominée par une autre π'.

Un optimum de Pareto signifie donc qu'aucun agent ne peut espérer obtenir plus sans occasionner de perte aux autres. Tout jeu possède au moins un optimum de Pareto.

Si l'objectif des agents consiste avant tout à optimiser la somme cumulée des retours, le concept du *bien-être social* peut préciser la solution.

Définition 12: Optimum de bien-être social

Le bien-être social d'une politique jointe π est défini par :

$$W(\pi) = \sum_{i \in \mathcal{I}} u_i(\pi) \tag{2.13}$$

Son optimum est atteint si $\pi \in \operatorname{argmax}_{\pi'} W\left(\pi'\right)$.

Si l'objectif des agents consiste à promouvoir l'équité, le concept de *justice sociale* peut s'appliquer.

Définition 13: Optimum de justice sociale

La justice sociale d'une politique jointe π est définie par :

$$W(\pi) = \prod_{i \in \mathcal{I}} u_i(\pi) \tag{2.14}$$

Son optimum est atteint si $\pi \in \operatorname{argmax}_{\pi'} F\left(\pi'\right)$. Attention, le symbole $\prod$

dénote ici le produit et non un ensemble de politiques.

Enfin, si, contrairement aux équilibres étudiés jusqu'à présent, on souhaite se libérer du principe de meilleures réponses mutuelles, alors on peut intégrer le concept du *regret* [Greenwald et Jafari, 2003]. Le regret mesure la différence entre les récompenses obtenues par une politique π_i et les récompenses supposées que rapporterait une politique alternative π_i'. Jouer *sans regret* signifie que pour un ensemble d'*épisodes* $e = 1, \ldots, z$, le regret d'un agent tend vers zéro lorsque $z \to \infty$:

$$\text{Regret}_i^z = \max_{\pi_i \in \Pi_i} \sum_{e=1}^{z} \left[u_i \left(\langle \pi_i, \pi_i^e \rangle \right) - u_i \left(\pi^e \right) \right] \tag{2.15}$$

Définition 14: Sans regret

Dans un jeu à somme générale composé de $|\mathcal{I}|$ agents, les agents sont sans regret si :

$$\forall i : \lim_{z \to \infty} \frac{1}{z} \text{Regret}_i^z \leq 0 \tag{2.16}$$

En résumé, fondé sur la théorie des jeux, un problème MARL se définit par un couple composé d'un modèle de jeu et d'un concept de solution. Le modèle de jeu décrit les interactions possibles, soit les règles du jeu. Le concept de solution précise les états du jeu qui forment un équilibre entre les gains accumulés par les agents. Dans ces équilibres, aucun agent ne peut unilatéralement changer de politique sans altérer les récompenses des autres. En raison de la taille parfois infinie des espaces d'équilibre, il est possible d'ajouter des critères à ces derniers afin d'en restreindre leur nombre. Pour les jeux à somme générale de forme normale, il n'existe aucun algorithme connu capable de résoudre un équilibre en un temps polynomial. Ce dernier point concerne *de facto* les algorithmes MARL.

Sans moyen de communication ou de coordination interagent, la sélection d'un équilibre s'avère complexe et les agents convergent parfois vers un équilibre sous-optimal. Si les politiques paraissent incertaines, les agents peuvent converger vers un équilibre à *dominance de risque* garantissant la plus haute récompense minimale, plutôt que vers un équilibre à *dominance de récompense* rapportant plus de récompenses qu'aucun autre. De plus, si différents équilibres rapportent différentes récompenses aux agents, des conflits surviennent quant à la sélection d'un équilibre et empêchent parfois toute convergence.

Si les modèles de jeux peuvent représenter les situations réelles comme le trafic, il en est autrement pour les concepts de solution. En effet, dans le cadre du trafic, les interactions dépendent en majeure partie de l'accès à une ressource limitée : l'espace roulable. Suivant l'offre et la demande pour cette ressource,

les conducteurs peuvent devenir plus ou moins compétitifs – offre inférieure à la demande – ou coopératifs – offre supérieure à la demande. L'hétérogénéité comportementale nous posera un véritable défi lors de la phase de conception des fonctions de récompense, qui, rappelons-le, sous-tendent les comportements (solutions) attendus.

2.2 Concevoir une solution

Maintenant que nous savons définir un problème MARL, voyons comment le résoudre. Afin d'éviter l'écueil de la dimensionnalité de l'espace d'état-action, nous constaterons de la nécessité d'adopter une méthode de généralisation (2.2.1). Nous découvrirons que les *réseaux de neurones* disposent de cette qualité et nous décrirons leurs différentes architectures ainsi que leur fonctionnement. Ensuite, nous listerons les différentes catégories d'algorithmes de RL (2.2.2). Nous détaillerons pour chacune ses avantages et ses inconvénients. Enfin, nous distinguerons les différents modes d'entraînement et d'exécution, leurs implications et leurs limites (2.2.3).

2.2.1 Concevoir un réseau de neurones

La complexité d'un problème MARL dépend majoritairement de la taille des espaces d'agents, d'états, d'actions et d'observations. Lorsque leur taille est relativement petite, comme c'est le cas pour le jeu *Pierre-feuille-ciseau*[1] à deux joueurs, il devient possible, pour les agents, de construire une politique en sauvegardant dans un tableau l'estimation des fonctions de valeur d'état ou de qualité d'action.

En pratique, rares sont les environnements aussi simplistes. La majorité d'entre-eux représentent des espaces d'états ou d'actions de manière continue, rendant l'*apprentissage tabulaire* caduc. Il n'est alors plus possible d'estimer précisément les récompenses futures. Ces dernières doivent donc être approximées par des fonctions.

Une *fonction d'approximation* doit être capable de cartographier les probabilités de transition $\mathcal{T}$ d'un environnement ou d'apprendre directement une politique π. Puisqu'aucune hypothèse ne peut être avancée *a priori*, la fonction d'approximation doit disposer d'un *biais inductif* quasi nul. Un biais inductif constitue l'ensemble des hypothèses utiles à la prédiction d'une sortie compte tenu des entrées qui n'ont pas encore été rencontrées [Mitchell, 1980]. Le nombre d'entrées et de sorties de la fonction d'approximation doit être paramétrable afin de s'adapter à la dimensionnalité des observations et des actions de différents jeux. Enfin, la fonction doit disposer d'une *capacité de généralisation* pour conserver les informations apprises sans garder en mémoire les *expériences* passées.

1. https://fr.wikipedia.org/wiki/Pierre-feuille-ciseaux

Les réseaux de neurones remplissent l'ensemble de ces critères, ce qui en fait une approche populaire dans le domaine de l'apprentissage machine. Nous utiliserons donc un réseau de neurones comme fonction d'approximation $f(x;\theta)$ où $x \in \mathbb{R}^{d_x}$ correspond à une entrée de dimension d_x et θ aux *paramètres entraînables* du réseau. Apprendre une fonction d'approximation implique un processus d'optimisation des paramètres θ dont l'objectif est d'approximer une *fonction cible* f^*. En apprentissage par renforcement, la fonction cible peut être la fonction de valeur d'état $V(s)$ représentée linéairement :

$$\hat{V}(s;\theta) = \theta^\intercal x(s) = \sum_{k=1}^{d} \theta_k x_k(s) \tag{2.17}$$

avec les vecteurs de paramètres $\theta \in \mathbb{R}^d$ et de représentation d'états $x(s) \in \mathbb{R}^d$.

L'apprentissage consiste à approximer la fonction $V(s)$. Une *fonction de perte* $\mathcal{L}(\theta)$ calcule la distance qui nous sépare de cet objectif, à savoir la distance entre l'approximation $\hat{V}(s;\theta)$ et la valeur réelle $V^\pi(s)$. Les *paramètres optimaux* θ^* sont obtenus par la *minimisation* de cette perte :

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\, \mathcal{L}(\theta) = \underset{\theta}{\operatorname{argmin}}\, \mathbb{E}_{s\in\mathcal{S}}\left[\left(V^\pi(s) - \hat{V}(s;\theta)\right)^2\right] \tag{2.18}$$

Puisque, contrairement aux techniques d'apprentissage supervisé, nous ne disposons pas de la valeur réelle $V^\pi(s)$, cette dernière est aussi approximée tout au long de l'apprentissage, telle que $V^\pi(s) \sim r_i + \gamma \hat{V}\left(s'_i;\theta\right)$.

Le processus d'*optimisation* s'effectue par *descente de gradient* $\nabla_\theta \mathcal{L}(\theta)$ [Amari, 1993]. Souvent, il s'agit d'une *descente de gradient par mini lots* $\mathcal{B}$ de tailles $|\mathcal{B}|$ composés de données d échantillonnées depuis un *ensemble d'entraînement* $\mathcal{D}$ avec une loi de probabilité $\mathcal{U}$:

$$\theta \leftarrow \theta - \alpha \nabla_\theta\, \mathcal{L}(\theta \mid \mathcal{B})\big|_{\mathcal{B}=\{d_i\sim\mathcal{U}(\mathcal{D})\}_{i=1}^{|\mathcal{B}|}} \tag{2.19}$$

La mise à jour des paramètres s'effectue progressivement selon un *taux d'apprentissage* α, dont la valeur est proche de zéro afin de ne pas *écraser* ce qui a préalablement été appris. L'optimisation est répétée jusqu'à vérifier l'un des critères suivants, définis par le concepteur :

- la perte $\mathcal{L}(\theta)$ passe sous un certain seuil ϵ
- l'algorithme converge, *i.e.*, les paramètres θ ne varient presque plus
- après N itérations

Plusieurs types de réseaux de neurones existent. Commençons par analyser la structure du *perceptron multicouche (MLP)*, également appelé *réseau de neurones entièrement connecté* [Popescu *et al.*, 2009]. Un MLP se compose de plusieurs *couches*, chacune constituée d'un ensemble de *neurones*. L'unité de base, le neurone, prend un ensemble d'entrées x et produit une sortie y. Un neurone possède autant

de poids $w \in \mathbb{R}^{d_x}$ qu'il compte d'entrée x. La sortie y scalaire produite résulte d'une multiplication vectorielle des entrées x et de leurs *pondérations* x à laquelle s'ajoute un *biais* scalaire $b \in \mathbb{R}$, le tout envoyé à une *fonction d'activation* $\mathcal{H} : \mathbb{R} \to \mathbb{R}$, tel que $y = \mathcal{H}\left(w^\intercal x + b\right)$. Les fonctions d'activation reconnues empiriquement pour leur efficacité sont :

- La *fonction unité linéaire rectifiée* $\text{ReLU}(x) = \max(0, x)$
- La *tangente hyperbolique* $\tanh(x) = \frac{\exp^x - \exp^{-x}}{\exp^x + \exp^{-x}}$
- La *sigmoïde* ou *fonction logistique* $\sigma(x) = \frac{1}{1+\exp^{-x}}$

La tangente hyperbolique et la sigmoïde permettent de conditionner la sortie y sur les intervalles $[-1;1]$ et $[0;1]$ respectivement. Ces fonctions s'avèrent donc utiles pour produire des sorties sur des intervalles fermés, comme une probabilité d'effectuer une action. En pratique, la fonction ReLU est préférée pour les couches intermédiaires du réseau.

Chaque couche d'un MLP traite les entrées qu'elle reçoit et propage les sorties aux couches suivantes. Grâce aux capacités de parallélisation des processeurs graphiques modernes, la propagation peut s'effectuer en temps $O(1)$ selon le nombre de neurones par couche. L'apprentissage des paramètres θ d'un MLP s'effectue par un algorithme de *rétropropagation* dont nous passerons les détails afin de ne pas alourdir cette section. Un MLP de k couches compte $\theta = \bigcup_k \theta^k$ paramètres où θ^k correspond au nombre de paramètres de la couche k. Chaque couche supplémentaire ajoute de la profondeur au réseau de neurones, on parle de *réseau de neurones profonds*.

Concevoir un réseau de neurones est simple. Concevoir un réseau adapté au problème traité l'est moins. Plus le réseau comporte de neurones, plus il peut modéliser des structures de données complexes. Toutefois, si la taille du réseau est disproportionnée par rapport à la complexité du problème, il existe un risque de *surajustement* [Hawkins, 2004] (figure 2.3). Le réseau modélisera si parfaitement les données d'apprentissage qu'il en perdra toute capacité de généralisation aux nouvelles données. Si la taille du réseau est sous-dimensionnée par rapport au problème abordé, il existe un risque de *sous-ajustement*. Dans ce cas de figure, le réseau échoue à modéliser les données d'apprentissage comme les nouvelles données.

Tant le surajustement que le sous-ajustement engendrent un biais inductif élevé et accroissent la *variance*, *i.e.*, la sensibilité du modèle aux variations dans les données d'entraînement. Malheureusement, aucune méthode ne permet de déterminer l'architecture optimale d'un réseau selon un problème donné. Ainsi, la majorité des concepteurs s'approprient les architectures dont l'efficacité est prouvée empiriquement sur une large gamme de problèmes.

Au-delà du biais inductif, le besoin d'ajustement incrémental des paramètres du réseau de neurones freine la convergence [Botvinick *et al.*, 2019]. Lorsque les paramètres d'un réseau de neurones sont mis à jour de manière trop brusque

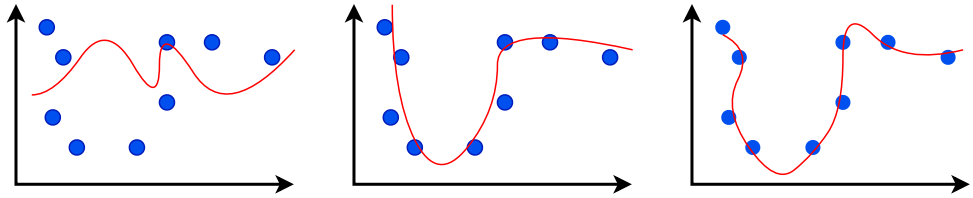


Figure 2.3 – Conséquence d'un sous-ajustement (gauche), d'un ajustement équilibré (milieu), d'un surajustement (droite) sur les données (points). Schéma inspiré de Badillo *et al.* [2020]

ou non incrémentale, l'apprentissage devient instable et cette instabilité altère la convergence. Ainsi, selon les auteurs, *pour apprendre rapidement, un réseau de neurones doit apprendre lentement*, *i.e.*, ses paramètres requièrent un ajustement progressif afin de ne pas effacer les informations préalablement apprises.

D'autres architectures de réseau de neurones existent. Parmi ces architectures, nous examinons les *réseaux de neurones récurrents (RNNs)*, en raison de leur spécialisation dans le traitement des séquences [Salehinejad *et al.*, 2017]. Les RNNs constituent une approche efficace pour la résolution de problèmes partiellement observables où l'exploitation de l'historique d'observations permet d'accéder à des informations cruciales pour le jeu à résoudre. Un RNN apprend les corrélations contenues dans les séquences temporelles. L'architecture dispose d'un *état caché* h, une représentation compacte sauvegardant les informations contenues dans les entrées préalablement traitées. À l'initialisation, l'*état caché initial* h^0 se constitue d'un vecteur rempli de zéros. La valeur de l'état caché h^t s'actualise à chaque instant t en tenant compte de l'entrée actuelle x^t et de la valeur courante de l'état caché h^{t-1}, tel que :

$$h^t = f\left(x^t, h^{t-1}; \theta\right) \tag{2.20}$$

L'apprentissage d'un RNN représente toutefois un défi de taille. La répétition d'étapes de rétropropagation, qui multiplie récursivement des dérivées, conduit souvent à une *disparition* ou à une *explosion* du gradient [Hanin, 2018]. Ces phénomènes s'observent également lorsqu'un MLP comporte trop de couches cachées. L'explosion du gradient trouve son origine dans la multiplication répétée de nombres supérieurs à 1. Le résultat tend alors vers l'infini. Pour remédier à ce problème, un concepteur doit tronquer le gradient pour le maintenir sous un certain seuil ou appliquer une *régularisation pondérée*.

La disparition du gradient découle du phénomène inverse, soit la multiplication répétée de nombres proches de 0. Le résultat tend alors vers l'infinitésimal. Ce problème s'atténue par l'utilisation de fonctions d'activation ReLu et par une initialisation appropriée des paramètres θ.

La disparition et l'explosion du gradient ont conduit les chercheurs à proposer

d'autres architectures RNNs plus performantes. Deux d'entre-elles, LSTM [2] et GRU [3], ont permis d'obtenir une meilleure rétention des informations à travers le temps en apprenant à mémoriser les informations utiles et à oublier celles qui ne le sont plus [Hochreiter et Schmidhuber, 1997; Cho *et al.*, 2014]. Ces deux architectures se distinguent par leur structure. Dans un LSTM, trois portes régulent le flux d'information (entrée, oubli et sortie) ainsi qu'une cellule de mémoire distincte. Dans un GRU, la structure se simplifie en deux portes (mise à jour et réinitialisation) sans cellule de mémoire distincte. Le tableau 2.1 synthétise les différences entre les architectures récurrentes.

Table 2.1 – Comparatif des réseaux de neurones à mémoire

	RNN	LSTM	GRU
Nombre de paramètres	faible	élevé	moyen
Performance séquences courtes	bonne	excellente	excellente
Performance séquences longues	faible	excellente	très bonne

Tout réseau de neurones, récurrent ou non, est sensible à la *distribution des données d'entraînement*. Lorsqu'il existe une forte corrélation entre les données d'entraînement, le réseau peut surajuster ses paramètres selon cette distribution. Le modèle se spécialisera alors dans la résolution d'un sous-jeu, au détriment de la résolution du jeu principal. Le réseau perd sa capacité de généralisation. Pour pallier ce problème, les données d'entraînement doivent respecter la propriété suivante.

Propriété 2: Variables indépendentes et identiquement distribuées

Des variables aléatoires $X_1, \ldots, X_n$ sont indépendantes si, pour tout réel $t_{i_1}, \ldots, t_{i_k}$:

$$\mathbb{P}\left[X_{i_1} \leq t_{i_1}, \ldots, X_{i_n} \leq t_{i_n}\right] = \mathbb{P}\left[X_{i_1} \leq t_{i_1}\right] \times \ldots \times \mathbb{P}\left[X_{i_n} \leq t_{i_n}\right] \quad (2.21)$$

et identiquement distribuées si $\forall i = 1, \ldots, n$:

$$\mathbb{P}\left[X_{i_i} \leq t_{i_i}\right] \leq \mathbb{P}\left[X_{i_1} \leq t_{i_1}\right] \quad (2.22)$$

Pour résumer, aucune méthode sûre ne permet de concevoir une architecture neuronale sur mesure au problème abordé. Les avantages apportés par les réseaux de neurones récurrents, notamment leur traitement séquentiel des données, en font de bons candidats pour les applications de RL. De plus, les architectures récurrentes se rapprochent des fondements théoriques qui stipulent que les politiques apprises se fondent sur un historique d'observations. Ce dernier étant naturellement inclus

2. Long-short time memory
3. Gated recurrent unit

dans les architectures récurrentes. L'architecture GRU, en particulier, semble offrir un compromis entre un faible nombre de paramètres et une bonne rétention d'informations sur les séquences à court terme, d'une durée équivalente à celle d'une interaction entre deux véhicules.

2.2.2 Choisir un algorithme

De nombreux algorithmes de RL existent. Afin de saisir leurs nuances, décrivons les catégories auxquelles ils appartiennent. Tout algorithme $\mathbb{L}$ de RL appartient nécessairement à trois catégories selon : (1) l'objet de la modélisation, (2) la méthode d'apprentissage de la politique et (3) la provenance des données d'entraînement.

L'objet de la modélisation peut être l'environnement ou la politique. Les approches *fondées sur un modèle* cherchent à apprendre la fonction de transition $\mathcal{T}$ [Moerland *et al.*, 2023]. Puisque cette fonction décrit les dynamiques de l'environnement, l'apprendre permet d'interagir de manière optimale en prédisant efficacement les états et les récompenses futurs atteignables en suivant une séquence d'actions précise (figure 2.4). Cependant, selon la taille de l'espace d'état-action, apprendre la fonction de transition peut s'avérer complexe, voire impossible. Ce, d'autant plus que la modélisation de $\mathcal{T}$ suppose que l'agent ait assez exploré l'environnement pour se le représenter précisément. De plus, la modélisation d'un environnement dans sa globalité nécessite plus d'espace mémoire. On peut alors légitimement se demander si conserver une représentation globale de l'environnement a un intérêt, sachant que seule une partie pourrait effectivement être utile à l'agent. Les algorithmes *sans modèle* répondent par la négative à cette question [Degris *et al.*, 2012]. Ces algorithmes modélisent donc directement une politique, qui, pour simplifier, correspond à un sous-ensemble de la fonction de transition qui évolue au gré des variations de la politique π durant la phase d'apprentissage.

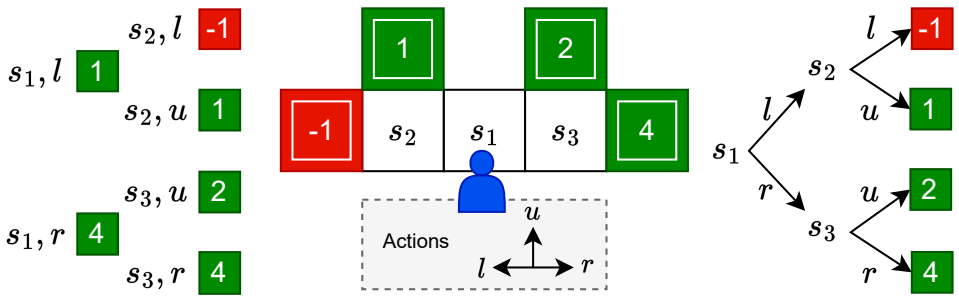


Figure 2.4 – Approches fondée politique (gauche) et fondée modèle (droite). Les cellules colorées dénotent les états finaux de l'environnement et leur récompense respective

La méthode d'apprentissage de la politique peut soit opter pour un apprentissage des fonctions de valeurs desquelles on dérive une politique, soit directement

apprendre cette dernière. Les approches *fondées sur une valeur* apprennent implicitement la politique par la sélection de l'action la plus prometteuse et en terme de gain $\pi_i = \text{argmax}_a Q(s,a)$ [McKenzie et McDonnell, 2022]. Toutefois, la dimensionnalité de l'espace d'état-action peut complexifier son approximation [Liu *et al.*, 2020]. Il peut alors être opportun de considérer une approche apprenant explicitement la politique. Les approches *fondées sur une politique* apprennent les probabilités de distribution des actions $\pi : \mathcal{S} \rightarrow \mathcal{A}$ [Sewak, 2019]. Ces méthodes souffrent cependant d'une variance élevée qui diminue le biais inductif et donc ralentit le processus d'apprentissage. Cette variance provient de la représentation de la politique qui n'indique pas les contributions respectives de chaque action au retour espéré. Afin de surmonter les défis posés par ces deux méthodes, des chercheurs ont proposé de les coupler [Grondman *et al.*, 2012]. Ces algorithmes, dits *acteur-critique*, se composent d'un critique estimant la fonction de valeur $Q(s,a)$ et d'un acteur apprenant une politique $\pi(s)$. Le critique sert de guide à l'acteur : grâce à son évaluation de la fonction de valeur, il évalue les décisions de l'acteur (figure 2.5). Une fois la politique apprise, le critique n'est plus nécessaire, seul l'acteur est conservé.

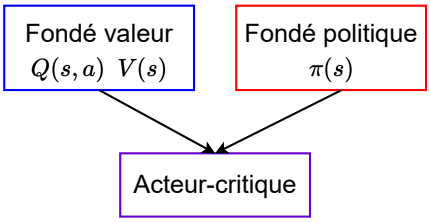


Figure 2.5 – Acteur-critique

La provenance des données d'entraînement caractérise également les algorithmes de RL [Hausknecht *et al.*, 2016]. Certains algorithmes sont conçus pour apprendre exclusivement à partir de données récentes issues de leur politique actuelle $\pi(s \mid \theta)$ [Andrychowicz *et al.*, 2021]. Ainsi, les approches *in-politique* sont réservées aux problèmes où la génération de données d'apprentissage est peu coûteuse (figure 2.6). Au contraire, les approches *hors politique* permettent d'apprendre avec des données récentes comme anciennes [Munos *et al.*, 2016]. Dans cette figure 2.6, l'approche hors politique maintient une mémoire tampon contenant les expériences collectées à différents moments de l'apprentissage, tandis que l'approche in-politique apprend uniquement à partir de ces expériences récentes.

Cette dernière catégorie – la provenance des données d'entraînement – inclut les approches de *RL inverse* et de *RL par imitation*, dont les objectifs respectifs reposent sur la déduction de la fonction de récompense $\mathcal{R}$ qui sous-tend un comportement et la reproduction d'un comportement par imitation d'une politique. La première approche requiert cependant de fortes hypothèses, incluant la linéarité de la fonction

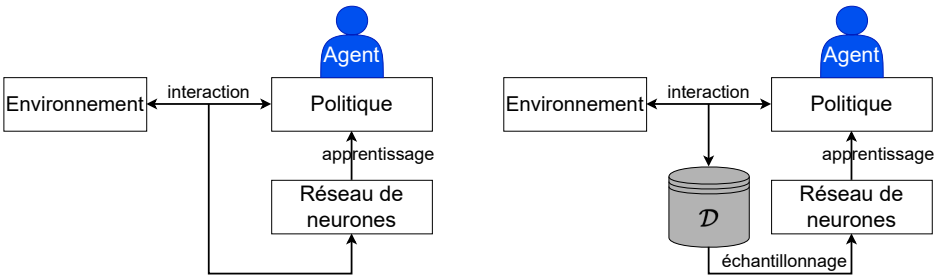


Figure 2.6 – Approches in-politique (gauche) et hors politique (droite).

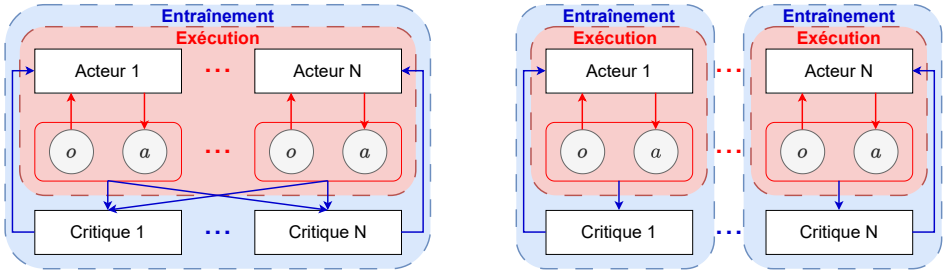
de récompense et une connaissance de la dynamique de l'environnement. La seconde nécessite que les comportements observés soient exemplaires, sinon, les agents pourraient répliquer les erreurs des conducteurs.

Le choix de l'algorithme dépend donc du problème. Dans notre cas, nous souhaitons apprendre à des robots de conduite à naviguer dans un trafic hétérogène. Compte tenu du nombre de véhicules qui composent le trafic et de la diversité de situations qui peuvent émerger du trafic, nous opterons pour un algorithme qui modélise directement la politique. Nous préférons un algorithme acteur-critique du fait des avantages évoqués quant à la résolution de problèmes du monde réel. Sachant qu'acquérir les données peut prendre un certain temps dans un environnement multiagent, nous préférons les approches hors politique. Enfin, nous renonçons aux approches fondées sur une imitation de politique dont les données proviendraient de démonstrations réelles de conducteurs humains. La raison principale étant que les données récoltées sont soit trop peu nombreuses pour que l'agent puisse apprendre à généraliser, soit manquent d'informations cruciales pour comprendre les décisions (motivations, contexte, *etc.*)

Pour résumer, nous souhaitons un algorithme acteur-critique qui apprenne une politique à partir de données générées hors politique. Nous détaillerons dans le prochain chapitre l'algorithme, noté $\mathbb{L}$, retenu. Pour le moment, concentrons-nous sur les différentes méthodes d'entraînement d'un système multiagent.

2.2.3 Modes d'entraînement et d'exécution

Tout algorithme d'apprentissage comporte deux phases : l'*entraînement* et l'*exécution*. Les systèmes multiagents peuvent adopter un mode *centralisé* ou *décentralisé* pour ces deux phases (figure 2.7). La centralisation de l'entraînement permet aux agents d'accéder à des *informations supplémentaires* par rapport à un mode décentralisé où chaque agent se contente de son *observation locale*. Le principe est globalement identique pour le mode d'exécution. Trois combinaisons de modes existent : entièrement centralisé, entièrement décentralisé, et apprentissage centralisé et exécution décentralisée.



(a) Centralisation **(b)** Décentralisation

Figure 2.7 – Modes centralisé (a) et décentralisé (b) pour un algorithme acteur-critique. Les éléments (bleu) de la phase d'entraînement sont abandonnés à l'exécution (rouge). o et a correspondent respectivement aux observations et actions.

Le mode *entièrement centralisé* utilise un mécanisme de partage d'informations interagent [Zhang *et al.*, 2021]. Cette information peut être un historique d'observations, une politique ou une modélisation de l'environnement. La centralisation vise à améliorer les politiques apprises en apportant des informations additionnelles qui atténuent l'observabilité partielle.

Si l'algorithme dispose de l'historique joint des observations $h^t = \left(h_1^t, \ldots, h_{|\mathcal{I}|}^t\right)$, le problème MARL se simplifie en un problème RL (agent unique) appelé *apprentissage central*. Cette approche convient aux problèmes nécessitant une coopération entre les agents. Elle se révèle néanmoins parfois impossible à mettre en place, car : (1) la récompense jointe $r = (r_1, \ldots, r_{|\mathcal{I}|})$ doit être scalarisée pour être apprise, (2) l'espace d'action joint $\mathcal{A}$ croît exponentiellement avec le nombre d'agents $|\mathcal{I}|$, et (3) centraliser autant d'informations prend trop de temps ou est matériellement impossible. Une approche décentralisée est alors envisagée.

Le mode *entièrement décentralisé* (figure 2.7b) n'autorise aucun partage centralisé d'informations [Gronauer et Diepold, 2022; Oroojlooy et Hajinezhad, 2023]. Ceci n'exclut en rien les processus d'influence et de communications directes ou indirectes interagents. Le trafic peut être considéré comme entièrement décentralisé : les conducteurs peuvent se coordonner et communiquer par des signaux (clignotants), mais n'accèdent pas aux observations, motivations et connaissances des autres usagers.

Ce mode peut aussi réduire un problème MARL en un problème RL appelé *apprentissage indépendant*. Dans ce cas, chaque agent dispose d'informations locales et ignore l'existence des autres. Les observations sont dépourvues d'informations directes sur les autres agents. Toutefois, leurs actions peuvent influencer la dynamique de l'environnement $\mathcal{T}$, et c'est seulement à travers celle-ci qu'un

agent i peut se représenter les autres $\bar{\imath}$:

$$\mathcal{T}_i\left(s^{t+1} \,\middle|\, s^t, a_i^t\right) \propto \underbrace{\sum_{a_{\bar{\imath}} \in \mathcal{A}_{\bar{\imath}}} \mathcal{T}\left(s^{t+1} \,\middle|\, s^t, \langle a_i^t, a_{\bar{\imath}} \rangle\right)}_{\text{dynamique environnementale}} \underbrace{\prod_{j \in \bar{\imath}} \pi_j\left(a_j \,\middle|\, s^t\right)}_{\text{influence des autres}} \tag{2.23}$$

La décentralisation évite naturellement les problèmes liés à la dimensionnalité de l'espace. Néanmoins, l'impossibilité de modéliser les autres agents peut altérer la convergence, car leur influence sur l'environnement ne peut être différenciée de sa stochasticité naturelle (éq. 2.23). Par ailleurs, Gupta *et al.* [2017] ont montré l'inefficacité de la décentralisation même pour un faible nombre d'agents. Compte tenu des limitations des approches centralisées et décentralisées, une combinaison de modes fut proposée (figure 2.7a).

Cette dernière consiste à *centraliser l'entraînement et à décentraliser l'exécution* [Lowe *et al.*, 2017]. Ainsi, les agents disposent d'informations additionnelles leur permettant d'apprendre des politiques plus robustes durant l'entraînement. Une fois leur politique apprise, les agents ne bénéficient plus de cette centralisation [Canese *et al.*, 2021; Schmidt *et al.*, 2022]. Afin de permettre une exécution décentralisée, la politique des agents ne doit reposer que sur des observations locales. Il s'agit d'une approche populaire dans le domaine MARL, car elle permet entre autres de modéliser la politique des autres agents et ainsi d'accélérer l'entraînement.

En résumé, la centralisation conjugue les informations individuelles pour bonifier les politiques apprises. Au contraire, la décentralisation rend indépendant les processus d'apprentissage des agents. Ces approches antagonistes souffrent toutes deux de problèmes de dimensionnalité. Le compromis de l'entraînement centralisé, d'exécution décentralisée, tire bénéfice des deux approches précédentes, tout en atténuant les problèmes de dimensionnalité, sans toutefois les éliminer totalement. Ce dernier s'impose donc comme une réponse possible à notre problématique.

2.3 Identifier les écueils

Précédemment, nous avons appris à définir un problème MARL et à identifier les éléments nécessaires à sa résolution. Nous allons maintenant nous pencher sur les défis inhérents au domaine MARL et examiner les approches qui permettent d'en atténuer les effets. Nous restreignons notre examen aux deux principaux écueils, à savoir : la non-stationnarité (2.3.1) et la dimensionnalité (2.3.2).

2.3.1 Non-stationnarité

La *non-stationnarité* constitue le problème central des approches MARL du fait de ses implications [Hernandez-Leal *et al.*, 2017; Guessoum, 2004]. Ce phénomène naît de l'apprentissage conjoint de politiques par les agents. L'objectif de tout

algorithme MARL consiste à apprendre une politique optimale fondée sur un apprentissage direct ou indirect des dynamiques de l'environnement $\mathcal{T}$. Une telle politique ne peut exister qu'au travers d'une approximation fiable des lois de probabilité qui la sous-tendent, comme celle de la dynamique $\mathcal{T}$ de l'environnement. Or, ces lois évoluent continuellement en raison de l'*adaptation conjointe* des agents due à l'apprentissage des politiques de chacun. Ainsi, du point de vue d'un agent, l'environnement est imprévisible, non stationnaire.

Ce phénomène est particulièrement notable au commencement de la phase d'apprentissage. À ce moment, la politique jointe résulte d'interactions purement aléatoires en raison d'une méconnaissance totale de l'environnement par les agents. On parle de *bruit exploratoire d'apprentissage*. Bien que la non-stationnarité puisse parfois s'atténuer une fois la *phase exploratoire* terminée, dans certains cas, elle perdure sous forme cyclique. Imaginez deux personnes jouant à *Pierre, feuille, ciseaux* et s'adaptant indéfiniment aux stratégies de l'une et de l'autre. Dans le pire des cas, la non-stationnarité condamne toute convergence vers une solution optimale.

Avant d'aller plus loin, définissons mathématiquement la *stationnarité*.

Définition 15: Stationnarité

Un processus stochastique $\left\{X^t\right\}_{t\in\mathbb{N}^0}$ est dit stationnaire si la probabilité de distribution X^{t+k} est indépendante de $k \in \mathbb{N}^0$ avec t, k des indices temporels.

Cette définition fait écho à la propriété de Markov définie en amont (voir éq. 2.1). Sachant que la phase d'apprentissage RL repose sur une mise à jour régulière de la politique $\pi^{t+1} = \mathbb{L}\left(d^t, \pi^t\right)$ où d^t dénote les données d'entraînement et $\mathbb{L}$ un algorithme, alors tout processus d'entraînement implique nécessairement de la non-stationnarité puisque la politique dépend de t.

Cette déduction s'applique aussi pour les algorithmes RL où un seul agent apprend. Toutefois, on comprend que ce problème est exacerbé par les approches MARL puisque la non-stationnarité n'est plus le produit d'un unique agent, mais bien de tous les agents apprenants. La différence étant qu'en RL la non-stationnarité se restreint à l'évaluation de la fonction de valeur, tandis qu'en MARL, c'est l'environnement tout entier qui apparaît comme non-stationnaire du point de vue des agents. L'environnement devient alors *non Markovien* en raison du non-respect de la propriété de Markov. En conséquence, les approches MARL ne disposent, à l'heure actuelle et selon nos connaissances, d'aucune *garantie théorique de convergence* pour des jeux à somme générale de forme normale.

La non-stationnarité place les problèmes MARL (POSG à somme générale) dans la catégorie des algorithmes de complexité temporelle NEXP-complet [Alsheikh *et al.*, 2015]. Cependant, sachant que la non-stationnarité résulte des actions des autres agents, modéliser leur influence permet de tempérer cet écueil. Dans

un article de synthèse, Hernandez-Leal *et al.* [2017] identifient cinq moyens d'y parvenir :

- **Ignorer**. La première approche considère les influences marginales de chaque agent comme inhérentes à la dynamique environnementale. Nous l'avons d'ailleurs évoquée plus haut sous l'appellation d'apprentissage indépendant (voir section 2.2.3). Les agents supposent donc que l'environnement est stationnaire.
- **Oublier**. La seconde approche consiste à s'adapter à la non-stationnarité en remplaçant les informations anciennes par de plus récentes. Cette technique s'applique aisément en définissant un taux d'apprentissage α élevé, de manière à pondérer plus fortement les mises à jour récentes par rapport aux informations anciennement acquises.
- **Répondre**. La troisième approche convient aux environnements où les agents interagissent avec un ensemble d'opposants fixes disposant d'un ensemble de politiques connues *a priori*. Dans ce contexte, l'objectif des agents consiste à déterminer la meilleure réponse face aux politiques de leurs opposants.
- **Modéliser**. La quatrième approche revient à modéliser les opposants afin de dériver une politique optimale. Ici, nul besoin d'informations préalables, excepté que l'ensemble des opposants est fixe. La modélisation des opposants peut s'effectuer par une approximation de l'action jointe ou par une reconstruction de leurs politiques.
- **Théorie de l'esprit**. Cette dernière catégorie suppose que chaque agent modélise ses opposants et suppose que ces derniers le modélisent également, entraînant un raisonnement récursif. Les approches de théorie de l'esprit deviennent inenvisageables dès lors que le nombre d'agents dépasse quelques unités, compte tenu de la complexité de tels calculs.

Les trois dernières approches supposent un ensemble d'opposants fixes, ce qui s'avère irréaliste dans le contexte de trafic où les usagers interagissant avec un individu changent continuellement.

En résumé, la non-stationnarité résulte de l'adaptation concurrentielle des agents et altère continuellement les lois de probabilité que les agents essaient d'apprendre. Ce défi prévient toute garantie de convergence pour les problèmes POSG à somme générale.

La modélisation de la politique permet à certains algorithmes de converger malgré la non-stationnarité. Cinq niveaux de modélisation des politiques existent, allant d'ignorer la non-stationnarité jusqu'à bâtir un modèle de théorie de l'esprit. Le dernier niveau, la théorie de l'esprit, a été abordé lors du chapitre précédent. Nous avions constaté que les conducteurs y font rarement appel compte tenu de son coût cognitif. Sachant que les troisième et quatrième niveaux requièrent des interactions prolongées avec un ensemble fixe d'agents pour distinguer leurs

influences de la dynamique de l'environnement (voir éq. 2.23), nous pouvons les exclure.

Par conséquent, seuls les deux premiers niveaux restent envisageables. Le premier, ignorer la non-stationnarité, amplifie les défis posés par cette dernière. Le deuxième, oublier les informations anciennes, nous semble donc être un bon compromis, d'autant plus que son implémentation paraît relativement aisée. Néanmoins, à ce niveau, la non-stationnarité posera assurément des problèmes de convergence.

L'absence d'approches mitigeant les conséquences de la non-stationnarité pour des environnements dont les propriétés s'apparentent à celles du trafic nous posera un problème majeur. Nous traiterons en détail ce problème lors du prochain chapitre. Pour le moment, essayons de comprendre dans quelles proportions la dimensionnalité accentue les problèmes engendrés par la non-stationnarité.

2.3.2 Dimensionnalité

De manière générale, la non-stationnarité, et les problèmes qui en découlent sont exacerbés par la dimensionnalité des différents espaces : d'agents, d'états et d'actions. La *malédiction de la dimensionnalité* concerne principalement le nombre d'agents. Ainsi, l'espace des états croît exponentiellement avec le nombre d'agents $|\mathcal{I}|^2 - |\mathcal{I}|$, puisque les états reçus par chaque agent incluent des informations sur l'ensemble des autres agents. L'observabilité partielle permet d'atténuer ce problème de dimensionnalité. Ainsi, les agents reçoivent des observations (états tronqués) plutôt que la totalité de l'état. Rappelons toutefois que l'observabilité partielle s'effectue au détriment de l'efficacité des politiques apprises.

Contrairement à l'espace d'états, la taille de l'espace des actions ne peut être diminuée. L'espace de l'action jointe, qui englobe les actions de tous les agents, croît donc de manière exponentielle $|\mathcal{A}| = \prod_{i \in \mathcal{I}} |\mathcal{A}_i|$. Néanmoins, dans certains cas, il est possible de faciliter l'apprentissage des agents en masquant certaines actions contre-productives ou impossibles à entreprendre. La mise en œuvre du masquage repose sur le concepteur et ses connaissances *a priori* des situations pouvant en bénéficier. Dans le cas du trafic, le masquage peut s'avérer pertinent dans les situations où certaines actions mèneraient inévitablement à un échec (figure 2.8). Généralement, il s'agit de transgressions du Code de la route. En définitive, le masquage accélère l'apprentissage en évitant à un agent de s'engager dans des actions dont la finalité est vouée à l'échec compte tenu des objectifs de l'agent.

En raison de la taille de l'espace d'état-action, la phase exploratoire, où les agents acquièrent des connaissances sur la dynamique de l'environnement et les politiques de leurs pairs, peut demander plus de temps aux agents. Un compromis entre l'exploration et l'exploitation s'impose pour que les agents convergent. Lorsque la taille de l'espace d'état-action explose, l'exploration devient superficielle et la probabilité de converger vers un optimum local plutôt que global s'accentue. L'efficacité de l'échantillonnage repose essentiellement sur les performances de la

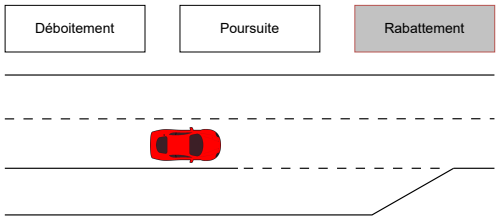


Figure 2.8 – Masquage d'une action contre-productive. Impossible pour le véhicule de sélectionner l'action de rabattement, vers la voie à sa droite, car cette manœuvre contrevient au Code de la route et mène irrémédiablement à un échec.

stratégie exploratoire, d'où son importance particulière dans les environnements où le coût d'interaction contraint les agents à minimiser la durée de la phase exploratoire.

La dimensionnalité se voit également affectée par la nature des agents. La nature de ces derniers, nous l'avons constaté, oscille entre l'homogénéité et l'hétérogénéité. Une population parfaitement homogène possède des caractéristiques, un ensemble d'actions et des fonctions de récompense identiques. Dans cette configuration, il devient plus aisé, pour un agent, de prédire les actions des autres, et la modélisation de leur politique ne nécessite pas une individualisation. À l'inverse, une population hautement hétérogène nécessitera certainement une modélisation individualisée des politiques des autres agents. Par conséquent, l'hétérogénéité accroît significativement les problèmes de dimensionnalité et, par extension, de non-stationnarité. Il s'agit probablement du défi le plus considérable que peuvent rencontrer les agents et leurs concepteurs.

Au-delà de l'hétérogénéité, la nature du jeu accentue les problèmes précédents [Oroojlooy et Hajinezhad, 2023]. Nous en avons déjà partiellement discuté, la nature du jeu oscille d'entièrement coopératif à entièrement compétitif. Les agents coopératifs tendent à converger plus rapidement vers une solution, car leur objectif ne repose pas sur la minimisation des récompenses de leurs pairs, mais, tout au plus, sur une stratégie de coordination. Dans les jeux compétitifs, notamment ceux à somme nulle, la convergence vers un équilibre est prolongée, car les agents doivent constamment s'adapter et contrer les politiques de leurs opposants.

Pour les environnements où l'horizon temporel tend vers l'infini, la convergence est d'autant plus complexe que les récompenses sont éparses. Pour beaucoup de jeux, la récompense ne survient qu'à la toute fin. Les agents doivent explorer l'environnement très longtemps avant de recevoir un retour quant à la qualité de leur politique. Pour atténuer ce problème, les concepteurs peuvent ajouter des pseudo-récompenses pour donner aux agents un retour plus fréquent. Par exemple, donner des récompenses à chaque but marqué par une équipe de football plutôt que de récompenser uniquement la victoire. Ces pseudo-récompenses peuvent toutefois mener à une politique sous-optimale, car elles sont ajoutées par le concepteur, qui

peut, par erreur, induire un mauvais comportement.

L'échelle de valeur des récompenses impacte également la convergence. Idéalement, les récompenses doivent être bornées entre $[-1, 1]$ et centrées en zéro. Une échelle de valeurs restreinte facilite les estimations des réseaux de neurones et favorise la convergence. En pratique, les récompenses négatives induisent parfois des comportements imprévisibles. Si dans une situation de trafic, un agent reçoit des récompenses négatives en raison du risque inhérent à sa situation, il pourra alors décider de provoquer un accident pour mettre fin aux récompenses négatives. Ce comportement découle de l'objectif du RL, qui consiste à accumuler un maximum de récompenses. Si l'agent juge qu'il ne pourra plus recouvrir les récompenses perdues, il tentera de mettre fin au jeu et provoquera un accident. La conception de fonctions donnant des récompenses négatives aboutit donc à des comportements imprévisibles dans certains jeux, car elles altèrent la nature du problème à résoudre.

Le nombre de fonctions de récompense par agent complexifie également l'apprentissage. Indépendamment de leur nombre, ces dernières sont converties en un scalaire pour l'apprentissage par réseau de neurones. Ce scalaire gomme les informations propres à chaque signal de récompense et nécessite plus de temps pour être appréhendé par les agents.

Enfin, l'espace des paramètres à apprendre impacte la mise à l'échelle des solutions multiagent. Nous avons détaillé les solutions par apprentissage centralisé permettant d'atténuer ce problème. D'autres paramètres, nommés *hyperparamètres*, ralentissent, voire condamnent la convergence. Les hyperparamètres englobent l'ensemble des paramètres qui ne peuvent s'apprendre. Entre autres, cela inclut : le facteur de réduction γ, le nombre de couches et de neurones par couches, le taux d'apprentissage α. Les algorithmes hors politique comportent plus d'hyperparamètres que ceux in-politiques. Leur valeur permet tantôt d'obtenir d'excellents résultats, tantôt de prévenir la convergence, sans qu'il existe de méthode fiable ou générale pour les régler. La recherche d'hyperparamètres efficaces est propre à chaque problème et devient chronophage lorsque leur nombre grandit.

Pour résumer, la dimensionnalité des espaces exacerbe les problèmes de non-stationnarité. Le nombre d'agents accroît exponentiellement la taille de l'espace d'état-action. Ce dernier rallonge la durée nécessaire de la phase exploratoire, laquelle est également exacerbée par la diversité des agents et la nature du jeu à résoudre. Aucune garantie de convergence vers une solution optimale n'existe pour les jeux compétitifs avec des agents hétérogènes [Nguyen *et al.*, 2020]. La conception de fonctions de récompense adaptées aux comportements attendus ainsi que le choix des hyperparamètres constituent des étapes cruciales de la résolution du problème. Malheureusement, aucune méthode fiable ne permet, à ce jour, de paramétrer efficacement ces éléments.

Conclusion

Dans ce chapitre, nous avons défini un problème MARL comme étant un couple composé d'un modèle de jeu et d'un ou plusieurs concepts de solutions. Le modèle de jeu régit les processus d'interaction, tandis que le concept énumère ses solutions possibles. La convergence vers une solution n'est garantie que pour certains cas spécifiques, souvent pour les jeux à deux agents reposant sur un horizon temporel court. La sélection d'un concept de solution reste complexe, car les solutions s'appliquent généralement à des jeux bien définis, comme les jeux de plateau, et leur transposition à des problèmes du monde réel n'a rien d'évident. Dans le trafic, par exemple, les interactions sont parfois éphémères et impliquent des nombres de conducteurs différents selon les situations.

Compte tenu de la diversité des situations rencontrées dans le trafic, composer une solution passera nécessairement par la conception d'une fonction d'approximation. En raison de son aptitude au traitement de données séquentielles, nous opterons pour une architecture récurrente. Plus précisément, nous avons détaillé les propriétés qui nous amenaient à penser que l'architecture GRU conviendrait à notre problématique.

Ensuite, nous avons précisé que nous souhaitions employer un algorithme acteur-critique qui apprenne une politique à partir de données générées hors politique. Rappelons les différents termes et les raisons qui ont orienté notre choix. Les algorithmes acteur-critique se composent d'un acteur qui apprend la politique et d'un critique qui évalue et guide la politique de l'acteur. Ces algorithmes s'adaptent généralement bien aux problèmes du monde réel. Du fait du coût des interactions avec un environnement composé de plusieurs dizaines d'agents, nous souhaitons apprendre à partir de données générées hors politique, réutilisables à différents moments de la phase d'entraînement. Enfin, compte tenu du nombre de situations possibles pouvant émerger d'un trafic, nous décidons d'apprendre directement une politique plutôt que d'apprendre les dynamiques d'un environnement influencées par des dizaines d'agents.

Concernant les modes d'entraînement et d'exécution, nous avons constaté que la centralisation permettait, dans une certaine mesure, d'accélérer la convergence, par l'avantage offert par des informations supplémentaires. Nous avons aussi remarqué que la décentralisation atténuait les problèmes de dimensionnalité. Le mode d'entraînement centralisé et d'exécution décentralisée nous paraissait être un bon compromis entre l'accélération de la convergence et l'atténuation des problèmes de dimensionnalité.

La dernière section a pointé les défis posés par l'implémentation d'une approche MARL pour répondre à un problème du monde réel. La non-stationnarité, exacerbée par la dimensionnalité des différents espaces, freine, voire condamne, la convergence. La non-stationnarité résulte des adaptations cycliques des agents aux politiques de leurs pairs. Nous avons examiné des méthodes de modélisation de population

permettant de l'atténuer. Ces approches restent néanmoins coûteuses et requièrent une longue phase d'apprentissage avec une population fixe d'agents. Compte tenu des caractéristiques du trafic, où les interactions sont parfois éphémères et où les usagers changent fréquemment, nous avons opté pour une approche peu coûteuse consistant à remplacer les informations trop anciennes par des nouvelles.

Enfin, la dimensionnalité des espaces d'agents, d'états, d'observations, d'actions, de paramètres et d'hyperparamètres sont autant d'écueils qui entraveront la convergence. Pour un environnement semblable au trafic, tantôt compétitif, tantôt coopératif et hautement hétérogène, aucune garantie de convergence n'existe. À ces difficultés, vient s'ajouter la conception de fonctions de récompense induisant des comportements conformément à ceux attendus. Là encore, aucune méthode fiable ne régit leur conception et, de ce fait, ces dernières induisent fréquemment des comportements inattendus, contre-intuitifs ou contre-productifs.

Grâce à cette première partie de contexte, nous disposons à présent de clés de compréhension sur la nature du trafic et sur les usagers qui le composent, ainsi que d'une vision globale de la conception d'une approche MARL. Abordons à présent à la partie contribution où nous bâtirons, dans un premier temps, une approche permettant d'atténuer les nombreux problèmes MARL susmentionnés.

Partie II

Contributions

CHAPITRE

Concilier hétérogénéité comportementale et passage à l'échelle

3

Sommaire du chapitre

Les deux premiers chapitres de l'état de l'art nous ont permis de comprendre quels étaient les verrous scientifiques propres à la simulation de trafic et au domaine de l'apprentissage par renforcement multiagent. Leurs conclusions énumèrent les différentes caractéristiques que nous souhaitions inclure dans notre modèle comportemental pour un robot de conduite. Nous avons jugé crucial de concevoir un modèle paramétrable et dont les comportements seraient hétérogènes. Toutefois,

le chapitre précédent insiste sur le fait que la diversité comportementale, lorsque appliquée à un grand nombre d'agents, augmente la complexité du système, l'empêchant parfois même de converger.

Concilier l'hétérogénéité comportementale et le passage à l'échelle d'un système multiagent apprenant s'avère donc être un verrou scientifique majeur que nous souhaitons lever dans ce premier chapitre de contribution.

La section 3.1 rappelle les objectifs de notre thèse et développe les problèmes posés par ce verrou scientifique (3.1.1). Cette section examine également les solutions proposées par la littérature (3.1.2).

Dans la section 3.2, nous proposons une solution afin de lever ce verrou. Nous commençons par donner une vision globale de notre approche (3.2.1). Puis, nous décrivons ses différentes composantes (3.2.2). Nous détaillons ensuite l'algorithme et l'architecture neuronale du modèle (3.2.3).

Enfin, la section 3.3 détaille les expériences et les résultats obtenus selon trois critères. Le premier critère s'assure de la capacité de passage à l'échelle du modèle proposé (3.3.1). Le second critère confirme la nature hétérogène des comportements résultants (3.3.2). Le troisième critère étudie la capacité de transfert d'apprentissage du modèle pour des comportements sur lesquels nous ne l'avons pas entraîné (3.3.3).

3.1 Définition du problème

Commençons cette section par rappeler nos objectifs et les verrous scientifiques freinant leur réalisation (3.1.1). Étudions ensuite les solutions proposées dans la littérature (3.1.2).

3.1.1 Verrous scientifiques

Lors du premier chapitre, nous avons constaté que le trafic se caractérise majoritairement par sa grande diversité interindividuelle. Chaque conducteur possède des attributs et des motivations propres, bien que certaines d'entre elles soient partagées par tous. Concernant notre thèse, l'objectif est de concevoir un système décisionnel essayant d'imiter les pratiques observées. Pour parvenir à un certain niveau de crédibilité, notre modèle doit nécessairement reproduire cette hétérogénéité comportementale.

Le chapitre précédent a néanmoins insisté sur la complexité inhérente aux systèmes multiagents, tels que le trafic, en particulier lorsque le nombre d'agents augmente. Nous avons noté que l'espace croît exponentiellement avec le nombre d'agents, ce qui tend à limiter la taille de la population simulée. La complexité croît d'autant plus quand les agents sont hétérogènes, puisque l'incertitude entourant les décisions des autres agents s'avère importante en l'absence de mécanisme permettant de l'atténuer.

Concilier l'hétérogénéité comportementale et le passage à l'échelle s'avère difficile dans le contexte MARL pour plusieurs raisons. Tout d'abord, plus la simulation comporte d'agents, plus la non-stationnarité et ses conséquences sont exacerbées. Ce défi devient d'autant plus contraignant que les agents compétitifs adaptent continuellement leur politique pour surpasser leurs opposants. Ce problème s'atténue dans le cas coopératif où les agents, certes, s'adaptent également aux comportements de leur équipe, mais jamais dans le but de minimiser leurs récompenses. Le trafic constitue un environnement n'encourageant ni les comportements totalement compétitifs ni ceux totalement coopératifs. Les comportements y sont mixtes, tantôt compétitifs lorsque les ressources, telles que l'espace routier, s'amenuisent, tantôt coopératifs autour des zones de fortes interactions, propices aux incidents.

Le second problème concerne l'hétérogénéité comportementale des agents RL. Chaque agent dispose de ses propres objectifs et donc de ses propres fonctions de récompense. Il devient alors difficile, voire impossible, de prédire avec certitude le comportement d'un agent. Pour tout agent souhaitant converger vers une politique optimale, il est crucial de comprendre les comportements des autres usagers, leurs motivations, et donc, leurs fonctions de récompense. S'il n'existe aucun signe distinctif entre les agents ou que ceux-ci sont trop nombreux, comme c'est le cas dans le trafic, la meilleure représentation consiste à déterminer la distribution de probabilité sous-tendant les comportements des agents. Ainsi, l'incertitude sur les comportements d'autrui est réduite de manière probabiliste en discriminant les comportements probables et improbables. Cette méthode convient lorsque la population d'agents partage certaines similitudes, mais peut apparaître détrimentaire dans les situations incidentogènes où une mauvaise anticipation accroît le risque d'accident.

Le troisième problème, spécifique au MARL de manière générale, concerne l'absence de garantie de convergence vers une solution optimale, à l'exception des scénarios avec un nombre d'agents restreint [Cui *et al.*, 2022]. C'est pourquoi la plupart des études menées avec des MARL se limitent, tout au plus, à une dizaine d'agents et attestent de la convergence ou non de leur apprentissage de manière empirique [Lowe *et al.*, 2017].

Enfin, le dernier problème provient de la consommation de ressources et du temps d'apprentissage. Sachant que chaque agent apprend sa politique par le biais d'un réseau de neurones, le temps de calcul s'accroît pour chaque agent supplémentaire présent dans la simulation. Comme l'augmentation du nombre d'agents va de pair avec celle de la non-stationnarité, le temps d'apprentissage s'en retrouve décuplé. De plus, on doit réapprendre depuis zéro chaque fois que l'on souhaite changer la distribution des caractéristiques de notre population d'agents.

Compte tenu des problématiques énoncées, la simulation d'une population hétérogène s'avère complexe. Voyons à présent comment les travaux de recherche ont tenté de surmonter ces difficultés.

3.1.2 Travaux similaires

Au travers d'une revue de littérature, nous avons identifié trois travaux de recherche qui abordent la problématique consistant à concilier le passage à l'échelle et l'hétérogénéité comportementale. Nous présentons ici leurs approches et leurs limites.

Toutes les approches que nous allons examiner se fondent sur la technique du *partage de paramètres* [Gupta *et al.*, 2017] (figure 3.1). Le partage de paramètres consiste à n'entraîner qu'un unique réseau de neurones pour l'ensemble des agents :

$$\theta_{\mathcal{I}} = \theta_1 = \ldots = \theta_n \tag{3.1}$$

De facto, l'approche appartient à la catégorie des entraînements centralisés. Ici, la centralisation apporte deux avantages majeurs. Le premier consiste à restreindre le nombre de paramètres apprenables. Ce dernier reste constant indépendamment du nombre d'agents entraînés. Le second avantage du partage de paramètres provient de la diversité des données d'entraînement. Ces données proviennent de l'ensemble des agents, et leur diversité vient améliorer la robustesse de l'apprentissage. Le partage de paramètres convient essentiellement aux agents homogènes, car le réseau partagé $\theta_{\mathcal{I}}$ prévient l'optimisation d'objectifs contradictoires. Cette contrainte implique une similarité entre les fonctions de récompense des agents.

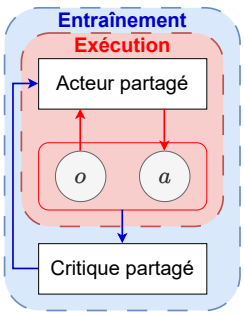


Figure 3.1 – Partage de paramètres

La première application du partage de paramètres au trafic fut proposée par Kaushik *et al.* [2020]. Leur méthode consiste à surcharger les observations des agents par un indice binaire induisant un comportement attendu. Selon sa valeur, l'indice peut encourager le véhicule à rester sur sa voie ($id = 0$) ou à dépasser un autre véhicule par un changement de voie ($id = 1$). La fonction de récompense des agents est conditionnée par cet indice :

$$\mathcal{R}(id) = \mathcal{R}_{\text{LK}} + id \cdot \mathcal{R}_{\text{LC}} \tag{3.2}$$

où $\mathcal{R}_{\text{LK}}$ et $\mathcal{R}_{\text{LC}}$ correspondent respectivement aux fonctions de récompense de maintien sur la voie et de changement de voie. Le partage de paramètres permet donc de générer des comportements hétérogènes en surchargeant le vecteur d'observation. Bien que les auteurs ne mentionnent jamais ces termes, nous considérons que leur approche correspond en réalité à un apprentissage de deux sous-politiques par partage de paramètres. Cette approche ne concerne en aucun cas l'hétérogénéité comportementale, car tous les agents adoptent ces deux sous-politiques indistinctement. Dans leurs expériences, les auteurs entraînent six véhicules par partage de paramètres. Lors de l'exécution, un maximum de quinze agents interagissent ensemble, portant la fréquence de collision à 11 %. Pour résumer cette approche, seuls deux comportements distincts résultent du partage de paramètres, et malgré un faible nombre d'agents, le taux de collision reste élevé.

Constatant les limitations de la première méthode, Christianos *et al.* [2021] ont tenté une seconde approche. Leur méthode de partage de paramètres sélectif consiste à apprendre plusieurs politiques, chacune fondée sur du partage de paramètres. Chaque politique correspond à un rôle et chaque agent se voit attribuer un rôle. L'hétérogénéité comportementale provient du nombre $K < |\mathcal{I}|$ de rôles, chacun correspondant à une politique π_k pour un ensemble d'agents homogènes. Autrement dit, l'hétérogénéité est interrôle et non interindividuelle. Ce compromis permet d'unir hétérogénéité comportementale et passage à l'échelle, car l'espace des paramètres apprenables dépend du nombre K de rôles et non du nombre d'agents $|\mathcal{I}|$. Toutefois, ce compromis sous-entend que la simulation d'une population hautement hétérogène sacrifie nécessairement le passage à l'échelle, ou *vice-versa*. Les expériences menées par les auteurs viennent d'ailleurs confirmer nos doutes, puisque, malgré l'entraînement simultané de 200 agents, seuls quatre rôles distincts sont appris.

Enfin, pointant le manque d'hétérogénéité comportementale des deux articles précédents, Yang *et al.* [2022] proposent une troisième approche où le partage de paramètres n'absorbe pas l'hétérogénéité comportementale des agents. Leur approche consiste en un algorithme acteur-critique où les poids du critique sont partagés tandis que chaque agent dispose de sa partie acteur. Les acteurs des agents tirent leurs données d'un ensemble d'entraînement partagé $\mathcal{D}_{\mathcal{I}}$, ce qui les catégorise parmi les approches par partage d'expériences sans partage de paramètres. Cette approche convient particulièrement lorsque l'homogénéité des agents s'avère trop faible pour envisager le partage de paramètres, mais que leur homogénéité permet toutefois d'apprendre à partir du même ensemble de données. Le partage d'expérience requiert un algorithme hors politique, car les données d'apprentissage proviennent de diverses politiques. Pour résumer la proposition de Yang *et al.* [2022], les auteurs favorisent la diversité au détriment de le passage à l'échelle. Leur expérience se restreint d'ailleurs à une quinzaine d'agents.

En définitive, les trois approches étudiées adoptent un compromis entre l'hétérogénéité comportementale de la population d'agents et son passage à l'échelle.

Aucune n'égale les performances du partage de paramètres, initialement conçu pour une population homogène, pour une population hétérogène. Par conséquent, aucune de ces méthodes ne répond à notre problématique, à savoir, la simulation d'un trafic hétérogène composé de dizaines d'agents. D'où la nécessité d'envisager une approche excluant tout compromis entre l'hétérogénéité comportementale et le passage à l'échelle, mais conciliant plutôt ces deux aspects.

3.2 Description du modèle GENEPI

Dans cette section, nous commençons par donner une vision globale du modèle que nous avons nommée GENEPI [Dinneweth *et al.*, 2023] (3.2.1). Puis, nous décrivons comment notre modèle apprend des comportements hétérogènes (3.2.2). Enfin, nous précisons les points techniques de son implémentation (3.2.3).

3.2.1 Vue globale

Notre proposition consiste à étendre le partage de paramètres, tout comme les approches citées ci-dessus, mais également à le combiner avec d'autres techniques d'apprentissage par renforcement. GENEPI se divise en deux phases : l'initialisation et l'entraînement (figure 3.2).

L'initialisation (à gauche sur la figure 3.2) attribue à chaque agent un profil de conduite distinct permettant d'obtenir une population hétérogène. Plus particulièrement, chaque agent reçoit un ensemble d'objectifs personnels échantillonnés depuis une distribution continue d'objectifs. Ces objectifs définissent le style de conduite des agents en définissant un comportement attendu. Par exemple, chaque agent reçoit une vitesse désirée qu'il doit chercher à atteindre. Ces objectifs conditionnent les fonctions de récompense respectives des agents. Ainsi, un agent avec une vitesse désirée de 70 km/h recevra une récompense maximale lorsque sa vitesse effective atteindra cet objectif.

Durant l'apprentissage (à droite sur la figure 3.2), les agents interagissent avec l'environnement, reçoivent leur observation, concatènent cette dernière avec leurs objectifs, envoient le résultat à une politique partagée qui détermine la meilleure action et, enfin, reçoivent leurs récompenses conditionnées par leurs objectifs respectifs ainsi que leur nouvelle observation. Précisons que l'environnement correspond au simulateur de trafic Archisim auquel s'ajoute les fonctions d'observation et de récompense. La politique repose sur du partage de paramètres et d'expériences. Le réseau de neurones partagé par les agents prend donc en entrée la concaténation de leur observation et de leurs objectifs. Contrairement aux approches de l'état de l'art distinguant différents comportements par un indice binaire, nos objectifs (les comportements attendus) sont définis par des valeurs continues. Une autre différence par rapport aux approches précédentes concerne la

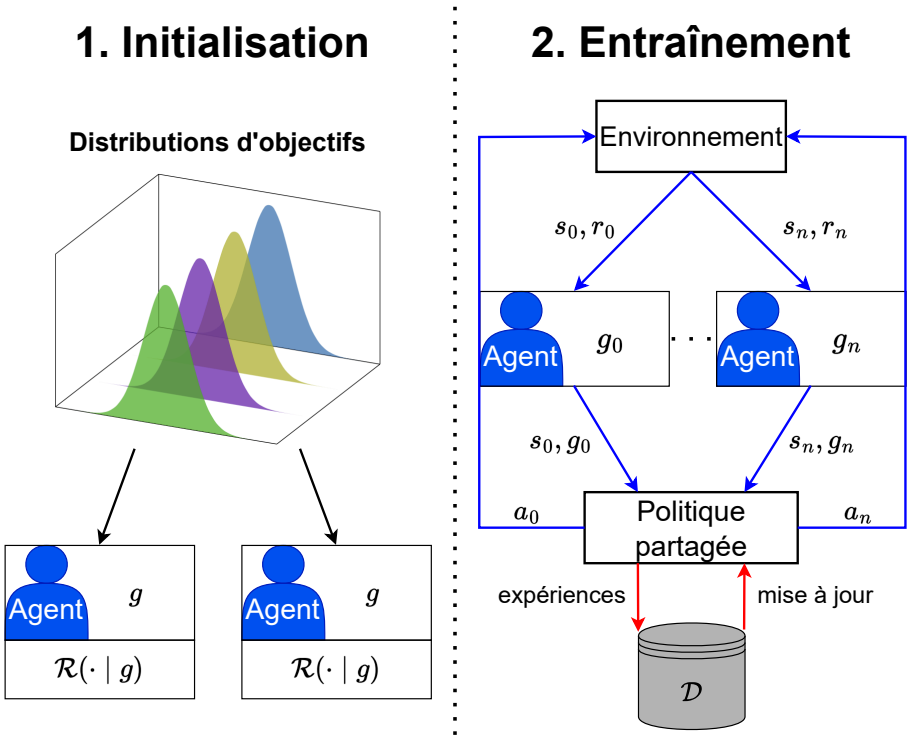


Figure 3.2 – Vue globale de GENEPI. Les flèches bleues et rouges décrivent respectivement les phases d'interaction avec l'environnement et d'apprentissage de la politique partagée. Nomenclature : jeu d'entraînement $\mathcal{D}$, objectif g, état s, action a, récompense r, fonction de récompense $\mathcal{R}$.

possibilité de combiner simultanément plusieurs objectifs. Ainsi, la taille de l'espace des comportements s'accroît exponentiellement avec le nombre d'objectifs.

Concernant la phase d'entraînement, nous utilisons un algorithme hors politique qui, rappelons-le, sauvegarde les expériences passées dans une mémoire tampon. Cette mémoire compense les coûts d'interaction avec l'environnement et permet de s'entraîner sur des données préalablement collectées. L'objectif de l'apprentissage consiste à optimiser des objectifs hétérogènes afin que ces derniers se traduisent en comportements hétérogènes.

Pour résumer, les agents reçoivent des objectifs continus caractérisant leur style de conduite et conditionnant leurs fonctions de récompense. L'objectif consiste à apprendre une politique partagée produisant des comportements hétérogènes. Le partage de paramètres permet d'atténuer le coût d'apprentissage. Sans cette approche, le temps d'apprentissage serait démultiplié, puisque chaque agent apprendrait sa propre politique. Examinons à présent la structure de notre modèle.

3.2.2 Générer une population hétérogène

Notre modèle repose en partie sur le concept d'approximateur de fonction de valeur universelle (UVFA [1]) pour générer une population hétérogène [Schaul *et al.*, 2015]. Le terme *approximateur* renvoie à l'utilisation d'un réseau de neurones plutôt que d'un tableau, la *fonction de valeur* désigne V ou Q et *universelle* caractérise la capacité de généralisation de l'apprentissage à plusieurs objectifs. Il s'agit d'une extension de la fonction de valeur généralisant non seulement sur les états s, mais aussi sur les objectifs $g \in \mathcal{G}$ d'un agent :

$$V_{g,\pi}(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \mathcal{R}_g\left(s^t, a^t, s^{t+1}\right) \prod_{k=0}^{t} \gamma_g(s^k) \,\middle|\, s^0 = s\right] \tag{3.3}$$

où $\mathcal{R}_g\left(s, a, s'\right)$ dénote une pseudo fonction de récompense et $\gamma_g(s)$ une pseudo fonction de réduction. Communément, une pseudo fonction de récompense estime ou approxime la récompense réelle afin de (1) faciliter l'apprentissage lorsque la récompense réelle renvoie rarement un signal ou (2) ajouter des connaissances supplémentaires pour guider l'agent. L'équation suivante correspond à la fonction équivalente pour la qualité d'action :

$$Q_{g,\pi}(s, a) = \mathbb{E}\left[\mathcal{R}_g\left(s, a, s'\right) + \gamma_g(s') \cdot V_{g,\pi}(s') \,\middle|\, s'\right] \tag{3.4}$$

Tout objectif g admet une politique optimale $\pi_g^*(s) = \operatorname{argmax}_a Q_{\pi,g}(s, a)$, une fonction de valeur optimale $V_g^* = V_{g,\pi_g^*}$ et une fonction de qualité d'action optimale $Q_g^* = Q_{g,\pi_g^*}$. L'UVFA consiste à représenter un ensemble de fonctions optimales de valeur par une unique fonction d'approximation généralisant sur les états et objectifs. Les fonctions d'approximation universelles $V(s, g; \theta) \approx V_g^*(s)$ et $Q(s, a, g; \theta) \approx Q_g^*(s, a)$ sont apprisent en optimisant les paramètres θ d'un réseau de neurones.

À notre connaissance, aucune approche ne combine l'UVFA et le partage de paramètres pour la simulation d'une population hétérogène. Notre modèle combine ces deux approches afin de profiter de leurs avantages en terme de passage à échelle. Ainsi, chaque agent i reçoit un ensemble d'objectifs $g_i = \{g_{i,1}, \ldots, g_{i,n}\}$ échantillonnés depuis des distributions continues d'objectifs $\mathcal{G} = \mathcal{G}_1 \times \ldots \times \mathcal{G}_n$ partagées par l'ensemble des agents. Chaque objectif $g_k \in \mathcal{G}_k$ est défini en termes d'états, *i.e.*, $\mathcal{G}_k \subseteq \mathcal{S}$ et conditionne une pseudo fonction de récompense $\mathcal{R}_{g_k}\left(s, a\right)$. La fonction de récompense $\mathcal{R}_i$ d'un agent i somme les pseudos récompenses associées à ses objectifs g_i, et éventuellement, applique des pondérations ω_k, tel que :

$$\mathcal{R}_i = \sum_{k=1}^{n} \omega_k \mathcal{R}_{g_{i,k}}\left(s, a, s'\right) \tag{3.5}$$

1. Universal Value Function Approximator

Décrivons à présent les objectifs de nos agents ainsi que leurs fonctions de récompenses associées. Nous dotons chacun de nos robots de conduite d'un couple d'objectifs, à savoir une vitesse désirée à atteindre et une marge de sécurité minimum à respecter. Chaque objectif est associé à une fonction de récompense, respectivement dénotée $\mathcal{R}_V$ et $\mathcal{R}_S$. Ces deux objectifs forment ensemble un compromis que tout conducteur en interaction doit s'efforcer d'équilibrer selon ses motivations. Afin de démontrer la faisabilité et l'efficacité de notre approche, nous nous concentrons sur un scénario basique dans lequel seul le mouvement longitudinal est autorisé. Ainsi, les agents peuvent uniquement atteindre leurs objectifs en modulant leur accélération longitudinale $\ddot{x}_i \in [-2;2]\ m.s^{-2}$.

La récompense de vélocité $\mathcal{R}_V \in [0;1]$ encourage l'agent à atteindre sa vitesse désirée $\widehat{\dot{x}}_i$, sachant sa vitesse actuelle $\dot{x}_i$:

$$\mathcal{R}_V = \left(\frac{\min\left(\dot{x}_i, \widehat{\dot{x}}_i\right)}{\max\left(\dot{x}_i, \widehat{\dot{x}}_i\right)} \right)^2 \tag{3.6}$$

La récompense est maximale lorsque l'écart entre sa vitesse désirée et sa vitesse actuelle $\dot{x}_i$ est nul. Les excès de vitesse comme les sous-régimes (par rapport à la vitesse désirée) sont punis de manière équivalente. La vitesse désirée d'un agent dépend de la limite de vitesse propre à la voie sur laquelle il se trouve, ainsi que d'une déviation par rapport à cette vitesse. La déviation est une caractéristique propre à l'agent, tirée d'une distribution lors de la phase d'initialisation.

La vitesse désirée d'un agent s'obtient donc par $\widehat{\dot{x}}_i = \dot{x}_{\ell_i}^{\max} \times \mathcal{N}\left(1;0{,}03\right)$ et tient compte des changements de limite de vitesse $\dot{x}_{\ell_i}^{\max}$ de la voie ℓ d'un agent i. Notons que le choix d'une distribution normale est motivé par le fait que la plupart des caractéristiques humaines, observées à l'échelle d'une population, tendent à former de telles distributions. Les valeurs choisies sont cependant arbitraires, puisque notre objectif est de proposer un modèle paramétrable par ses futurs utilisateurs, le rendant de fait compatible avec la simulation de trafics divers, indépendamment de toute considération culturelle et contextuelle.

La distribution est centrée en 1 ce qui signifie qu'en moyenne, la vitesse désirée des conducteurs correspondra à la limite de vitesse en vigueur. Toujours en moyenne, la moitié des conducteurs roulera à une vitesse inférieure à celle définie comme limite, tandis que l'autre moitié transgressera cette limite.

La seconde récompense $\mathcal{R}_S \in [-1;0]$ décourage l'agent d'être en trop forte interaction en punissant ce dernier lorsque le risque perçu $RP_i\left(j\right)$ lors d'une interaction avec un véhicule j qui le précède excède son seuil maximal de risque $RP_i^{\max}$ acceptable. Nous attribuons une récompense nulle lorsque le risque descend en deçà du seuil afin d'éviter d'induire un comportement excessivement prudent :

$$\mathcal{R}_S = \begin{cases} RP_i^{\max} - RP_i\left(j\right) & \text{si } RP_i\left(j\right) > RP_i^{\max} \\ 0 & \text{sinon} \end{cases} \tag{3.7}$$

Nous définissons une interaction forte comme dépassant un certain seuil de risque. Le risque étant une notion subjective, nous fondons son estimation sur l'étude

de Kondoh *et al.* [2008] (voir 1.2.2) dont nous avons décrit la formule dans le premier chapitre (éq. 1.9). Nous considérons ce calcul du risque comme pertinent, car il tient compte à la fois du différentiel de vitesse intervéhiculaire et du différentiel temporel avant une éventuelle collision. Considérés individuellement, l'intertemps et le temps avant collision retranscrivent médiocrement les situations incidentogènes. L'intertemps masque la dangerosité des situations avec des différentiels de vitesse conséquents, tandis que le temps avant collision néglige l'insécurité propre à l'adoption d'infimes marges de sécurité. Leur association et leur pondération donnent cependant une meilleure compréhension de la nature des interactions. Ces deux variables possèdent également l'avantage de dépendre du temps, ce qui signifie que leur interprétation est quasi indépendante des situations observées.

La subjectivité de la perception du risque correspond à la seconde caractéristique personnelle. Ainsi, nous dotons chaque véhicule d'un seuil de risque maximal qu'il peut accepter $RP_i^{\max} \sim \mathcal{N}(1{,}5; 0{,}2)$. Pour des raisons analogues à la caractéristique précédente, le seuil de risque maximal est échantillonné depuis une distribution Gaussienne. Dans leur étude, Kondoh *et al.* [2008] notent que le risque maximal toléré par les conducteurs de son étude excède rarement $RP = 2$. Notre distribution tente donc approximativement de se conformer à leurs observations. Toutefois, comme la caractéristique précédente, ces valeurs correspondent à des paramètres ajustables par les utilisateurs du modèle.

Enfin, nous pondérons identiquement ces deux fonctions qui forment la récompense globale $\mathcal{R}_i \in [-1; 1]$, soit le compromis entre vitesse et sécurité $\mathcal{R}_i = \mathcal{R}_V + \mathcal{R}_S$.

Décrivons à présent les variables qui composent les observations egocentrées transmises aux agents. Une observation se compose de six entrées normalisées dans l'intervalle $[-1; 1]$, conformément aux besoins du réseau de neurones les traitant. Notons que les deux objectifs sont aussi normalisés. Quatre variables décrivent les interactions entre les véhicules et les deux variables restantes apportent des informations macroscopiques sur le trafic.

Ces observations proviennent du simulateur de trafic Archisim sur lequel nous fondons notre approche. Au niveau macroscopique, Archisim divise le trafic en zones (figure 3.3) dont la dimension s'adapte à la vitesse de l'agent. Ces zones forment un quadrillage latéral des voies et longitudinal des portions de voies.

Dans notre cas, nous conservons uniquement les deux zones précédant l'agent (figure 3.4), car notre étude se restreint aux manœuvres longitudinales. Chacune de ces zones est décrite par une variable donnant la vitesse moyenne des véhicules présents à l'intérieur. Si la zone ne contient aucun véhicule, la variable indique sa vitesse limite.

Les quatre variables restantes décrivent deux interactions entre des paires de véhicules. La première interaction correspond à celle entre l'agent et le véhicule qui le précède, tandis que la seconde décrit l'interaction entre le véhicule qui le

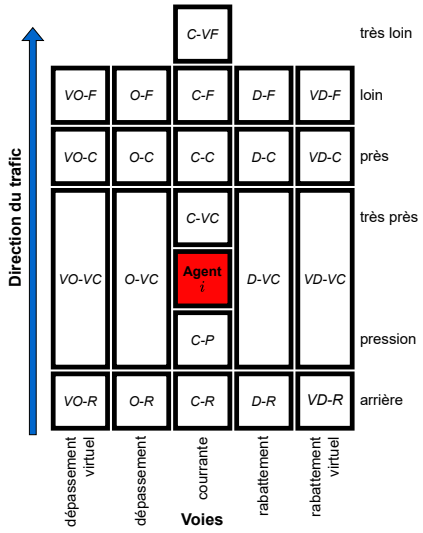


Figure 3.3 – Zones d'ArchiSim égocentrées sur l'agent (rouge). Les voies virtuelles concentrent l'ensemble les voies au-delà des voies adjacentes

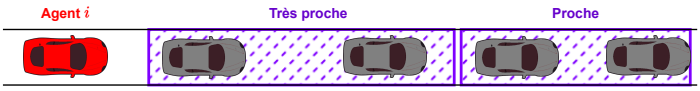


Figure 3.4 – Zones observées (violet) par l'agent (rouge)

précède et celui précédant ce dernier (figure 3.5). Une interaction se décrit selon deux variables : l'intertemps TH et le temps avant collision TTC.

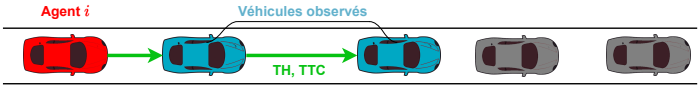


Figure 3.5 – Description des interactions (flèches vertes) entre l'agent (rouge) et les autres véhicules impliqués (bleus)

Les entrées sont ensuite traitées par un algorithme de renforcement que nous décrivons à présent.

3.2.3 Algorithme et architecture neuronale

Concernant le choix de l'algorithme, nous avons vu au chapitre précédent que les algorithmes acteur-critique offrent un bon compromis. Dans ces algorithmes, le critique apprend à associer des récompenses aux transitions et cette information permet ensuite d'évaluer la qualité de la politique de l'acteur qui associe des actions

aux états. Sachant que le critique sert exclusivement à guider les décisions de l'acteur durant l'entraînement, il devient superflu une fois l'apprentissage terminé. Ainsi, lors de l'exécution de la politique apprise, le coût calculatoire est doublement réduit, car (1) le critique n'est plus appelé et (2) le réseau de neurones de l'acteur n'effectue plus de rétropropagation, soit l'étape la plus coûteuse.

Parmi les algorithmes acteur-critique existants, nous souhaitons obtenir en sortie une action continue représentant l'accélération longitudinale. Compte tenu du coût d'interaction avec un système composé de plusieurs dizaines d'agents, nous recherchons un algorithme *hors-politique* afin de pouvoir s'entraîner sur des expériences anciennes, générées par une version antérieure de la politique courante. Ces contraintes nous amènent à opter pour l'algorithme D4PG dont le critique présente l'avantage de cartographier plus finement la fonction de valeur via une représentation distributionnelle [Barth-Maron *et al.*, 2018].

L'approche distributionnelle évite les erreurs décisionnelles [Bellemare *et al.*, 2017]. Prenons l'exemple d'un conducteur qui hésite entre deux chemins différents pour rallier un point d'arrivée le plus rapidement possible et surtout en moins de vingt minutes. D'après son expérience, le premier itinéraire s'effectue en moyenne en quinze minutes, mais au gré des conditions de circulation, le temps de trajet est parfois doublé. Le second itinéraire s'effectue en moyenne en dix-sept minutes et les conditions de circulation sont toujours favorables. Une représentation distributionnelle du choix d'itinéraire nous indiquera que le second itinéraire s'avère préférable au premier du fait de l'absence d'incertitude quant au respect de la contrainte temporelle de vingt minutes. Une représentation scalaire nous indiquera cependant le premier itinéraire comme préférable au second en moyenne.

Ainsi, la récompense espérée indiquée par un scalaire cache une réalité parfois complexe que seule une distribution peut habilement déceler. Dans certains cas, une évaluation par un scalaire peut empêcher l'apprentissage d'une politique optimale et produire des décisions contre-productives. La figure 3.6 illustre ce phénomène où la valeur moyenne prédite par un scalaire (à gauche) se révèle également être la moins probable lorsque l'on représente la même situation de manière distributionnelle (à droite). Cette dernière permet donc d'éviter les erreurs de jugement.

Toutefois, l'approche distributionnelle s'avère plus coûteuse, car l'algorithme apprend à minimiser les erreurs d'estimation probabilistes pour chaque quartile plutôt que pour une seule valeur comme l'approche scalaire.

Notre architecture neuronale acteur-critique s'inspire de *Deep Dense Architectures in Reinforcement Learning* (D2RL) [Sinha *et al.*, 2020] et de l'*architecture de récompense hybride* (HRA) [Van Seijen *et al.*, 2017], le tout adapté pour correspondre à la représentation distributionnelle de l'algorithme D4PG. Ces architectures permettent de stabiliser l'apprentissage et d'améliorer la convergence. Détaillons leurs mécanismes.

Les auteurs de l'architecture D2RL constatent que plus le nombre de couches cachées d'un réseau augmente, plus ses performances se détériorent. Selon eux,

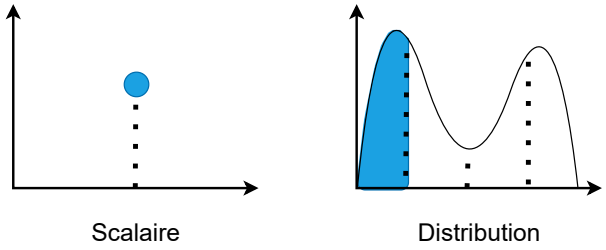


Figure 3.6 – Représentations scalaire et distributionnelle des probabilités. La zone bleue correspond à la valeur la plus problable.

chaque étape de propagation efface une partie des informations contenues dans l'entrée initiale. Pour contrecarrer cette tendance et conserver l'avantage des réseaux denses pour la modélisation de données complexes, l'architecture D2RL concatène l'entrée initiale à chaque couche du réseau (figure 3.7). Ainsi, D2RL préserve l'information initiale indépendamment du nombre de couches cachées du réseau.

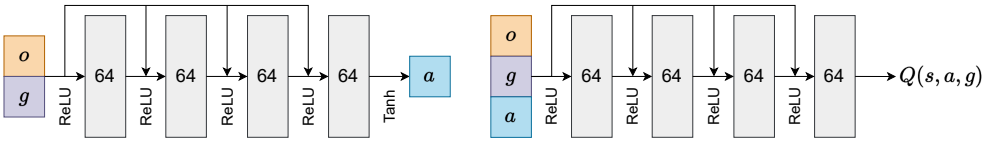


Figure 3.7 – Architecture D2RL avec les réseaux acteur (gauche) et critique (droite). Les blocs gris représentent des couches neuronales entièrement connectées et leur valeur indique leur nombre de neurones.

L'architecture de récompense hybride (HRA) se propose d'éliminer les instabilités de l'apprentissage et ainsi d'accélérer la vitesse de convergence des algorithmes. Selon ses auteurs, les algorithmes de RL peinent à généraliser l'approximation d'une fonction de valeur optimale $V(s)$ ou $Q(s,a)$. Pour pallier ce problème, les auteurs proposent d'apprendre une fonction de valeur simplifiée, décomposant la fonction de récompense en n sous-fonctions de récompense différentes (figure 3.8). Chacune de ces fonctions devient alors un problème indépendant et dispose d'une fonction d'approximation particulière. Leurs approximations sont ensuite aggrégées, produisant une valeur scalaire pour chaque action. Pour résumer, l'architecture HRA accélère l'apprentissage de problèmes multi-objectifs, sous réserve que la fonction de récompense soit décomposable.

Tout comme l'algorithme D4PG originel, nous utilisons une mémoire tampon prioritisée et distribuée $\mathcal{D}_p$ de taille $\left|\mathcal{D}_p\right|$ [Horgan *et al.*, 2018]. L'algorithme 1 décrit les cycles d'interaction avec l'environnement et d'apprentissage d'une politique

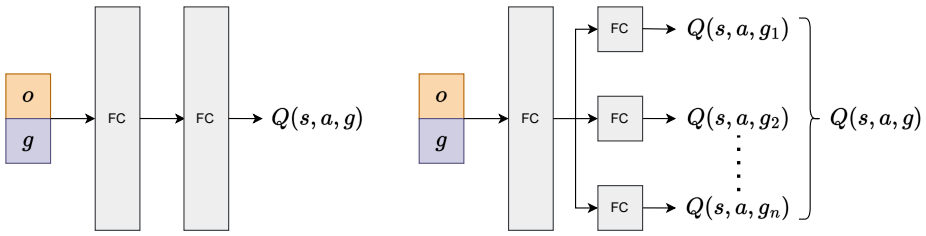


Figure 3.8 – Architectures par défaut (gauche) et HRA (droite). FC dénote une couche complètement connectée.

optimale de l'algorithme D4PG modifié, incluant l'apprentissage multi-objectif HRA. L'algorithme commence par initialiser aléatoirement les paramètres des réseaux de neurones de l'acteur et du critique (ligne 2) et de copier leurs valeurs vers les réseaux cibles (ligne 3). Chaque itération (ligne 4) comprend une phase d'interaction avec l'environnement (lignes 5-8). Un agent échantillonne une action en se fondant sur la politique partagée et sur un facteur d'exploration aléatoire ϵ (ligne 6). Il exécute cette action et reçoit une récompense r conditionnée à ses objectifs g (ligne 7). Ces interactions alimentent le jeu de données d'entraînement $\mathcal{D}_p$ (ligne 8). Vient ensuite une phase d'apprentissage (lignes 9-13). Durant cette phase, les données sont échantillonnées depuis le jeu d'entraînement $\mathcal{D}_p$ (ligne 10). À partir de ces données, les distributions cibles Y_{i,g_k} – correspondant à l'estimation des valeurs d'état-action – sont calculées pour chaque objectif (ligne 11). Puis, les paramètres des réseaux de neurones de l'acteur et du critique sont mis à jour (ligne 12).

Dans cet algorithme, le retour est défini par la variable aléatoire Z_π, telle que :

$$Q_\pi(s,a) = \mathbb{E}\left[Z_\pi(s,a)\right] \tag{3.8}$$

où s dénote la concaténation de l'observation et les objectifs $s = (o,g)$ avec $g = \{g_1,\ldots,g_n\}$.

Algorithme 1 : GENEPI

Entrées : Taille du lot $|\mathcal{B}|$, longueur d'une trajectoire N, mémoire tampon prioritisée et distribuée $\mathcal{D}_p$ de taille $|\mathcal{D}_p|$, taux d'apprentissage α_μ et α_Q, fonction de mesure de l'erreur distributionnelle $\mathcal{L}$, un taux d'exploration ϵ.

Initialiser aléatoirement les poids des réseaux acteur et critique (θ_μ, θ_Q)

Initialiser les poids cibles $(\theta^*_\mu, \theta^*_Q) \leftarrow (\theta_\mu, \theta_Q)$

pour *chaque iteration* **faire**

  pour *chaque interaction avec l'environnement* **faire**

    Échantillonner une action $a = \pi_{\theta_\mu}(s) + \epsilon\mathcal{N}(0;1)$

    Exécuter l'action a, recevoir la récompense $r = \{r_{g_1}, \ldots, r_{g_n}\}$ et la concaténation de l'observation suivante et les objectifs $s' = (o', g)$

    Sauvegarder la transition $\mathcal{D}_p \leftarrow \mathcal{D}_p \cup \{(s, a, r, s')\}$

  pour *chaque mise à jour du gradient* **faire**

    Échantillonner un lot de transitions $\mathcal{B} = \left\{d_i \sim \mathcal{U}\left(\mathcal{D}_p\right)\right\}_{i=1}^{|\mathcal{B}|}$ avec une priorité p_i où $d_i = \left(s^{i:i+N}, a^{i:i+N-1}, r^{i:i+N-1}\right)$

    Construire les distributions cibles de chaque objectif g_k :

$$Y_{i,g_k} = \left(\sum_{n=0}^{N-1} \gamma^n r^{i+n}_{g_k}\right) + \gamma^N Z_{\theta^*_\mu, g_k}\left(s^{i+N}, \pi_{\theta^*_\mu}\left(s^{i+N}\right)\right)$$

    Mettre à jour les paramètres $\theta_Q \leftarrow \theta_Q + \alpha_Q \delta_{\theta_Q}$ et $\theta_\mu \leftarrow \theta_\mu + \alpha_\mu \delta_{\theta_\mu}$ où :

$$\delta_{\theta_Q} = \frac{1}{|\mathcal{B}|} \sum_i \nabla_{\theta_Q} \left(|\mathcal{D}_p| \cdot p_i\right)^{-1} \mathbb{E}\left[\sum_{g_k} \mathcal{L}\left(Y_{i,g_k}, Z_{\theta_Q, g_k}\left(s^i, a^i\right)\right)\right]$$

$$\delta_{\theta_\mu} = \frac{1}{|\mathcal{B}|} \sum_i \nabla_{\theta_\mu} \pi_{\theta_\mu}\left(s^i\right) \mathbb{E}\left[\nabla_a \sum_{g_k} Z_{\theta_Q, g_k}\left(s^i, a\right)\right]\Big|_{a=\pi_{\theta_\mu}(s^i)}$$

    Mettre à jour les paramètres cibles $(\theta^*_\mu, \theta^*_Q) \leftarrow (\theta_\mu, \theta_Q)$

retourner θ_μ

3.3 Expériences et résultats

Dans cette section, nous décrivons les expériences conduites et les résultats obtenus. Nous évaluons GENEPI selon trois critères : le passage à l'échelle (3.3.1), l'hétérogénéité comportementale (3.3.2) et le transfert d'apprentissage (3.3.3).

Commençons par décrire le contexte expérimental. Toutes les expériences sont menées dans le simulateur de trafic *ArchiSim* avec une carte graphique *@NVIDIA*

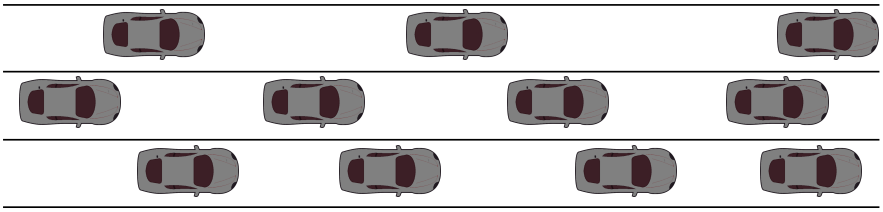


Figure 3.9 – Scénario GENEPI

RTX 3080 TI. Le scénario étudié consiste en une autoroute à trois voies où les agents ne peuvent changer de voie (figure 3.9). Ainsi, les agents disposent seulement d'une action continue leur permettant d'accélérer ou de ralentir pour maximiser les récompenses induites par leurs objectifs respectifs.

Pour établir un lien entre le simulateur *ArchiSim* et le module d'apprentissage, nous avons réalisé un travail préliminaire qui a nécessité plusieurs mois de développement. Ce travail a consisté, entre autres, à établir des échanges entre le simulateur *ArchiSim*, programmé en *Modula-2*, et la partie apprentissage, programmée en *Python*. L'échange de données porte sur (1) l'envoi des observations aux agents du simulateur vers le module d'apprentissage et (2) la réception de leurs actions après que celles-ci aient été déterminées par le réseau de neurones. L'optimisation des échanges de données fut cruciale pour obtenir des résultats avec des délais raisonnables.

Le tableau 3.1 liste les hyperparamètres retenus. Nous employons un taux d'apprentissage élevé afin de mitiger les effets de la non-stationnarité sur l'apprentissage. Cette approche appartient à la catégorie *oublier* du chapitre précédent (2.3.1). Concernant le facteur de réduction γ, nous constatons que les entraînements menés avec des valeurs proches de 1 échouent à converger. Nous attribuons cet échec au fait qu'en apprentissage multiagent, chaque agent peut influencer les récompenses de ses pairs, augmentant ainsi l'incertitude relative à l'obtention de récompenses futures. Nous optons donc pour une valeur plus faible $\gamma = 0.85$.

Table 3.1 – Hyperparamètres GENEPI

Optimiseur	Adam
Taux d'apprentissage α (GENEPI)	$1e-3$
Taux d'apprentissage α (référence)	$3e-4$
Taille de la mémoire tampon $\left\|\mathcal{D}_p\right\|$	$5e+4$
Facteur de réduction γ	0.85
Taille de lot $\left\|\mathcal{B}\right\|$	64

3.3.1 Passage à l'échelle

La première expérience vise à estimer l'avantage procuré par GENEPI en termes de passage à l'échelle vis-à-vis d'une approche décentralisée, *i.e.* sans partage de paramètres où chaque agent apprend individuellement à optimiser ses objectifs [Bernstein *et al.*, 2002]. L'approche de référence est identique à l'algorithme GENEPI, excepté qu'elle ne dispose pas de partage de paramètres et que son taux d'apprentissage diffère (tableau 3.1). Pour l'approche de référence, nous remarquons qu'un taux d'apprentissage sensiblement plus faible donne de meilleurs résultats. Nous espérons que GENEPI sera capable d'apprendre avec un plus grand nombre d'agents qu'avec l'approche de référence.

Nous expérimentons les deux approches avec différents nombres d'agents (3, 6, 10, 30, 60 et 90) et vérifions chaque minute la convergence. Nous considérons que l'algorithme converge lorsque les agents accumulent au minimum 90% des récompenses d'un épisode, soit 40 secondes de simulation. Ce score signifie qu'aucun accident n'est survenu au cours de la phase de test. Nous supposons que l'algorithme échoue à converger si, après vingt heures d'apprentissage (temps utilisateur), les agents n'ont toujours pas atteint ce seuil de récompense. Chaque expérience est répétée cinq fois avec des objectifs g initiaux différents, des positions x_i initiales des agents différentes et une initialisation aléatoire des paramètre θ_Q et θ_μ.

Pour les deux approches, nous mesurons le temps moyen de convergence et son écart-type (tableau 3.2). Les résultats montrent que l'approche de référence échoue à converger en moins de vingt heures lorsque plus de six agents apprennent simultanément. Nous attribuons cet échec à la non-stationnarité. En effet, avec l'approche décentralisée, chaque agent apprend indépendamment des autres et nécessite plus d'interactions avec l'environnement pour accumuler un nombre d'expériences équivalent à l'approche centralisée par partage de paramètres (GENEPI). Or, les comportements de ses pairs évoluent à un rythme plus élevé que celui auquel un agent accumule des expériences. Les agents apprennent donc à s'adapter à des comportements qui ont déjà changés. Avec GENEPI, les agents accumulent des expériences à un rythme supérieur, car ces dernières sont partagées par tous.

Comme attendu, l'algorithme GENEPI converge plus rapidement que l'approche décentralisée de référence. Son temps de convergence reste presque constant ($2{,}4$ minutes en moyenne) indépendamment du nombre d'agents. Cette constance est probablement la conséquence du partage de paramètres dont l'espace des paramètres apprenables reste invariant quel que soit le nombre d'agents entraînés. Nous n'avons pas simulé plus de 90 agents, car nous considérons que dans un contexte restreint à la conduite en file, l'augmentation du nombre d'agents n'a que peu de chances de produire des situations différentes et donc des résultats différents.

Le partage de paramètres permet également d'atténuer l'hétérogénéité com-

Table 3.2 – Temps (en minutes) avant convergence. Le symbole * dénote l'absence de convergence.

# agents	GENEPI	Référence
3	$\mathbf{3{,}4 \pm 1{,}5}$	14 ± 5
6	$\mathbf{2{,}6 \pm 0{,}5}$	28 ± 26
10	$\mathbf{2 \pm 0{,}6}$	*
30	$\mathbf{1{,}6 \pm 0{,}8}$	*
60	$\mathbf{2{,}2 \pm 1{,}2}$	*
90	$\mathbf{2{,}6 \pm 0{,}8}$	*

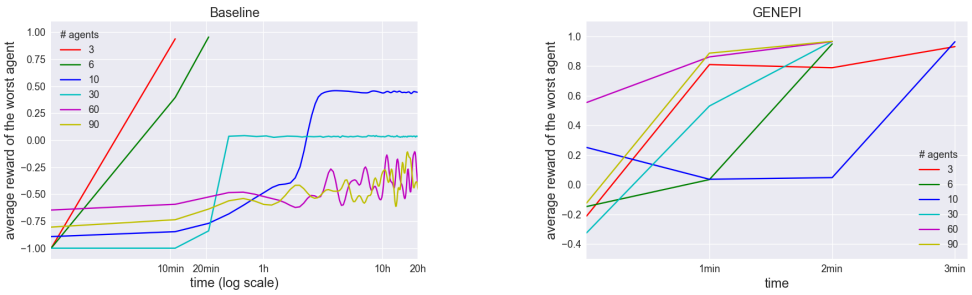


Figure 3.10 – Temps de convergence de l'approche de référence et de GENEPI

portementale et son impact sur la non-stationnarité. Les agents se comportent de manière plutôt homogène, car ils partagent la même politique. Ainsi, les agents apprennent à s'adapter à une hétérogénéité comportementale restreinte, accélérant la vitesse de convergence. L'effet inverse se produit pour l'approche décentralisée. La figure 3.10 nous permet d'observer ce phénomène. Avant même que l'entraînement n'ait commencé, à $t = 0$, la pire récompense moyenne obtenue par les agents de l'algorithme GENEPI est clairement plus importante ($-0{,}4$) que celle obtenue par l'approche de référence (-1).

Maintenant que nous avons montré la capacité de GENEPI à entraîner plusieurs dizaines d'agents simultanément, nous devons vérifier que leurs comportements sont effectivement hétérogènes. Dans les expériences qui suivent, nous réduisons le nombre d'agents à cinq afin de gagner en clarté. Par souci de clarté, les courbes du véhicule roulant en tête ne seront pas affichées, car ce dernier ne peut jamais être en interaction avec un autre véhicule.

3.3.2 Hétérogénéité comportementale

Pour évaluer l'hétérogénéité comportementale, nous concevons un scénario dans lequel chaque agent a une vitesse désirée supérieure à celle de l'agent qui le

précède. Ainsi, pour accumuler un maximum de récompenses, les agents doivent atteindre leur limite de risque maximale.

La figure 3.11 montre le risque estimé des agents au cours du temps. Notons que le risque du cinquième agent est omis, car n'ayant aucun prédécesseur, son risque est nul. Les agents 0, 1 et 2 atteignent leur limite de risque respective $RP_0 = 2$, $RP_1 = 1{,}8$, $RP_2 = 1{,}6$ avant la fin de la simulation, tandis que l'agent 3 n'a pas encore atteint sa limite $RP_3 = 1{,}4$. Nous observons que les agents cessent d'accélérer une fois leur limite de risque atteinte. Ces résultats confirment que GENEPI peut produire des comportements hétérogènes selon des objectifs renseignés.

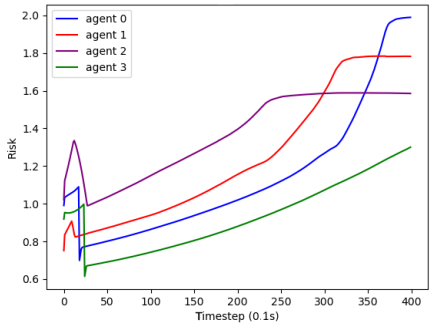


Figure 3.11 – Risques perçus par quatre agents durant la simulation

3.3.3 Transfert d'apprentissage

Enfin, le dernier critère évalué concerne le transfert d'apprentissage. Cette notion se réfère à la capacité de généralisation d'un algorithme d'apprentissage sur des nouveaux objectifs, différents de ceux sur lesquels il a appris. Cette capacité est évaluée sans réapprentissage.

Initialement, les objectifs des agents étaient tirés d'une distribution $RP_i \in [1, 2]$. Pour évaluer la capacité de transfert de GENEPI, nous tirons une distribution $RP_i \in [2, 3]$ et observons si les agents adoptent les comportements attendus.

La figure 3.12 confirme nos attentes et montre que les agents s'adaptent correctement à cette nouvelle distribution, puisque les agents tendent à maintenir un risque RP dans l'échelle de valeur de la nouvelle distribution. Notons que les agents 0 et 1 (courbes bleue et rouge) observent un pic de risque peu avant la fin de l'expérience. Ce pic est dû à une accélération trop brusque de la part de ces agents. Puisque la fonction de risque croît de manière exponentielle à mesure que le temps avant collision diminue, une accélération trop brusque provoque ce pic.

Notre proposition jouit donc d'une capacité de transfert d'apprentissage, ce qui permettra de modifier les caractéristiques des agents sans devoir réentraîner le modèle.

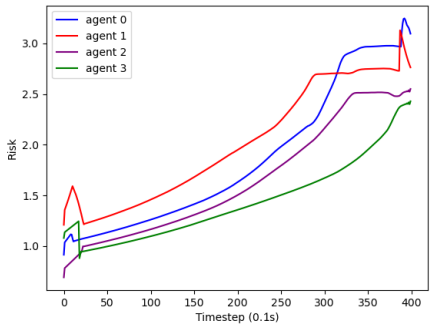


Figure 3.12 – Capacité de transfert d'apprentissage

Conclusion

L'objectif de cette première contribution était de concevoir un modèle d'agent capable de comportements hétérogènes selon des caractéristiques renseignées par un concepteur. Les verrous scientifiques existants insistaient sur la difficulté de concevoir simultanément un modèle hétérogène et pouvant être mis à échelle.

Notre contribution a permis de concilier hétérogénéité comportementale et passage à l'échelle dans un contexte spécifique. Dans un scénario de conduite sur route, nos résultats montrent que notre approche converge là où l'approche de référence échoue, et ce, même lorsque nous augmentons drastiquement le nombre d'agents. La mesure de l'hétérogénéité comportementale résultante correspond aux caractéristiques données aux agents en entrée. De plus, le modèle affiche une capacité de transfert d'apprentissage, permettant d'éviter la nécessité d'un réentraînement chaque fois que les distributions de caractéristiques changent.

Nous avons à présent un système décisionnel de robot de conduite pouvant constituer un trafic hétérogène, soit l'une des caractéristiques principales observées dans les trafics composés uniquement de conducteurs humains. Le premier objectif de notre thèse étant atteint, nous passons donc au second : intégrer des comportements sociaux à notre modèle.

CHAPITRE

Gestion sociale des interactions

4

Sommaire du chapitre

Lors du premier chapitre de contexte, nous avons mis en exergue les défis propres à l'introduction de robots de conduite (RC) dans le trafic. Nous avons mentionné la nécessité d'opérer une phase de transition maîtrisée afin d'assurer la sécurité d'une phase transitoire voyant un trafic composé exclusivement de conducteurs humains s'automatiser progressivement. Parmi les pistes souhaitables et *a priori* réalistes, nous avons souligné que l'appropriation de comportements humains par les RC permettrait d'améliorer les conditions de sécurité de cette phase transitoire. En d'autres termes, concevoir des RC socialement désirables minimiserait les incompréhensions entre les protagonistes d'un trafic mixte et

améliorerait de fait la sécurité d'un éventuel trafic mixte.

Poursuivant cette réflexion, ce chapitre propose un modèle de gestion des interactions sociales pour des RC. Plus précisément, nous visons à doter ces véhicules d'une capacité d'empathie situationnelle, similaire à celle observée chez les conducteurs humains.

La section 4.1 définira le cadre du problème. Nous entamerons cette section par l'introduction de certains concepts d'apprentissage par renforcement, préliminaires à la compréhension globale de notre approche (4.1.1). Nous rappellerons ensuite les défis posés par la conception de RC socialement désirables et les limites des approches existantes (4.1.2).

La section 4.2 décrira l'approche envisagée. Nous commencerons par donner une vue globale du modèle (4.2.1). Notre modèle s'inspirant de la structure décisionnelle hiérarchique des conducteurs humains, nous décrirons la logique des niveaux tactiques (4.2.2) et opérationnels (4.2.3).

Enfin, la section 4.3 présentera les expériences menées et les résultats obtenus. Nous y décrirons les conditions expérimentales (4.3.1), puis discuterons des résultats (4.3.2).

4.1 Robots de conduite socialement désirables

Cette section introduira certaines approches d'apprentissage par renforcement qui composeront notre modèle (4.1.1). Puis, nous identifierons les défis posés par la conception de RC socialement désirables ainsi que les limites des approches existantes (4.1.2).

4.1.1 Préliminaires

Avant de nous attaquer au sujet central de la conception de RC socialement désirables, introduisons quelques concepts d'apprentissage par renforcement qui nous seront utiles ultérieurement. Nous avons discuté, au second chapitre (2.2.3), de certaines stratégies d'apprentissage MARL. Examinons à présent les stratégies d'apprentissage RL applicables au cas multiagent. Plus précisément, nous nous concentrerons sur les stratégies inspirées des mécanismes cognitifs humains accélérant la convergence vers une solution.

Les approches que nous présenterons appartiennent au champ de recherche de l'apprentissage par curriculum [Elman, 1993]. Ce domaine vise à simplifier le problème à résoudre sans toutefois en altérer la nature. Nous présentons deux approches majeures, à savoir : la difficulté adaptative et l'apprentissage hiérarchique.

Les approches par difficulté adaptative supposent que, parallèlement aux êtres humains, les algorithmes de RL maîtriseraient plus rapidement un problème si la difficulté de ce dernier augmentait graduellement, commençant par une version simplifiée du problème initial et ajoutant de la complexité au fur et à mesure de

l'acquisition de compétences par les agents [Bengio *et al.*, 2009]. Par exemple, un joueur de Go novice évoluera plus facilement s'il s'entraîne avec des joueurs de son niveau que s'il joue constamment face à un grand maître. C'est d'ailleurs en suivant ce paradigme que l'agent RL AlphaGo a battu le champion du monde en titre [Silver *et al.*, 2017]. L'agent jouait uniquement contre lui-même, un adversaire constamment à sa portée.

Pour le cas du trafic, la difficulté peut se mesurer selon la densité. Ainsi, la difficulté de l'environnement peut s'accroître progressivement à mesure que l'agent améliore ses compétences (figure 4.1).

La difficulté adaptative est principalement utilisée en RL, mais nous soutenons que cette dernière pourrait mitiger les problèmes de non-stationnarité dans un environnement multiagent. Moins de véhicules signifie une plus grande stationnarité. À mesure que les politiques de chacun se stabilisent, l'environnement peut accepter plus de véhicules et se rapprocher du problème initial plus rapidement qu'avec l'approche classique.

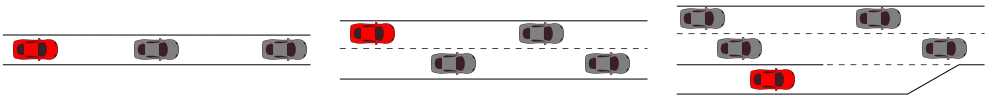


Figure 4.1 – Apprentissage par curriculum pour la conduite. De gauche à droite : les agents commencent à apprendre dans un trafic fluide, puis le trafic se densifie, et les interactions se multiplient

La difficulté adaptative repose donc sur un concept simple permettant de résoudre des problèmes complexes qui nécessiteraient plus de temps pour être maîtrisés, voire impossible dans certains cas.

La seconde approche se nomme l'apprentissage hiérarchique (HRL [1]) [Pateria *et al.*, 2021]. Le HRL repose sur le concept de *diviser pour mieux régner*. Ce paradigme suppose que la politique d'un agent peut s'apprendre plus efficacement si cette dernière est subdivisée en plusieurs sous-politiques, chacune dédiée à un problème particulier (figure 4.2). Chaque sous-politique ainsi apprise peut également être réutilisée pour des tâches connexes. Par exemple, la connaissance acquise pour un déboîtement (pour dépasser un véhicule) pourrait être réutilisée par une sous-politique adressant le problème du rabattement, puisque les deux sous-politiques concernent un changement de voie.

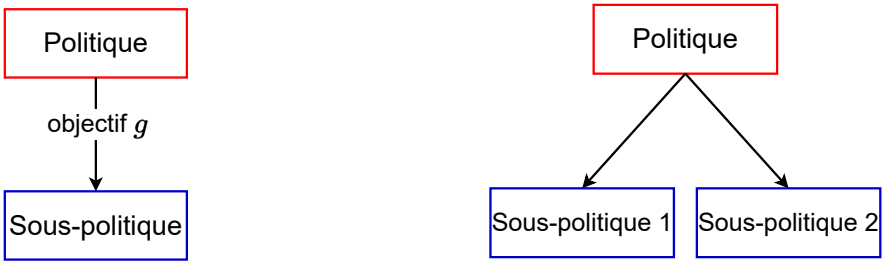
La division en sous-politiques bénéficie également à la résolution de tâches gangrenées par un problème de dimensionnalité, car chacune d'elles disposerait alors d'une partie restreinte de l'espace d'état-action, confinée à ses besoins. Un autre avantage concerne l'apprentissage concurrent de divers objectifs, chacun appartenant à un niveau décisionnel distinct. Par exemple, une politique de haut niveau peut définir la manière dont gérer les interactions dans le trafic, tandis qu'une

1. Hierarchical Reinforcement Learning

politique de bas niveau contrôlerait le véhicule. Cet exemple correspond d'ailleurs à la description du système décisionnel des conducteurs par les psychologues [Michon, 1985].

Deux politiques peuvent ainsi évoluer selon des temporalités différentes. Cette abstraction temporelle permet de limiter la fréquence d'appel à certaines politiques et donc d'engager des tâches plus complexes. Par ailleurs, le lien entre HRL et les sciences cognitives a été mis en évidence par Botvinick *et al.* [2009].

Plusieurs implémentations d'HRL ont été proposées. Parmi elles, nous distinguons deux structures : celles conditionnées par un objectif et celles par sélection d'options (figure 4.2).



(a) Conditionné à un objectif g **(b)** Sélection d'options

Figure 4.2 – Apprentissage par renforcement hiérarchique

Pour la première famille, HRL conditionné par un objectif, une politique de haut niveau définit un objectif qu'une politique de bas niveau devra atteindre (figure 4.2a). Imaginons un robot devant se déplacer d'un point A à un point B à l'intérieur d'une maison et que des murs contraignent l'agent à emprunter un chemin indirect, car différent d'une ligne droite. Dans ce cas, une politique de haut niveau peut découper l'itinéraire initial en une série d'itinéraires directs. Ces derniers deviennent des objectifs que la politique de haut niveau transmet à celle de bas niveau. La complexité de la tâche initiale est ainsi minimisée, chaque niveau HRL est dédié à la réalisation d'une tâche plus simple. L'objectif défini par la politique de haut niveau peut être statique (une position fixe) ou mobile (une balle rebondissante). Mathématiquement, les structures conditionnées par un objectif g s'apparentent à l'approximateur universel de valeur de fonction (UVFA) discuté au chapitre précédent, du moins pour les niveaux inférieurs de la hiérarchie [Schaul *et al.*, 2015].

Pour la seconde famille, HRL par sélection d'options, les politiques de niveau supérieur remplacent leurs actions par des politiques de niveau inférieur (figure 4.2b). Les politiques de bas niveau sont appelées des options et les politiques de haut niveau se contentent d'opter pour l'une de ces options. Par exemple, dans le cas de la conduite, une voiture souhaitant dépasser une autre pourrait disposer d'une politique de haut niveau optant pour entre deux sous-politiques de dépassement par la droite ou par la gauche. Dans cette configuration, les politiques de haut

niveau doivent anticiper les résultats probables de leurs sous-politiques afin de les sélectionner efficacement.

Indépendamment de la famille HRL déployée, les différents niveaux peuvent être appris soit séquentiellement, soit parallèlement.

Dans le premier cas, les étages inférieurs de la hiérarchie sont appris avant ceux des niveaux supérieurs. L'apprentissage se simplifie, puisque les politiques de haut niveau font appel à des politiques de bas niveau stables, car non-apprenantes.

Dans le second cas, tous les étages sont appris en même temps. L'apprentissage se complexifie, car le résultat des politiques de niveau inférieur évolue au cours de l'entraînement, alors que l'objectif des politiques de niveau supérieur consiste justement à prédire le résultat de leur politique. L'apprentissage concurrent de plusieurs niveaux nécessite donc plus d'interactions avec l'environnement, bien que certaines approches atténuent ces difficultés [Levy *et al.*, 2017; Nachum *et al.*, 2018].

La prévalence de l'une des deux formes d'apprentissage dépend du problème traité ainsi que de la structure hiérarchique retenue. Si la résolution du problème dépend principalement de la qualité des interactions entre les différents niveaux de la hiérarchie, alors l'apprentissage HRL parallèle peut s'avérer plus approprié. Cependant, une hiérarchie complexe requerra certainement un apprentissage HRL séquentiel. Certaines approches hybrident ces deux formes d'apprentissage afin de trouver un compromis entre efficacité et complexité.

Pour résumer, l'apprentissage par curriculum vise à faciliter l'apprentissage d'une tâche sans toutefois la dénaturer. L'apprentissage par curriculum revêt plusieurs formes, allant de la gradation adaptative de la difficulté du problème au découpage hiérarchique de la politique. Ces deux formes pouvant être conjointement appliquées, nous les emploierons d'ailleurs toutes deux dans notre modèle.

4.1.2 Comportement socialement désirable

Nous avons discuté au premier chapitre de la volonté politique de se diriger vers un trafic automatisé. Nous avions alors insisté sur le fait que la transition du trafic actuel vers un trafic automatisé doit s'effectuer de manière maîtrisée et réfléchie pour limiter l'émergence de situations incidentogènes. Parmi les options envisageables, nous avons suggéré que le trafic mixte serait éventuellement plus acceptable par les conducteurs si les RC montraient des capacités sociales similaires à celles des humains [Dinneweth *et al.*, 2022]. Cette voie, nous l'espérons, favorisera une bonne interprétation mutuelle des intentions de chacun et rendra le trafic mixte plus sûr.

Cependant, cette dernière hypothèse va au-delà des ambitions de la thèse. La vérifier nécessiterait de réaliser des expérimentations en condition réelle avec un véhicule autonome piloté par notre modèle naviguant sur la même route que des

conducteurs humains. Ici, nous nous concentrerons donc sur la conception d'un module d'empathie.

L'insertion sur autoroute est indéniablement l'un des scénarios où les capacités sociales des conducteurs influencent le plus le trafic. Les zones d'insertion sont propices aux conflits du fait que les conducteurs luttent pour une ressource finie : l'espace. L'appropriation de l'espace devient d'autant plus conflictuelle que deux flux de véhicules ayant des vélocités inégales fusionnent en un seul et unique flux. Dans la plupart des zones d'insertion, hors périphérique parisien où la priorité à droite prévaut, les véhicules du flux principal sont prioritaires. Cette norme formelle n'empêche cependant aucunement l'émergence de comportements sociaux, donc informels, car absents ou contraires au Code de la route. Les comportements sociaux peuvent avoir pour but de diminuer les contraintes des autres usagers, le cas altruiste, ou de leur signifier notre indifférence, le cas égoïste. L'influence des comportements sociaux sur les interactions est donc prépondérante autour des zones d'insertion. Ils permettant un fonctionnement "apaisé" et relativement efficace : cas de l'effet "zip" où les conducteurs de l'axe prioritaire et ceux de la bretelle d'insertion s'auto-organisent.

Les psychologues ont investigué ces questions d'altruisme et d'égoïsme chez l'être humain par le concept de valeur d'orientation sociale (SVO) [Murphy *et al.*, 2011]. Selon leur définition de la SVO, un individu altruiste sacrifie une partie de son utilité au profit des autres, tandis que l'égoïste refuse ce compromis. Dans le cadre de la conduite automobile, l'utilité d'un conducteur peut se mesurer de diverses manières et inclure, par exemple, la vitesse du conducteur ou encore son niveau d'interaction. Schwarting *et al.* [2019] ont d'ailleurs tenté d'estimer la SVO des conducteurs en temps réel. Nous estimons toutefois que leurs travaux estiment le résultat d'interactions quantifié par un calcul de la SVO plutôt que la véritable intention en termes de SVO. Par là, nous signifions que le résultat d'une interaction ne suit pas toujours le déroulé intentionnellement prévu.

Le modèle de conducteur MOBIL, dont nous avons déjà parlé, fut le premier à introduire un concept proche de la SVO [Kesting *et al.*, 2007]. MOBIL inclut un facteur de politesse influant sur les décisions de changement de voie. Par un ajustement de ce facteur, les motivations des agents peuvent varier de l'altruisme à l'égoïsme. Les agents altruistes chercheront à minimiser le freinage des véhicules proches induit par leur changement de voie, tandis que les agents égoïstes déconsidèreront les conséquences de leurs actions sur les autres usagers.

[Sadigh *et al.*, 2018] ont simulé les influences réciproques entre un RC et des conducteurs humains. Leur étude révèle que les RC peuvent influencer les décisions des conducteurs humains. Cette preuve de concept a ouvert la voie à la conception de RC dotés de SVO et donc désirables socialement. Également menée en simulation, une étude décrit les conducteurs humains comme plus propices à l'empathie envers leurs pairs qu'envers des RC [Sun *et al.*, 2024]. Ces travaux semblent indiquer que l'intégration de RC au trafic serait plus harmonieuse si les

comportements exhibés par ces derniers se rapprochaient de ceux des conducteurs humains.

Plusieurs recherches ont intégré une forme de valeur d'orientation sociale aux RC, révélant au passage que l'excès d'altruisme ou d'égoïsme produit des effets néfastes sur le trafic autour des zones d'insertion [Toghi *et al.*, 2021a,b, 2022]. Plus récemment, une autre étude récente s'est concentrée sur les comportements sociaux associés à la traversée de piétons [Li *et al.*, 2022].

Nous considérons que la majorité des travaux présentés jusqu'alors souffrent de certaines limitations qui les éloignent des pratiques humaines et qui, par conséquent, pourraient compromettre la sécurité du trafic.

Premièrement, ces travaux conçoivent la SVO comme un moyen d'améliorer l'utilité des usagers proches sans identifier qui parmi eux a réellement besoin d'assistance. De ce fait, les actions empathiques ne sont pas obligatoirement dirigées vers les véhicules les plus nécessitants. Imaginons qu'un RC approche d'une zone d'insertion depuis les voies principales et qu'il soit considéré comme gênant par un véhicule souhaitant s'insérer. Si ce RC ralentit pour faciliter l'insertion de l'usager, il minimise par la même occasion l'utilité de tous les véhicules le suivant. Dans ce cas, difficile de prédire la décision du RC. Si ce dernier avait identifié le véhicule de la rampe d'insertion comme étant l'objet de son empathie compte tenu de ses contraintes, alors la décision de faciliter son insertion aurait été plus claire. L'optimisation de l'utilité locale peut donc causer de l'incertitude dans certaines situations. En outre, sa mise en pratique peut se révéler difficile à interpréter par les conducteurs humains qui, habituellement, se concentrent sur un lot restreint de véhicules à aider.

Deuxièmement, ces travaux omettent d'étudier la faisabilité de leurs décisions empathiques. Dans certains cas, malgré la nécessité d'une action empathique, celle-ci est irréalisable par le RC, soit, car les contraintes spatio-temporelles la préviennent, soit, car cette action dépend d'un autre usager. En cherchant à optimiser l'utilité de tous les usagers proches, le RC sera « puni » pour une situation dont la responsabilité lui échappe. Pire encore, il pourrait tenter une manœuvre désespérée, donc peu prévisible, pour tenter de rétablir la situation.

Confiner les actions empathiques là où elles sont nécessaires et effectives soulève de nombreux défis, mais pourrait éventuellement renforcer la sécurité des interactions en trafic mixte. Ce sont ces défis que nous abordons avec le modèle Archicool et sa composante principale : l'empathie sélective.

4.2 Modèle Archicool

Archicool tente de se rapprocher des pratiques humaines et introduit le concept d'empathie sélective. Notre modèle s'inspire de la structure décisionnelle hiérarchique des conducteurs étudiée au premier chapitre. Cette section commence

par une explication globale du modèle (4.2.1), puis nous détaillons le processus hiérarchique aux niveaux tactique (4.2.2) et opérationnel (4.2.3).

4.2.1 Vue globale

Au premier chapitre, nous avons souligné les travaux de psychologues de la conduite indiquant une division hiérarchique à trois niveaux du processus décisionnel des conducteurs : le niveau stratégique dédié à la planification du parcours, le niveau tactique gérant les interactions et le niveau opérationnel responsable de l'exécution des manœuvres [Michon, 1985].

Archicool s'inspire de cette hiérarchie (figure 4.3), à l'exception du niveau stratégique que nous laissons de côté, puisque la planification de trajet sort du cadre de nos travaux. Le niveau tactique du modèle émet des préférences sur la manière d'interagir avec les autres usagers. Ses préférences sont ensuite transmises au niveau opérationnel qui s'efforcera de réaliser les manœuvres spécifiées par la conduite du véhicule si les conditions de sécurité sont propices. Par exemple, le niveau tactique peut anticiper une longue interaction avec le véhicule précédant et soumettre au niveau opérationnel une demande de changement de voie.

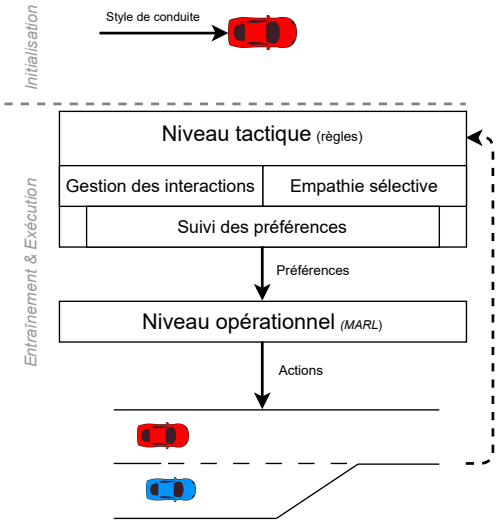


Figure 4.3 – Archicool

À l'instar du modèle GENEPI, les agents reçoivent un ensemble de quatre préférences caractérisant leur style de conduite à l'initialisation : une vitesse désirée $\widehat{x}_i$, d'une limite maximale de risque toléré $RP_i^{\max}$, d'un jerk maximum toléré $\widetilde{j}_i^{\max}$ et d'un niveau de valeur d'orientation sociale (SVO) ϕ_i. Le jerk correspond à la dérivée des accélérations longitudinales comme latérales. Pour un rappel sur la

SVO et l'interprétation des valeurs, voir la figure 1.3. Ces attributs influencent les processus tactiques et opérationnels.

Les décisions tactiques émanent exclusivement de bases de règles, tandis que celles du niveau opérationnel résultent de l'apprentissage par renforcement. Cette hybridation tire avantage de la capacité de résolution de problèmes du RL tout en conservant les spécificités propres aux décisions humaines.

Le niveau tactique comprend deux processus mutuellement exclusifs : la gestion des interactions et l'empathie sélective. La gestion des interactions minimise les interactions avec les autres usagers, tandis que l'empathie sélective assiste les autres usagers dans leurs manœuvres. Quel que soit le processus s'exécutant, la sortie du niveau tactique correspond à un ensemble de deux préférences indiquant la vitesse et la voie désirées. Afin d'assurer leur pertinence à travers le temps, nous introduisons un processus de suivi des préférences. Ce dernier permet d'annuler les préférences qui ne font plus sens au regard d'une situation qui aurait évolué.

Le niveau opérationnel reçoit les préférences retenues par le niveau tactique. Il tente de satisfaire ces préférences, et celles du style de conduite définies à l'initialisation, en pilotant le véhicule. Par exemple, si le niveau tactique indique une préférence pour un changement de voie, le niveau opérationnel doit trouver un créneau spatio-temporel dans la voie cible pour réaliser la manœuvre demandée. En sortie, le niveau opérationnel détermine les accélérations longitudinale et lattérale nécessaires pour piloter le véhicule.

Détaillons à présent les niveaux tactique et opérationnel.

4.2.2 Niveau tactique

Le niveau tactique comprend deux processus mutuellement exclusifs : la gestion des interactions et l'empathie sélective. Le premier minimise les interactions de l'agent avec les autres usagers, tandis que le second minimise l'interaction d'autres usagers avec l'agent. Une interaction (courante ou anticipée) est un problème d'accès à une ressource partagée, en l'occurrence l'espace routier, qui peut se terminer par un conflit ou une coopération. Le choix du processus à suivre dépend en partie de la SVO, car pour exécuter une action empathique, un agent devra probablement concéder une part de son utilité personnelle au profit de l'amélioration de l'utilité d'autres usagers. Par utilité, nous désignons le degré d'interaction avec les autres usagers et les conséquences en termes de vitesse et de risque perçu. Bien que visant des objectifs contraires, ces processus peuvent converger vers des solutions identiques. Ce serait le cas d'un changement de voie permettant à un agent d'atteindre sa vitesse désirée tout en facilitant l'insertion d'usagers sur la voie que ce dernier prévoit de quitter.

Le processus de gestion des interactions se fonde sur les travaux d'une psychologue de la conduite et de l'implémentation de ces derniers sur le simulateur de trafic *Archisim* [Espié et Saad, 2000]. Nous avons détaillé les travaux de F. Saad lors du premier chapitre (1.2). Rappelons néanmoins quelques points essentiels. F.

Saad a identifié un ensemble de règles motivationnelles partagées par l'ensemble des conducteurs. Pour résumer simplement, les conducteurs cherchent à minimiser la durée de leurs interactions soit par un changement de voie lorsque l'interaction actuelle (ou anticipée) dure plusieurs secondes, soit par une modulation de la vitesse et de l'intertemps véhiculaire si l'interaction semble éphémère ou en l'absence d'opportunités pour changer de voie.

Le processus d'empathie sélective comporte quatre étapes :

1. Identifier des véhicules cibles pouvant bénéficier d'une action empathique
2. Déterminer la capacité de l'agent à aider le véhicule cible
3. Estimer le coût de l'action empathique
4. Décision de réaliser ou non l'action empathique

Notons une interaction Ψ, sa durée d'une interaction t_Ψ et une interaction de longue durée, $t_\Psi^+ = 2{,}5$ s. Désormais, nous prenons le point de vue d'un seul agent $i \in \mathcal{I}$, bien que tous les véhicules $j \in \mathcal{I} \setminus \{i\}$ présents se comportent selon le modèle Archicool. Nous employons le mot agent pour référer à i et le mot véhicule pour référer à tout autre véhicule j proche de l'agent i.

La première étape de l'empathie sélective, l'identification de véhicules cibles, se concentre sur les véhicules des voies adjacentes à l'agent i présents dans son champ visuel avant (figure 4.4). Dans cette zone, l'agent i identifie les véhicules cibles pouvant bénéficier d'une action empathique de sa part. L'agent doit s'assurer que ces véhicules ont effectivement l'intention de venir sur sa voie pour éviter toute action empathique vaine. Cette intention se déduit à partir d'indices visuels, parmi lesquels se trouve (1) l'approche d'une fin de voie ℓ_j par le véhicule, (2) la durée de l'interaction $t_\Psi(j, \cdot)$ subie par le véhicule et (3) un différentiel de vitesse $\Delta_{\dot{x}}(\ell_i, \ell_j)$ défavorable entre la voie ℓ du véhicule j et celle de l'agent i.

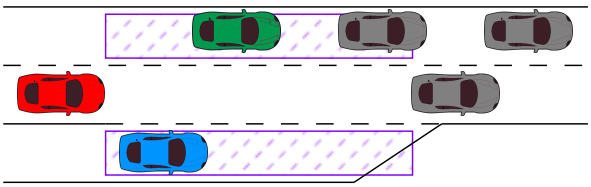


Figure 4.4 – Identification de véhicules cibles (vert et bleu) pour l'empathie sélective selon les zones de recherche (violet) du point de vue de l'agent (rouge)

Parmi les cibles potentielles pour une action empathique, l'agent i concentre son attention sur le véhicule subissant la contrainte la plus haute $\hat{\imath} = \operatorname{argmax}_j \hat{c}_i^t(j)$, estimée par :

$$\hat{c}_i^t(j) = \omega_{RP} \cdot \sigma\left(\ln\left(RP_j\left(\overline{\ell_j}\right)\right)\right) + \omega_\Psi \cdot \sigma\left(\ln\left(\frac{t_\Psi(j,z)}{t_\Psi^+} \cdot \left(\dot{x}_{\ell_j}^{\max} - \dot{x}_j\right)^+\right)\right) \tag{4.1}$$

où $\hat{c}_i^t(j) \in [0;1]$ est la contrainte de j estimée par i, $\dot{x}_{\ell_j}^{\max}$ est la vitesse limite de la voie de j, $\dot{x}_j$ est la vitesse de j, $RP_j\left(\overline{\ell_j}\right)$ le risque perçu par j par rapport à la fin de sa voie $\overline{\ell_j}$, et σ dénote la fonction sigmoïde. Le symbole $(\cdot)^+$ réfère à la partie positive, soit $\max(0,\cdot)$. Nous accordons une pondération plus importante au risque perçu $\omega_{RP} = \frac{2}{3}$ qu'à la durée d'interaction $\omega_\Psi = \frac{1}{3}$ pour prioriser la sécurité.

La décision de poursuivre le processus d'empathie sélective envers le véhicule $\hat{\imath}$ repose en partie sur la SVO de l'agent. Contrairement aux approches existantes, nous utilisons la contrainte des véhicules plutôt que leur utilité pour déterminer si l'agent doit se montrer altruiste avec le véhicule cible. Si l'inéquation suivante est vraie, l'agent passe à la seconde étape du processus :

$$\cos(\phi_i) \cdot c_i < \sin(\phi_i) \cdot c_{\hat{\imath}} \tag{4.2}$$

où ϕ_i est la SVO (degré d'utilité accordé à soi *versus* aux autres) de l'agent i. Sinon, il applique le processus de gestion d'interaction (voir le second paragraphe de la section 4.2.2).

Cette seconde étape s'assure de la capacité de l'agent à assister le véhicule cible $\hat{\imath}$ dans son changement de voie. L'agent détermine quelles manœuvres lui permettent d'atteindre cet objectif. Deux options s'offrent à lui : freiner ou changer de voie. Pour le freinage, l'agent détermine sa durée et son intensité, tandis que pour le changement de voie, l'agent évalue sa capacité à effectuer la manœuvre sous contrainte spatio-temporelle.

Un freinage empathique consiste à décélérer jusqu'à ce que le véhicule $\hat{\imath}$ puisse reconnaître l'intention via la création d'un créneau sécurisé. Du point de vue de l'agent i, la manœuvre consiste à atténuer le risque perçu entre lui et le véhicule cible $\hat{\imath}$. Un créneau sécurisé entre les deux véhicules se définit comme une interaction dont le risque perçu passe sous un certain seuil $RP_i^t(\hat{\imath}) < \widehat{RP}_i(\hat{\imath})\ \forall\, t \in [0, t_i^{b,\max}]$. La limite haute, $t_i^{b,\max}$, correspond à la durée maximum de freinage. Dans le cas où la voie du véhicule se termine, cette limite désigne le temps restant avant que $\hat{\imath}$ n'atteigne la fin de sa voie compte tenu de sa dynamique actuelle :

$$t_i^{b,\max} \leq TTC\left(\hat{\imath}, \overline{\ell_{\hat{\imath}}}\right) - \hat{t}_i^{\ell_{\hat{\imath}} \to \widehat{\ell}_{\hat{\imath}}} \tag{4.3}$$

où $\hat{t}_i^{\ell_{\hat{\imath}} \to \widehat{\ell}_{\hat{\imath}}}$ désigne l'estimation par i du temps nécessaire à $\hat{\imath}$ pour atteindre sa voie cible. Cette durée dépend de la distance latérale $\Delta_y\left(\hat{\imath}, \widehat{\ell}_{\hat{\imath}}\right)$ séparant le véhicule $\hat{\imath}$ de sa voie cible $\widehat{\ell}_{\hat{\imath}}$, et de son accélération latérale maximum $\ddot{y}_{\hat{\imath}}^{\max}$:

$$\hat{t}_i^{\ell_{\hat{\imath}} \to \widehat{\ell}_{\hat{\imath}}} = \frac{1}{\ddot{y}_{\hat{\imath}}^{\max}} \left(-\dot{y}_{\hat{\imath}} + \sqrt{\dot{y}_{\hat{\imath}}^2 + 2\Delta_y\left(\hat{\imath}, \widehat{\ell}_{\hat{\imath}}\right) \cdot \ddot{y}_{\hat{\imath}}^{\max}} \right) \tag{4.4}$$

L'algorithme 2 résume ces calculs et détermine l'intensité de la décélération permettant de proposer un créneau sécurisé. Afin d'éviter un freinage trop brusque,

l'algorithme (ligne 3) teste plusieurs décélérations $\ddot{x}_i \in [-2;0]$ et détermine la valeur minimum créant un créneau sécurisé pour le véhicule $\hat{\imath}$. L'algorithme retourne ∞ (ligne 10) si le freinage excède les capacités de l'agent étant donné les contraintes situationnelles. Ce qui peut survenir lorsque la distance séparant l'agent du véhicule est faible et que le différentiel de vitesse est élevé. Dans ce cas, le freinage serait trop brusque et il semble plus pertinent que le véhicule change de voie après le passage de l'agent. Sinon, il retourne la perte de vitesse estimée (ligne 7).

Algorithme 2 : CoutFreinage

1 **Entrées** : Agent i, véhicule $\hat{\imath}$, durée maximale de freinage $t_i^{\flat,\max}$.
2 $\ddot{x}_i \leftarrow 0$
3 **tant que** $\ddot{x}_i \geq -2$ **faire**
4 $\quad t_i^{\flat} \leftarrow \text{DureeFreinage}\left(\widehat{RP}_i\left(\hat{\imath}\right), \ddot{x}_i, t_i^{\flat,\max}\right)$
5 $\quad$ **si** $t_i^{\flat} \leq \min\left(5, t_i^{\flat,\max}\right)$ ***et*** $\ddot{x}_i \neq 0$ **alors**
6 $\quad\quad \widehat{\dot{x}}_i \leftarrow \dot{x}_i + \ddot{x}_i \cdot t_i^{\flat}$
7 $\quad\quad$ **retourner** $\left(\dot{x}_i - \widehat{\dot{x}}_i\right)^+$
8 $\quad$ **sinon**
9 $\quad\quad \ddot{x}_i \leftarrow \ddot{x}_i - \frac{1}{2}$
10 **retourner** ∞

Pour effectuer ses estimations, cet algorithme (ligne 4) en appelle un second (3) qui renvoie le temps que doit durer la décélération de l'agent afin de créer un créneau sécurisé. Dans ce second algorithme (lignes $3-5$), si la durée de décélération vaut zéro, le véhicule $\hat{\imath}$ ne requiert aucune aide pour effectuer sa manœuvre et le temps de freinage requis est donc égal à zéro. Si la dérivée estimée du risque à l'instant $t_i^{\flat}$ est inférieure à 0 (lignes $8-10$), *i.e.* que freinage pourrait diminuer le risque perçu, alors l'algorithme renvoie la durée de freinage correspondante. L'algorithme retourne ∞ (ligne 13) lorsqu'aucun freinage ne peut créer de créneau sécurisé selon les contraintes temporelles courantes.

Le niveau tactique doit communiquer ses préférences (voie, vitesse) au niveau opérationnel. Le freinage empathique consiste donc à diminuer la vitesse désirée $\widehat{\dot{x}}_i$ de l'agent. Cette vitesse se calcule à partir de l'intensité et de la durée du freinage données par les deux algorithmes.

La seconde manœuvre empathique envisagée par l'agent est le changement de voie, sous condition que le marquage au sol l'autorise. Comme pour le freinage, l'agent estime la perte de vélocité engendrée par un changement de voie selon les différentiels de vitesse entre sa voie courante et sa voie cible (celle se trouvant à l'opposé du véhicule $\hat{\imath}$).

Une fois ces deux manœuvres considérées, l'agent passe à la troisième étape du

Algorithme 3 : DureeFreinage

Entrées : Seuil de risque maximal $\widehat{RP}_i(\hat{\imath})$, décélération souhaitée $\ddot{x}_i$, durée maximale de freinage $t_i^{\flat,\max}$
$a, b, c \leftarrow$ FormePolynomial($\widehat{RP}_i(\hat{\imath})$)
$\vec{s} \leftarrow$ Résoudre(a, b, c) // solutions de $at^2 + bt + c = 0$
si $|\vec{s}| = 0$ ***ou*** $a < 0$ **alors**
  retourner 0 // freinage inutile
pour $t_i^\flat$ ***dans*** $\vec{s}$ **faire**
  si $t_i^\flat \geq 0$ **alors**
    $d_{\text{risque}} \leftarrow RP_i^{t_i^\flat\,\prime}(\hat{\imath} \mid \ddot{x}_i)$ // dérivée du risque à $t_i^\flat$
    si $d_{risque} < 0$ **alors**
      retourner $t_i^\flat$
    sinon si $t_i^\flat > t_i^{\flat,\max}$ **alors**
      retourner 0 // freinage inutile
retourner ∞ // aucun freinage empathique possible $\hat{\imath}$

processus d'empathie sélective. Celle-ci consiste à déterminer l'action empathique la moins contraignante parmi celles restantes. L'agent associe donc à chaque manœuvre un coût correspondant à la perte de vitesse attendue et retient la moins coûteuse.

Nous arrivons à la dernière étape du processus. L'agent décide ou non d'exécuter l'action la moins coûteuse à condition que cette dernière soit inférieure à un seuil maximal fixé à 10 km/h. Si à quelque étape, le processus empathique se voit abandonner, l'agent applique les règles de gestion des interactions.

Au premier chapitre, nous avons discuté de l'influence des émotions et des ressentis sur les comportements des conducteurs (1.2.1). Nous souhaitons simuler l'un d'entre eux, l'impatience, pour deux raisons. La première étant que si l'agent effectue un freinage empathique, mais que le véhicule cible ne réagit pas promptement ou n'a simplement pas l'intention de changer de voie, alors il devient nécessaire d'annuler cette manœuvre empathique pour limiter la perturbation sur le trafic. La seconde étant que des conducteurs humains malicieux pourraient entrevoir une faille dans le RC et décider d'en tirer avantage au détriment de la fluidité, voire de la sécurité, du trafic. Par exemple, dans le cas où deux flux de véhicules convergent et qu'un comportement social émerge laissant alternativement la priorité aux deux flux, un RC pourrait céder indéfiniment la priorité à un flux au détriment de l'autre, produisant ainsi un blocage. Pour ces deux raisons, nous simulons l'impatience en diminuant progressivement la SVO de l'agent lorsque ce dernier effectue une action empathique. Nous fixons cette baisse à $\phi_- = \frac{\pi}{36}$ par

seconde.

L'impatience n'est pas l'unique mécanisme capable de réviser une décision tactique. L'agent s'assure de la viabilité de ses décisions antérieures en les réévaluant continuellement. Par exemple, nous considérons que la décision de changer de voie émanant du processus de gestion des interactions doit être maintenue plusieurs secondes avant d'être exécutée. De plus, la voie cible doit présenter un avantage minimum d'un mètre par seconde par rapport à la voie courante de l'agent.

Aussi, à l'instar de MOBIL, tout changement de voie est précédé par une estimation des contraintes (éq. 4.1) subies par le véhicule suiveur de la voie cible $i_{-1}^{\widehat{\ell}_i}$ afin de minimiser l'impact sur le trafic. Cette contrainte s'estime pour la durée totale du changement de voie $t = \hat{t}_i^{\ell_i \to \widehat{\ell}_i}$ de l'agent. La décision finale d'exécuter un changement de voie doit ainsi satisfaire l'inéquation suivante :

$$\overbrace{\cos(\phi_i) \cdot \hat{c}_i^t \left(i \,\middle|\, \hat{t}_i^{\ell_i \to \widehat{\ell}_i} \right)}^{\text{contrainte de } i \text{ sachant le temps requis pour changer de voie}} > \sin(\phi_i) \cdot \underbrace{\hat{c}_i^t \left(i_{-1}^{\widehat{\ell}_i} \,\middle|\, \hat{t}_i^{\ell_i \to \widehat{\ell}_i} \right)}_{\text{contrainte du véhicule suivant } i \ (i_{-1}) \text{ sur la voie cible } \widehat{\ell}_i} \tag{4.5}$$

L'agent peut également révoquer sa décision de changer de voie si : (1) le véhicule cible a déjà réalisé sa manœuvre avec succès, (2) la voie cible n'est plus avantageuse, ou (3) si l'agent n'est plus en interaction. Néanmoins, l'agent poursuivra sa manœuvre s'il se situe déjà entre deux voies. Enfin, le niveau tactique soumet ses préférences (voie, vitesse) au niveau opérationnel en contrôle du véhicule.

4.2.3 Niveau opérationnel

Le niveau opérationnel s'efforce de satisfaire les préférences du niveau tactique tout en considérant les préférences du style de conduite spécifiées à l'initialisation. Ce niveau se fonde sur de l'apprentissage par renforcement (RL) pour piloter le véhicule. Décrivons ses composants, à savoir, l'observation de l'environnement, les actions et fonctions de récompenses.

Une observation comporte 87 entrées regroupées en trois catégories : les préférences à satisfaire, le trafic local et une description des interactions. Nous avons déjà discuté des préférences à plusieurs reprises. Ces dernières incluent une vitesse désirée, un risque maximum toléré, un jerk maximum toléré et une voie cible. Notons que la SVO sert exclusivement au niveau tactique. Les données relatives au trafic local englobent les vitesses longitudinales et latérales des véhicules proches, ainsi que leur distance latérale par rapport à l'agent i. Les interactions décrivent les relations entre paires de véhicules en termes d'intertemps (TH) et de temps avant collision (TTC). L'intertemps correspond au temps avant qu'un premier véhicule rejoigne la position actuelle d'un second véhicule. Le tout constitue une représentation égocentrée du trafic local, similaire aux descriptions de psychologues de la conduite [Espié et Saad, 2000].

Nous représentons le trafic local par un graphe dont les nœuds et les arêtes correspondent respectivement aux véhicules et à leurs interactions (figure 4.5). La représentation par graphe présente l'avantage non négligeable de s'accommoder à divers scénarios de trafic. Si la modélisation du trafic local ne requiert qu'une portion du graphe, ce dernier peut être tronqué et les entrées substituées par des valeurs spécifiques. Par exemple, si aucun véhicule n'occupe la voie située à gauche de l'agent, la description de leur interaction peut fixer des valeurs proches de l'infini pour le TH et le TTC, indiquant ainsi l'absence d'interaction.

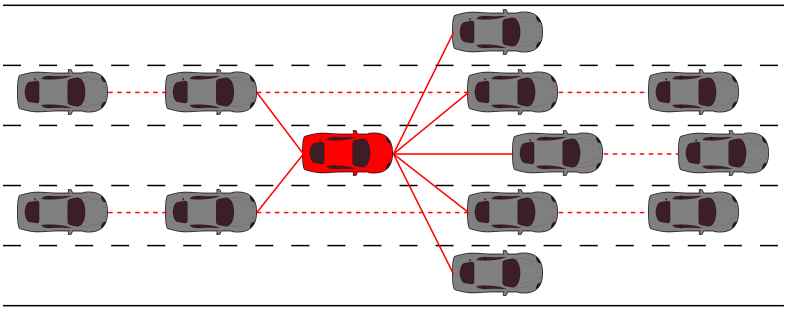


Figure 4.5 – Représentation par graphe du trafic local. Les lignes continues et discontinues représentent les interactions directes et indirectes avec l'agent (véhicule rouge), respectivement.

À partir de ces entrées, la politique de l'agent, fondée sur un réseau de neurones, détermine les accélérations longitudinale $\ddot{x}_i \in [-3;3]$ et latérale $\ddot{y}_i \in [-1;1]$ (en $m.s^{-2}$) à adopter. Par ces deux actions continues, l'agent vise à optimiser son retour u_i^t, conditionné selon quatre fonctions de récompenses quantifiant la satisfaction de préférences en termes de : sécurité $\mathcal{R}_S$, positionnement latéral $\mathcal{R}_T$, vitesse $\mathcal{R}_V$ et confort $\mathcal{R}_C$.

Nous pondérons ces quatre fonctions de récompense comme suit :

$$\mathcal{R}_i = \omega_S \cdot \mathcal{R}_S + \omega_T \cdot \mathcal{R}_T + \omega_V \cdot \mathcal{R}_V + \omega_C \cdot \mathcal{R}_C \tag{4.6}$$

Cette pondération reflète l'étude de Cnossen *et al.* [2000] qui indique que les conducteurs humains priorisent les différents aspects de la conduite selon le même ordre que nous les avons introduits, à savoir : $\omega_S > \omega_T > \omega_V > \omega_C$.

La première fonction de récompense, $\mathcal{R}_S$, optimise l'élément le plus crucial de la conduite, à savoir, la sécurité. Les conducteurs, et plus globalement les passagers, ressentent de l'insécurité lorsque leur perception du risque excède un certain seuil [Kondoh *et al.*, 2008]. Suivant cette description au pied de la lettre, notre fonction de récompense pénalise un agent i qui excèderait son risque maximal toléré $RP_i^{\max}$:

$$\mathcal{R}_S\left(\cdot \mid RP_i^{\max}\right) = 1 - \tanh\left(\sum_{z \in Z} \left(RP_i\left(z\right) - RP_i^{\max}\right)^+\right) \tag{4.7}$$

Le risque peut émaner de diverses sources $z \in Z$ incluant les autres véhicules ou les limites de l'infrastructure (fermeture de voie), longitudinalement comme latéralement. Notons que le seuil de risque maximal dépend du contexte et des contraintes de l'agent. Ainsi, un agent peut provisoirement accepter un risque élevé s'il approche une fin de voie et qu'aucun créneau ne correspond à ses critères de sécurité initiaux. Par exemple, quand un agent se rapproche de la fin d'une rampe d'insertion, et qu'aucun créneau ne se situe à son niveau, il peut décider d'accélérer pour atteindre un créneau sécurisé situé en amont. Dans cette fonction de récompense, comme pour les suivantes, la tangente hyperbolique contraint sous une certaine échelle d'intensité.

La seconde fonction de récompense, $\mathcal{R}_T$, concerne la réalisation de manœuvres latérales. Nous notons que les décisions de se maintenir sur sa voie ou d'en changer induisent toutes deux une position latérale à atteindre. Se maintenir sur sa voie requiert de rester à la même position latérale, le centre de sa voie, tandis que le changement de voie implique de se déporter vers le centre d'une voie adjacente. Ainsi, la fonction de réalisation de manœuvres récompense l'agent selon la distance le séparant de la position latérale cible $\widehat{\ell}_i$ définie au niveau tactique :

$$\mathcal{R}_T\left(\cdot \,\middle|\, \widehat{\ell}_i\right) = 1 - \tanh\left(\frac{\left|\Delta_y\left(i, \widehat{\ell}_i\right)\right|}{w_\ell}\right) \tag{4.8}$$

Cette fonction pénalise la distance latérale Δ_y entre l'agent i et sa voie cible $\widehat{\ell}_i$ proportionnellement à la largeur de voie w_ℓ.

La troisième fonction de récompense, $\mathcal{R}_V$, encourage l'agent à atteindre sa vitesse désirée $\widehat{x}_i$, définie à l'initialisation et éventuellement modulée au niveau tactique :

$$\mathcal{R}_V\left(\cdot \,\middle|\, \widehat{\dot{x}}_i\right) = 2^{-\left|\dot{x}_i - \widehat{\dot{x}}_i\right|} \tag{4.9}$$

Cette fonction applique de lourdes pénalités à l'agent lorsqu'il s'écarte trop de sa vitesse désirée $\widehat{\dot{x}}_i$. Afin d'éviter de punir l'agent lorsque ce dernier ne peut accélérer sans outrepasser son seuil maximal de risque, la vitesse désirée de l'agent est recalculée continuellement pour tenir compte de la vitesse maximale qu'il peut atteindre selon les contraintes imposées par le véhicule qui le précède i_{+1} :

$$\widehat{\dot{x}}_i = \min\left(\widehat{\dot{x}}_i, \frac{1}{5}\left(RP_i^{\max} \cdot \Delta_x\left(i, i_{+1}\right) + 4\dot{x}_{i_{+1}}\right)\right) \tag{4.10}$$

Ainsi, la fonction punit exclusivement la politique opérationnelle en déconsidérant l'impact des décisions tactiques dans le calcul de la récompense. Par exemple, si le niveau tactique décide de se maintenir derrière un véhicule plus lent, la vitesse désirée se calquerait sur celle de ce véhicule et le niveau opérationnel conserverait les mêmes récompenses.

La quatrième et dernière fonction de récompense, $\mathcal{R}_C$, vise à améliorer le confort des passagers. Nous supposons que le confort des conducteurs dépend

principalement du jerk qu'ils ressentent. Chaque individu ressent de l'inconfort si ce jerk ressenti excède un certain seuil. Ainsi, la fonction de récompense du confort pénalise toute manœuvre induisant un jerk $\tilde{j}_i^t$ qui excède un seuil maximal $\tilde{j}_i^{\max}$:

$$\mathcal{R}_C\left(\cdot \mid \tilde{j}_i^{\max}\right) = 1 - \tanh\left(\left(\tilde{j}_i^t - \tilde{j}_i^{\max}\right)^+\right) \tag{4.11}$$

Au-delà du confort ressenti, cette fonction de récompense permet de restreindre les manœuvres brusques et promeut donc une conduite plus *douce*.

En résumé, le modèle Archicool s'inspire de la structure décisionnelle des conducteurs humains telle que rapportée par les psychologues. Le modèle se compose d'un niveau tactique gérant les interactions et d'un niveau opérationnel pilotant le véhicule. Contrairement aux descriptions des psychologues, nous omettons le niveau stratégique de planification du trajet, bien que cette composante puisse être intégrée au modèle sans difficulté majeure par toute personne souhaitant l'étendre.

Le niveau tactique comprend deux processus mutuellement exclusifs fondés sur des bases de règles : la gestion des interactions et l'empathie sélective. La décision d'appliquer l'un ou l'autre dépend de la valeur d'orientation sociale des agents. La gestion des interactions vise à accroître l'utilité de l'agent selon une approche opportuniste, tandis que l'empathie sélective permet d'adopter une approche altruiste.

Contrairement aux approches existantes considérant l'empathie comme une volonté d'améliorer le trafic dans sa globalité, l'empathie sélective se veut plus proche de celle observée chez les conducteurs humains. Ainsi, l'assistance se concentre sur un ensemble de véhicules identifiés comme en ayant probablement besoin.

Le niveau opérationnel satisfait diverses préférences à travers le pilotage du véhicule qui s'effectue par apprentissage par renforcement. Ces préférences dépendent du style de conduite défini à l'initialisation et du niveau tactique.

La séparation des niveaux tactiques et opérationnels permet une prise de décision interprétable, car fondée sur des règles comportementales inspirées de descriptions de psychologues au niveau tactique que le niveau opérationnel s'évertue à appliquer.

Comme nous l'avons décrit au premier chapitre (1.2), les styles de conduite diffèrent entre les régions et les cultures. Considérant cet aspect crucial, propre au trafic humain, notre modèle comporte de nombreux paramètres dont l'ajustement est censé refléter les pratiques locales des conducteurs. Il appartient aux utilisateurs du modèle de le calibrer selon leurs besoins de modélisation.

4.3 Expériences et résultats

Toutes les expériences et les résultats sont effectués avec le simulateur de trafic *Archisim* [Doniec *et al.*, 2008b]. Nos expériences se concentrent sur l'évaluation de l'empathie sélective dans un scénario d'insertion sur autoroute comprenant deux voies principales (figure 4.4). Dans ce scénario, les limites de vitesse de la rampe d'insertion et des voies principales sont fixées à 100 et 130 km/h, respectivement.

Commençons par décrire le protocole d'entraînement du modèle (4.3.1), puis définissons le contexte expérimental et les résultats obtenus (4.3.2).

4.3.1 Entraînement

Nous effectuons l'entraînement avec une *@Nvidia RTX 3080*. Les styles de conduite des agents sont initialisés selon les caractéristiques suivantes :

- vitesse désirée $\widehat{\dot{x}}_i \sim \dot{x}_{\ell_i}^{\max} \times \mathcal{N}\left(1, \frac{1}{25}\right)$ où $\dot{x}_{\ell_i}^{\max}$ correspond à la limite de vitesse de la voie courante
- risque maximal toléré $RP_i^{\max} \sim \mathcal{N}\left(\frac{13}{10}, \frac{3}{20}\right)$
- jerk maximal toléré $\ddot{j}_i^{\max} \sim \mathcal{N}\left(1, \frac{1}{10}\right)$
- valeur d'orientation sociale $\phi_i \sim \mathcal{N}\left(0, \frac{\pi}{10}\right)$

avec $\mathcal{N}$ symbolisant une distribution gaussienne.

L'approche reprend la structure du modèle GENEPI afin d'optimiser le temps d'apprentissage de la population hétérogène. Nous y ajoutons une couche récurrente GRU à l'architecture, motivée par les raisons évoquées au second chapitre. Nous adaptons en conséquence la mémoire tampon pour sauvegarder des séquences de transitions, nécessaires à l'apprentissage d'une couche récurrente. Enfin, nous suivons les recommandations de l'architecture R2D2 afin d'optimiser l'apprentissage de cette couche [Haarnoja *et al.*, 2018; Kapturowski *et al.*, 2018].

Table 4.1 – Hyperparamètres

Facteur de réduction γ	0,95
Taux d'apprentissage α	$3e-4$
Optimiseur	Adam
Taille de la mémoire tampon	$1e+6$
Taille de lot $\lvert\mathcal{B}\rvert$	64

Nous pondérons les fonctions de récompense telles que : $\omega_S = \frac{4}{10}$, $\omega_T = \frac{3}{10}$, $\omega_V = \frac{2}{10}$, $\omega_C = \frac{1}{10}$. Ainsi, la récompense théorique maximale reste dans l'intervalle $\mathcal{R}_i \in [0, 1]$ à chaque pas de temps $\Delta t = \frac{1}{10}$. Sachant qu'un épisode dure au

maximum 45 secondes et se termine prématurément en cas d'accident, tout agent peut théoriquement accumuler une récompense épisodique maximum de 450.

L'apprentissage comme les expériences se déroulent sur un scénario d'insertion sur une autoroute à deux voies principales (figure 4.6). L'insertion des véhicules s'effectue depuis la rampe d'insertion. Conformément à nos propos de la section 4.1.1, l'apprentissage s'effectue par curriculum [Narvekar *et al.*, 2020]. Nous concevons ici une difficulté adaptative à quatre niveaux. Cette difficulté se mesure en termes de densité de trafic des voies principales du scénario d'insertion à l'initialisation. Les quatre niveaux de difficulté se caractérisent donc par un intertemps $\overline{TH}$ moyen entre les véhicules des voies principales :

- *Vide* — aucun véhicule sur les voies principales
- *Léger* — 5 véhicules sur chaque voie principale, $\overline{TH} = 6 \pm \frac{1}{2}$ s
- *Moyen* — 10 véhicules sur chaque voie principale, $\overline{TH} = 3 \pm \frac{1}{2}$ s
- *Dense* — 15 véhicules sur chaque voie principale, $\overline{TH} = 2 \pm \frac{1}{2}$ s

Ainsi, les marges de sécurité se réduisent proportionnellement à l'accroissement de la densité sur les voies principales. Plus cette densité augmente, plus l'insertion devient difficile pour les véhicules de la rampe d'insertion. Durant toutes les phases d'entraînement et d'expérience, 10 agents commencent sur la rampe d'insertion et devront s'insérer sur les voies principales l'un après l'autre. Les agents commencent par apprendre selon le premier niveau de difficulté et accèdent au suivant lorsqu'ils accumulent 75% de la récompense théorique maximale. L'apprentissage se conclut une fois ce seuil atteint au dernier palier de difficulté.

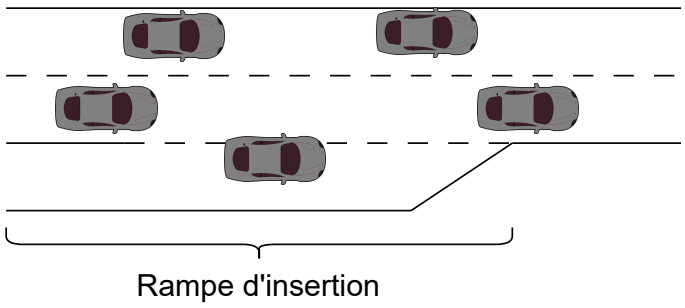


Figure 4.6 – Scénario d'insertion sur autoroute

4.3.2 Scénarios et résultats

L'apprentissage s'effectue approximativement en trois heures (temps utilisateur). Nous conduisons ensuite deux expériences afin d'évaluer le processus d'empathie sélective. La première consiste en une étude d'ablation évaluant l'impact de l'empathie sélective sur la sécurité des véhicules s'insérant. La seconde consiste

à mesurer l'efficacité des changements de voie empathiques en se référant aux journaux de décision émis par le niveau tactique.

Concernant la première expérience, nous définissons deux comportements types pour les véhicules des voies principales :

- *Altruiste* — les agents accordent plus d'importances aux autres usagers qu'à eux-mêmes $\phi_i = \frac{\pi}{2}$
- *Égoïste* — les agents déconsidèrent les autres usagers $\phi_i = 0$

La SVO des agents commençant sur la rampe d'insertion reste inchangée. Nous quantifions l'impact sur le trafic de ces deux comportements types en termes de vitesse de flux et de sécurité selon plusieurs densités de trafic.

Chaque expérience est répétée cinquante fois avec une position initiale des véhicules aléatoire. Au total, cela correspond à 500 insertions sur autoroute. Nous évaluons (1) le nombre d'accidents, (2) le risque perçu par les agents au moment de leur insertion et (3) la vitesse des flux en amont et en aval de la zone d'insertion. Pour le point (2), ce moment correspond à l'instant où le centre du véhicule passe au-dessus des marquages au sol délimitant la voie d'insertion des voies principales. Nous espérons que l'empathie sélective améliorera la sécurité des véhicules s'insérant — cas altruiste — sans dégrader le flux de trafic.

La figure 4.7 illustre les résultats de cette expérience. Elle montre les trajectoires des véhicules s'insérant. Notons que seuls les véhicules s'insérant sont représentés. La couleur de ces trajectoires représente le risque perçu par les agents. Les trajectoires de gauche correspondent au comportement type altruiste, tandis que celles de droite correspondent au comportement type égoïste. De haut en bas, les trajectoires correspondent aux densités de trafic légère, moyenne et dense. Les flèches indiquent la vitesse du flux (en m/s) en amont et en aval de la zone d'insertion.

En analysant ces trajectoires, nous constatons que le risque perçu provient essentiellement de deux facteurs : l'intertemps (TH) et le temps avant collision (TTC). À mesure que les véhicules se rapprochent de la fin de la rampe d'insertion, leur TTC par rapport à cette fin de rampe diminue significativement, augmentant le risque perçu. Ces situations sont mises en exergue par les couleurs les plus chaudes (d'orange à rouge) sur les trajectoires précédant l'insertion. Une fois ces véhicules insérés, le risque perçu diminue fortement. La couleur des trajectoires tend alors vers des couleurs plus froides (du bleu au vert). Ceci signifie que le principal danger du scénario étudié provient de la manœuvre d'insertion en elle-même.

Dans les scénarios avec la plus forte densité (trajectoires du bas), les véhicules éprouvent plus de difficulté à trouver des créneaux sécurisés pour s'insérer sur les voies principales. On observe alors une augmentation du risque perçu, principalement lorsque le comportement type est égoïste (trajectoires en bas à droite), où les agents priorisent leurs récompenses et ne sont pas enclins à aider les véhicules s'insérant. Ce comportement mène parfois à des accidents (trajectoires rouges se finissant prématurément), ce qui constitue un échec de la manœuvre d'insertion.

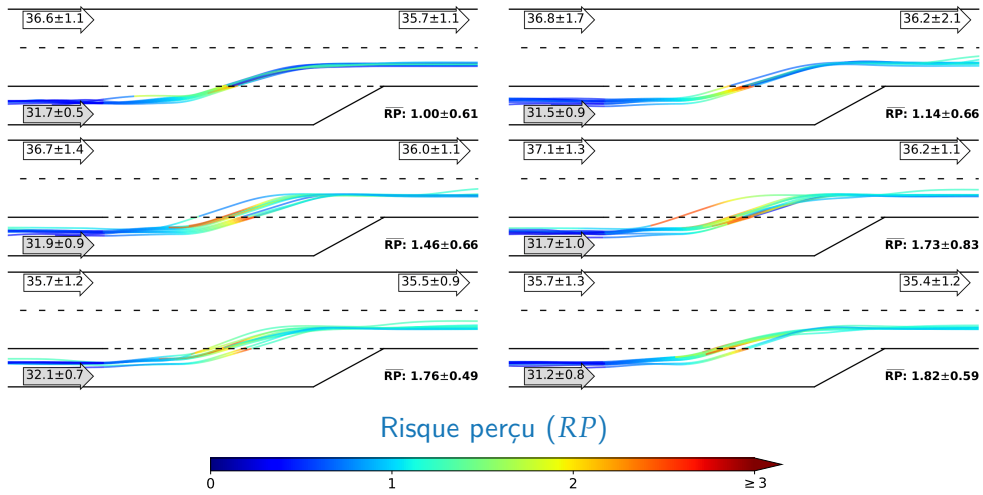


Figure 4.7 – Trajectoires des véhicules s'insérant sur les voies principales. De haut en bas, les densités faible, moyenne et dense. De gauche à droite, agents altruistes et égoïstes. Avec $\overline{RP}$ le risque perçu à l'insertion. Les flèches indiquent le flux de trafic (en m/s) en amont et en aval de la zone d'insertion.

Concernant la fluidité du trafic, l'expérience montre que les vitesses des flux en amont et en aval de la zone d'insertion restent cohérentes par rapport aux observations du monde réel. Une fois encore, il doit être noté que les données du monde réel dépendent en partie du lieu de leur collecte, car les pratiques des conducteurs varient selon les régions. La fusion de deux flux de trafic produit naturellement une légère baisse de la vitesse en aval de la zone d'insertion. Mais aucune différence significative n'est observée entre les comportements types altruiste et égoïste.

La table 4.2 décrit le pourcentage d'accidents selon la densité du trafic et divers comportements types. Nous observons deux tendances. Premièrement, les accidents surviennent plus fréquemment lorsque les conducteurs des voies principales agissent de manière égoïste. Deuxièmement, le taux d'accidents s'accroît avec la densité. Le taux d'accidents est donc directement lié à la difficulté du scénario. On pourrait s'étonner en constatant que le taux d'accidents augmente quand les comportements types deviennent extrêmes (altruiste ou égoïste). Cependant, ces résultats sont aussi observés par Toghi *et al.* [2022]. L'explication est qu'un comportement trop altruiste ($\frac{\pi}{2}$) a plus de chance de perturber le flux de trafic et d'augmenter le risque d'incidents.

La figure 4.8 nous permet de mieux apprécier le risque perçu par les agents au moment de l'insertion pour les comportements types altruiste et égoïste. Dans cette figure, les croix noires dénotent le risque perçu en moyenne. Comparé au comportement type égoïste, l'altruiste permet de diminuer le risque perçu par les véhicules s'insérant de 14%, 16% and 12% pour les densités légère, moyenne et

Table 4.2 – Taux d'accidents (en %)

	Léger	Moyen	Dense
Altruiste	6,13	5,63	6,57
Pro-altruiste ($\frac{3\pi}{8}$)	4,77	3,19	7,08
Prosocial ($\frac{\pi}{4}$)	5,91	5,71	7,34
Pro-égoïste ($\frac{\pi}{8}$)	5,71	4,34	5,77
Égoïste	4,91	7,14	7,16

élevée, respectivement. Logiquement, le risque perçu s'accroît à mesure que le trafic se densifie. On constate à travers cette figure que l'écart-type est élevé. Cela est du à la nature du risque qui croît exponentiellement à mesure que la probabilité d'accident augmente.

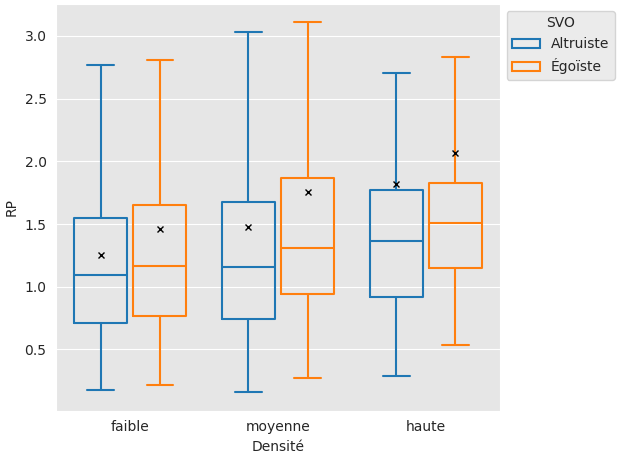


Figure 4.8 – Risque perçu lors de l'insertion

La table 4.3 compare le taux d'accidents de notre modèle Archicool avec les études similaires. Notons qu'il est difficile de tirer une quelconque conclusion à partir de cette table, car les conditions expérimentales divergent entre les études. Certains auteurs étudient l'insertion avec un scénario ne comportant qu'une voie principale, tandis que les autres en considèrent deux. La plupart des auteurs utilisent un espace d'actions discret, des méta-actions comprenant cinq actions (accélérer, ralentir, dépasser, se rabattre, ne rien faire). Le contrôle du véhicule est ainsi facilité, mais les possibilités en termes de comportements sont réduites. C'est pourquoi nous avons opté pour un espace d'action continu, malgré le défi que cela représente. Bien que cette table compare les approches en termes d'accidents, nous pensons que le risque perçu est plus pertinent. En effet, le taux d'accidents est une métrique discrète qui masque les presqu'accidents, tandis que le risque perçu est une métrique continue qui traque toutes les interactions (incidentogènes

ou non). Rappelons surtout que la particularité de notre approche réside dans le fait de réaliser des actions empathiques de manière indépendante, là où les autres approches reposent sur des mécanismes de coordinations intervéhiculaires. Cette particularité vise à rendre possible des actions empathiques de la part de véhicules autonomes lors des premières phases du trafic mixte, où le nombre de ces véhicules sera probablement trop réduit pour permettre des actions coordonnées.

Table 4.3 – Comparaison du taux d'accidents (en %) avec les études existantes

	# Voies	Espace d'action	Altruiste	Égoïste
Archicool	2	Continu	5,63	7,14
Toghi *et al.* [2022]	1	Méta-actions	13,9	41,8
Toghi *et al.* [2021b]	2	Méta-actions	2,6	58,3
Toghi *et al.* [2021a]	2	Méta-actions	16,3	78,2
Valiente *et al.* [2022]	2	Méta-actions	0,0	0,1

Comme attendu, l'empathie sélective d'Archicool atténue le risque des véhicules s'insérant. Indépendamment de la densité du trafic des voies principales, le risque est plus atténué par les agents altruistes que ceux égoïstes. L'empathie sélective s'avère plus efficace lorsque le trafic est modérément dense. Cela suit une certaine logique. Quand la densité est faible, le trafic dispose naturellement de créneaux sécurisés pour les véhicules s'insérant. Les interactions étant plus éparses, les conditions qui déclenchent les actions empathiques sont plus rarement remplies, ce qui minimise l'impact de l'empathie sur le risque perçu. Quand la densité est élevée, les distances de sécurité entre les véhicules sont faibles, ce qui restreint les opportunités d'atténuer le risque. Ainsi, c'est lorsque le trafic est modérément dense que les opportunités empathiques sont les plus nombreuses et que leur effet est le plus important.

Concernant la seconde expérience, nous nous concentrons sur un trafic dense avec des agents égoïstes sur les voies principales. La SVO des agents commençant sur la rampe d'insertion reste inchangée, elle est donc hétérogène, comme lors de la phase d'entraînement. Ces conditions rendent l'insertion des véhicules particulièrement ardue. L'empathie ne peut émaner que des véhicules ayant préalablement réussi à s'insérer, car les véhicules des voies principales sont égoïstes. Les décisions tactiques étant fondées sur des règles, nous les sauvegardons pour chaque agent dans des journaux de décision. Ces journaux servent alors à expliquer les comportements observés *a posteriori*.

Pour cette expérience, nous nous concentrons sur le risque intervéhiculaire et écartons donc de l'affichage tout risque associé à l'infrastructure comme l'approche d'une fin de voie (figure 4.9). Nous observons que deux véhicules rencontrent des difficultés à l'insertion, leur trajectoire étant mise en exergue par les cercles

rouges sur la figure. En examinant les journaux de décision, nous notons que deux agents, s'étant préalablement insérés et ayant un comportement altruiste, ont effectué un changement de voie altruiste pour faciliter l'insertion de ces véhicules. La figure 4.10 illustre ce comportement en proposant un découpage en trois temps. Ces comportements émergents montrent la pertinence de l'empathie sélective, et plus particulièrement du changement de voie empathique, dans les situations critiques où un freinage seul ne peut créer les conditions sécuritaires nécessaires compte tenu des contraintes spatio-temporelles.

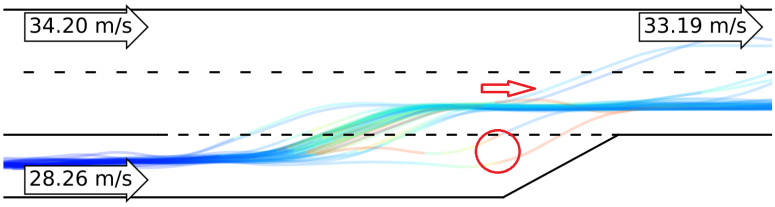


Figure 4.9 – Manœuvres altruistes des véhicules s'insérant. Pour cette figure, le risque perçu par rapport à l'infrastructure est omis.

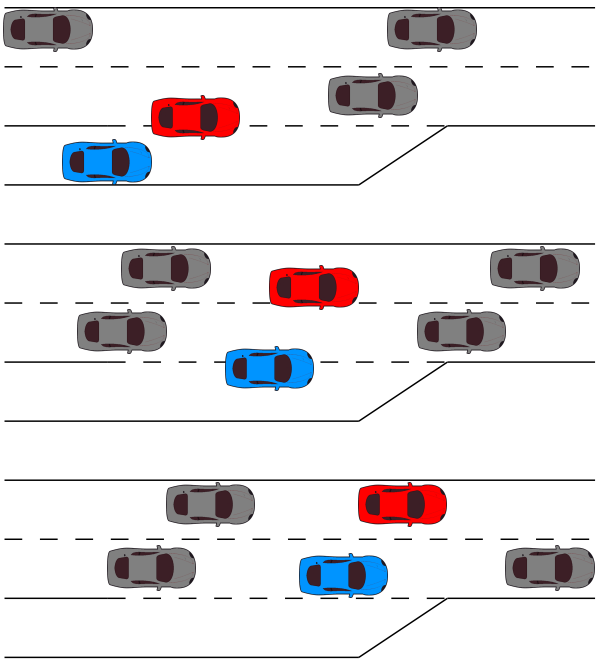


Figure 4.10 – Analyse temporelle des comportements observés (de haut en bas)

D'ailleurs, cet exemple illustre le potentiel des journaux de décision pour

l'explicabilité du comportement des RC. La taille de ces journaux dépend 1) des actions des agents, 2) de la durée de simulation, 3) de la complexité du scénario et 4) des informations que l'on souhaite sauvegarder. L'explicabilité permet de raffiner le modèle lorsque ce dernier produit des comportements indésirables. Notons que la finesse de ces explications résulte essentiellement du fondement à base de règles du niveau tactique. Une approche par apprentissage du niveau tactique n'aurait pas permis de telles explications.

Conclusion

La phase transitoire, où le trafic actuel s'automatiserait graduellement, soulève de nombreux défis relatifs à la cohabitation entre conducteurs humains et RC. Partant du principe que ces difficultés peuvent être atténuées par la minimisation des différences opposant ces deux protagonistes, nous avons proposé un modèle de RC doué de capacités sociales. Au regard des approches existantes sur l'implémentation de la valeur d'orientation sociale dans la conduite, nous avons identifié deux limitations majeures. La première concerne la non-désignation de véhicules cibles pour les actions empathiques, créant un flou décisionnel. La seconde concerne l'absence d'évaluation sur la faisabilité en amont de l'engagement des manœuvres. Archicool, notre modèle de RC, répond à ces deux problématiques. Sa composante SVO, l'empathie sélective, identifie les véhicules souhaitant changer de voie et nécessitant une assistance particulière dans l'exécution de cette manœuvre.

La hiérarchisation du processus décisionnel, calquée sur les études de psychologues de la conduite, participe à clarifier les décisions du modèle. Le niveau tactique, fondé sur des règles, opte pour une gestion des interactions ou pour une action altruiste selon le contexte et la situation. Le niveau tactique définit des préférences à satisfaire par le niveau opérationnel, chargé de la conduite et fondé sur l'apprentissage par renforcement. Le niveau opérationnel veille ainsi à satisfaire une vitesse désirée, à ne pas excéder un seuil de risque ou de jerk, et à manœuvrer latéralement selon les indications du niveau supérieur.

La cohérence des manœuvres, réévaluée continuellement, prévient les comportements aberrants et irrationnels. Par l'introduction d'un mécanisme d'impatience, les actions empathiques, ne se concrétisant pas promptement par la réaction du véhicule cible, sont abrogées.

Le fondement à base de règles du niveau tactique permet de conserver des journaux de décision servant ensuite à expliquer les comportements observés. Ces journaux permettent également de raffiner les comportements indésirables résultants.

Nos résultats suggèrent que l'empathie sélective atténue le risque perçu par des conducteurs s'insérant jusqu'à 16%, sans altérer la vitesse du flux. L'efficacité de l'empathie sélective s'observe davantage dans un trafic modérément dense, car un trafic fluide ne nécessite que rarement des actions empathiques et un trafic

dense restreint les possibilités d'assistance à autrui.

Archicool comprend de nombreux paramètres à calibrer. Ces derniers sont toutefois cruciaux, puisqu'ils permettent d'ajuster le style de conduite des RC à des pratiques locales conformes aux considérations de ses utilisateurs.

Conclusion générale

La volonté politique de s'orienter vers un trafic sans accident tend à promouvoir l'automatisation du trafic routier. Avant de parvenir à une automatisation totale des véhicules, une phase de cohabitation, d'une durée indéterminée, pourrait voir robots de conduite (RC) et conducteurs humains cohabiter. Cette cohabitation soulève de nombreuses questions d'ordre sécuritaire en raison des différences comportementales opposant ces deux acteurs. Notre thèse a visé à atténuer ces différences en proposant un modèle de RC aux inspirations psychologiques et fondé sur l'apprentissage par renforcement.

Dans un premier temps, nous avons étudié ce qui composait le trafic routier et les diverses tentatives de le simuler. Distinguant deux échelles, nous avons constaté que les interactions au niveau microscropique donnaient parfois lieu à des phénomènes émergents au niveau macroscopique. De nombreux travaux ont tenté de reproduire ces phénomènes, et ce, par diverses approches. Les approches par base de règles motivationnelles nous semblaient plus pertinentes que celles par modèles mathématiques ou par base de règles décisionnelles, du fait de leur fondement psychologique.

Nous avons donc orienté la suite de notre étude sur la psychologie. Nous avons relevé deux points. Le premier est que l'une des motivations principales des conducteurs consiste à minimiser leurs interactions avec les autres usagers tout en satisfaisant certains critères de performance tels qu'atteindre une vitesse désirée. Le second concerne la nature hautement hétérogène des comportements observables dans le trafic. Cette hétérogénéité nous est parue comme un défi de taille pour les systèmes de décision des RC, rendant difficilement appréhendables les situations qu'ils rencontreraient quotidiennement.

Les premiers travaux abordant la navigation autonome en trafic mixte ont produit des RC aux comportements axés sur la défensive. Or, en restreignant la marge de manœuvre d'un RC, nous le rendons vulnérable aux conduites opportunistes des conducteurs humains. L'exploitation de cette vulnérabilité peut alors

se transformer en un véritable problème pour la fluidité du trafic, mais surtout pour sa sécurité. La conclusion que nous en avons tirée était que le comportement d'un RC ne devrait pas être entièrement prévisible pour être viable en trafic mixte. Cela impliquait de concevoir des RC adoptant des styles de conduite divers, ce qui ferait l'objet de nos premiers travaux.

En étudiant les travaux actuels sur les RC, nous avons constaté, de la part de certains chercheurs, une volonté d'inclure l'altruisme dans les processus décisionnels. Étonnamment, les travaux existants ne concevaient l'altruisme que sous un angle coopératif entre plusieurs RC, mais jamais d'un point de vue individuel. Cependant, dans l'éventualité d'un trafic mixte, les RC seraient trop peu nombreux pour effectuer des manœuvres coordonnées entre eux. Nous avons donc décidé d'orienter nos seconds travaux vers la conception d'un module d'altruisme, fonctionnel pour un RC non-communiquant.

Une fois nos deux axes de recherche définis, nous nous sommes demandé quelle serait l'approche la plus adaptée. Parmi les options envisagées, nous avons retenu l'apprentissage par renforcement, et ce, pour deux raisons. La première étant que l'apprentissage par renforcement permet de résoudre des problèmes séquentiels de prise de décision dont les objectifs sont renseignés sous la forme de motivations, ce qui s'approche du mode de fonctionnement humain. La seconde est que l'apprentissage par renforcement s'avère fertile à l'émergence. Ainsi, nous pourrions simuler de nouveaux types d'interactions spécifiques au trafic mixte.

Dans un second temps, nous avons donc étudié l'apprentissage par renforcement dans un contexte multiagent (MARL). Nous avons défini un problème MARL par un modèle de jeu auquel on applique un concept de solution. Le modèle de jeu régit les processus d'interaction, tandis que le concept énumère ses solutions possibles. Compte tenu de sa nature hétérogène, aucun concept de solution ne semblait convenir au trafic. En effet, les concepts de solution existants s'appliquent généralement aux jeux strictement définis, tels que les jeux de plateau, et leur transposition vers des problèmes complexes ne semble pas toujours pertinente.

L'étude des algorithmes MARL nous a permis d'apercevoir l'étendue des défis auxquels nous ferions face lors de nos expérimentations. En particulier, le défi posé par la non-stationnarité – l'adaptation continue des agents aux politiques de leurs pairs – nous privait assurément de garanties de convergence. De plus, les problèmes de dimensionnalité des espaces devaient nous restreindre dans le nombre de véhicules simulés simultanément.

Dans un troisième temps, nous avons ainsi souhaité lever le verrou de dimensionnalité et simuler un trafic de plusieurs dizaines de véhicules. Cet objectif devait s'inscrire également dans le cadre de la simulation d'un trafic hétérogène. Notre contribution, le modèle GENEPI, a permis de concilier hétérogénéité comportementale et passage à l'échelle dans un trafic composé de RC. Dans ce modèle, l'hétérogénéité résulte d'un ensemble d'objectifs qui correspondent à des paramètres ajustables par l'utilisateur. Là où l'approche de référence échoue à entraîner

plus de six agents hétérogènes, GENEPI est parvenu à entraîner 90 agents hétérogènes simultanément. L'avantage non négligeable de ce modèle réside dans sa durée d'apprentissage, qui semble constante indépendamment du nombre d'agents entraînés. De plus, nous avons confirmé la capacité de transfert du modèle, soit la capacité de généralisation d'un algorithme d'apprentissage sur des nouveaux objectifs, différents de ceux sur lesquels il a appris, et ce, sans phase d'apprentissage supplémentaire. Avec ce modèle, nous disposons des éléments nécessaires pour simuler un trafic de RC aux traits humains.

Dans un quatrième et dernier temps, nous avons intégré l'altruisme au processus décisionnel du RC en fondant nos recherches sur la notion psychologique de valeur d'orientation sociale (SVO). Pour y parvenir, nous avons fait évoluer l'architecture du modèle en distinguant un niveau tactique et un niveau opérationnel. Cette hiérarchisation des processus décisionnels s'inspire des descriptions de psychologues. Le niveau tactique correspond à la gestion des interactions par les conducteurs, tandis que le niveau opérationnel correspond au pilotage du véhicule. C'est donc au niveau tactique que sont décidées les manœuvres altruistes visant à atténuer le risque perçu par les véhicules s'insérant sur une autoroute. Nos résultats montrent une diminution du risque perçu allant jusqu'à 16% dans un trafic modérément dense lorsque le module d'empathie est activé. L'impact sur la sécurité du trafic est donc positif, tandis que l'impact sur la fluidité est quasi nul.

Évoquons à présent des perspectives de recherche à nos travaux. Nous avons restreint le champ de recherche sur les comportements sociaux au concept d'empathie, mais d'autres attitudes sociales sous-tendent les situations que nous observons quotidiennement dans le trafic. L'une d'entre elles concerne l'imitation. Les conducteurs humains ont souvent tendance, consciemment ou inconsciemment, à imiter, dans une certaine mesure, les comportements de leurs pairs. Par exemple, un conducteur peut s'aligner sur la vitesse du véhicule qui le précède, sans pour autant que cette vitesse soit précisément celle qu'il désire. Un tel comportement participe à stabiliser le trafic et donc à sa fluidité et sa sécurité. Étudier cet aspect de la conduite et le reproduire permettrait aux RC de s'intégrer plus facilement dans un trafic mixte.

Puisque nous parlons de trafic mixte, nous avons restreint nos expériences à la simulation d'un trafic composé exclusivement de RC. Ce choix découle de limites techniques inhérentes à l'approche MARL. En effet, en ajoutant un second type de véhicule simulant la conduite humaine, la complexité de l'apprentissage aurait significativement augmenté, car les RC auraient dû déterminer le type de conducteur – humain ou robot – et, à partir de cette information, anticiper les comportements éventuels. Compte tenu des limites actuelles des approches MARL et souhaitant conserver un modèle – réseau de neurones – relativement simple, nous nous sommes restreints à un trafic de RC. Des travaux futurs pourraient toutefois s'atteler à la simulation d'un trafic mixte.

Pour l'instant, nous avons uniquement considéré les niveaux tactique et opérationnel, mais comme nous l'avons plusieurs fois mentionné, notre modèle est pensé pour accueillir un niveau stratégique. Le niveau stratégique – pour la planification de trajet – pourrait s'avérer utile si les situations étudiées devenaient plus complexes. Par exemple, dans un environnement urbain, le véhicule devrait considérer quelles rues il peut emprunter selon la signalisation.

Enfin, sachant que nos travaux ont été réalisés exclusivement avec un simulateur de trafic, notre modèle pourrait être validé en situation réelle. Cette évaluation du modèle permettrait de mesurer son impact réel sur la sécurité et la fluidité d'un trafic mixte, particulièrement autour des zones d'insertion.

Publications

Nos travaux de recherche ont donné lieu à deux publications :

— En 2022, nous avons publié un article de synthèse portant sur l'apprentissage multiagent pour les RC [Dinneweth *et al.*, 2022].

— En 2023, nous avons présenté à un atelier de la conférence AAMAS un article correspondant à notre première contribution [Dinneweth *et al.*, 2023].

En 2024, nous avons soumis l'article correspondant à la contribution sur l'altruisme.

Bibliographie

Ahmed, H. U., Huang, Y. et Lu, P. (2021). A review of car-following models and modeling tools for human and autonomous-ready driving behaviors in micro-simulation. *Smart Cities*, 4(1):314–335.

Albrecht, S. V., Christianos, F. et Schäfer, L. (2023). Multi-agent reinforcement learning : Foundations and modern approaches. *Massachusetts Institute of Technology : Cambridge, MA, USA*.

Alsheikh, M. A., Hoang, D. T., Niyato, D., Tan, H.-P. et Lin, S. (2015). Markov decision processes with applications in wireless sensor networks : A survey. *IEEE Communications Surveys & Tutorials*, 17(3):1239–1267.

Amari, S.-i. (1993). Backpropagation and stochastic gradient descent method. *Neurocomputing*, 5(4-5):185–196.

Andrychowicz, M., Raichuk, A., Stańczyk, P., Orsini, M., Girgin, S., Marinier, R., Hussenot, L., Geist, M., Pietquin, O., Michalski, M. *et al.* (2021). What matters in on-policy reinforcement learning ? a large-scale empirical study. *In ICLR 2021-Ninth International Conference on Learning Representations*.

Arvin, R., Kamrani, M., Khattak, A. J. et Rios-Torres, J. (2018). Safety impacts of automated vehicles in mixed traffic. Rapport technique, Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (United States).

Aumann, R. J. (1987). Correlated equilibrium as an expression of bayesian rationality. *Econometrica : Journal of the Econometric Society*, pages 1–18.

Badillo, S., Banfai, B., Birzele, F., Davydov, I. I., Hutchinson, L., Kam-Thong, T., Siebourg-Polster, J., Steiert, B. et Zhang, J. D. (2020). An introduction to machine learning. *Clinical pharmacology & therapeutics*, 107(4):871–885.

Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B. et Mordatch, I. (2019). Emergent tool use from multi-agent autocurricula. *arXiv preprint arXiv :1909.07528*.

Bando, M., Hasebe, K., Nakayama, A., Shibata, A. et Sugiyama, Y. (1995). Dynamical model of traffic congestion and numerical simulation. *Physical review E*, 51(2):1035.

Barlow, H. B. (1989). Unsupervised learning. *Neural computation*, 1(3):295–311.

Barth-Maron, G., Hoffman, M. W., Budden, D., Dabney, W., Horgan, D., Dhruva, T., Muldal, A., Heess, N. et Lillicrap, T. (2018). Distributed distributional deterministic policy gradients. *ArXiv*, abs/1804.08617.

Bazzan, A. L. et Klügl, F. (2009). *Multi-agent systems for traffic and transportation engineering*. Information Science Reference.

Bellemare, M. G., Dabney, W. et Munos, R. (2017). A distributional perspective on reinforcement learning. *In Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 449–458. JMLR.org.

Bengio, Y., Louradour, J., Collobert, R. et Weston, J. (2009). Curriculum learning. *In Proceedings of the 26th Annual International Conference on Machine Learning*, ICML'09, page 41–48, New York, NY, USA. Association for Computing Machinery.

Bergemann, D. et Valimaki, J. (2018). *Bandit problems*, pages 665–670. Palgrave Macmillan UK, London.

Bernstein, D. S., Givan, R., Immerman, N. et Zilberstein, S. (2002). The complexity of decentralized control of markov decision processes. *Mathematics of operations research*, 27(4):819–840.

Botvinick, M., Ritter, S., Wang, J. X., Kurth-Nelson, Z., Blundell, C. et Hassabis, D. (2019). Reinforcement learning, fast and slow. *Trends in cognitive sciences*, 23(5):408–422.

Botvinick, M. M., Niv, Y. et Barto, A. G. (2009). Hierarchically organized behavior and its neural foundations : A reinforcement learning perspective. *Cognition*, 113(3):262–280.

Bouha, N., Morvan, G., Abouaïssa, H. et Kubera, Y. (2015). A first step towards dynamic hybrid traffic modeling. *In 29th European Conf. on modelling and simulation (ECMS)*.

Canese, L., Cardarilli, G. C., Di Nunzio, L., Fazzolari, R., Giardino, D., Re, M. et Spanò, S. (2021). Multi-agent reinforcement learning : A review of challenges and applications. *Applied Sciences*, 11(11):4948.

Champion, A., Espié, S. et Mandiau, R. (2002). Let interactions live–how to improve the behavior of simulated drivers approaching a crossroad. *In Proceedings of the Summer Computer Simulation Conference, San Diego, USA*, pages 82–96.

Chao, Q., Bi, H., Li, W., Mao, T., Wang, Z., Lin, M. C. et Deng, Z. (2020). A survey on visual traffic simulation : Models, evaluations, and applications in

autonomous driving. *In Computer Graphics Forum*, volume 39, pages 287–308. Wiley Online Library.

Chen, D., Chen, K., Li, Z., Chu, T., Yao, R., Qiu, F. et Lin, K. (2021). Powernet : Multi-agent deep reinforcement learning for scalable powergrid control. *IEEE Transactions on Power Systems*, 37(2):1007–1017.

Cho, K., van Merriënboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H. et Bengio, Y. (2014). Learning phrase representations using rnn encoder–decoder for statistical machine translation. *In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

Christianos, F., Papoudakis, G., Rahman, M. A. et Albrecht, S. V. (2021). Scaling multi-agent reinforcement learning with selective parameter sharing. *In* Meila, M. et Zhang, T., éditeurs : *Proceedings of the 38th International Conference on Machine Learning*, volume 139 de *Proceedings of Machine Learning Research*, pages 1989–1998. PMLR.

Cnossen, F., Rothengatter, T. et Meijman, T. (2000). Strategic changes in task performance in simulated car driving as an adaptive response to task demands. *Transportation Research Part F : Traffic Psychology and Behaviour*, 3(3):123–140.

Cui, K., Tahir, A., Ekinci, G., Elshamanhory, A., Eich, Y., Li, M. et Koeppl, H. (2022). A survey on large-population systems and scalable multi-agent reinforcement learning. *ArXiv*, abs/2209.03859.

Cunningham, P., Cord, M. et Delany, S. J. (2008). Supervised learning. *In Machine learning techniques for multimedia : case studies on organization and retrieval*, pages 21–49. Springer.

Degris, T., Pilarski, P. M. et Sutton, R. S. (2012). Model-free reinforcement learning with continuous action in practice. *In 2012 American Control Conference (ACC)*, pages 2177–2182. IEEE.

Deschamps, J.-C. (1974). L'attribution, la catégorisation sociale et les représentations intergroupes. *Bulletin de psychologie*, 27(312):710–721.

Dinneweth, J., Boubezoul, A., Mandiau, R. et Espié, S. (2022). Multi-agent reinforcement learning for autonomous vehicles : A survey. *Autonomous Intelligent Systems*, 2(1):27.

Dinneweth, J., Boubezoul, A., Mandiau, R. et Espié, S. (2023). Genepi : a generic parameter-sharing for intrinsically motivated marl agents. *In Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023)*.

Doniec, A., Mandiau, R., Piechowiak, S. et Espié, S. (2008a). Anticipation based on constraint processing in a multi-agent context. *Autonomous Agents and Multi-Agent Systems*, 17:339–361.

Doniec, A., Mandiau, R., Piechowiak, S. et Espié, S. (2008b). A behavioral multi-agent model for road traffic simulation. *Engineering Applications of Artificial Intelligence*, 21(8):1443–1454.

Drogoul, A. et Ferber, J. (2018). Multi-agent simulation as a tool for studying emergent processes in societies. *In Simulating societies*, pages 127–142. Routledge.

Duval, C., Piolino, P., Bejanin, A., Laisney, M., Eustache, F. et Desgranges, B. (2011). La théorie de l'esprit : aspects conceptuels, évaluation et effets de l'âge. *Revue de neuropsychologie*, 3(1):41–51.

El Fallah-Seghrouchni, A., Haddad, S. et Mazouzi, H. (1999). A formal study of interactions in multi-agent systems. *In Proceedins of ISCA International Conference in Computer and their Applications*, CATA'99. Citeseer.

El Hadouaj, S. (2004). *Conception de comportements de résolution de conflits et de coordination : Application à une simulation multi-agent du trafic routier*. Thèse de doctorat, Paris 6.

Elman, J. L. (1993). Learning and development in neural networks : The importance of starting small. *Cognition*, 48(1):71–99.

Espié, S. et Saad, F. (2000). Driver behaviour modelling and traffic simulation. *In Proceedings of the human factors and ergonomics society annual meeting*, volume 44, pages 3–251. SAGE Publications Sage CA : Los Angeles, CA.

Fudenberg, D. et Tirole, J. (1991). *Game theory*. MIT press.

Gazis, D. C., Herman, R. et Rothery, R. W. (1961). Nonlinear follow-the-leader models of traffic flow. *Operations research*, 9(4):545–567.

Greenwald, A. et Jafari, A. (2003). A general class of no-regret learning algorithms and game-theoretic equilibria. *In Learning Theory and Kernel Machines : 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003. Proceedings*, volume 2777, pages 2–12. Springer.

Gronauer, S. et Diepold, K. (2022). Multi-agent deep reinforcement learning : a survey. *Artificial Intelligence Review*, 55(2):895–943.

Grondman, I., Busoniu, L., Lopes, G. A. et Babuska, R. (2012). A survey of actor-critic reinforcement learning : Standard and natural policy gradients. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(6):1291–1307.

Guériau, M. et Dusparic, I. (2020). Quantifying the impact of connected and autonomous vehicles on traffic efficiency and safety in mixed traffic. *In 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, page 1–8. IEEE Press.

Guessoum, Z. (2004). Adaptive agents and multiagent systems. *IEEE Distributed Systems Online*, 5(7).

Gupta, J. K., Egorov, M. et Kochenderfer, M. (2017). Cooperative multi-agent control using deep reinforcement learning. *In International conference on autonomous agents and multiagent systems*, pages 66–83. Springer.

Haarnoja, T., Zhou, A., Abbeel, P. et Levine, S. (2018). Soft actor-critic : Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *In International conference on machine learning (ICML)*, pages 1861–1870. PMLR.

Hamila, M. A. (2012). *Planification multi-agents dans un cadre markovien : les jeux stochastiques à somme générale*. Thèse de doctorat, Université de Valenciennes et du Hainaut-Cambresis.

Hanin, B. (2018). Which neural net architectures give rise to exploding and vanishing gradients ? *In Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 580–589, Red Hook, NY, USA. Curran Associates Inc.

Hausknecht, M., Stone, P. et Mc, O.-p. (2016). On-policy vs. off-policy updates for deep reinforcement learning. *In Deep reinforcement learning : frontiers and challenges, IJCAI 2016 Workshop*. AAAI Press New York, NY, USA.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of chemical information and computer sciences*, 44(1):1–12.

Hernandez-Leal, P., Kaisers, M., Baarslag, T. et De Cote, E. M. (2017). A survey of learning in multiagent environments : Dealing with non-stationarity. *arXiv preprint arXiv :1707.09183*.

Hochreiter, S. et Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hoogendoorn, S. et Knoop, V. (2013). Traffic flow theory and modelling. *The transport system and transport policy : an introduction*, pages 125–159.

Horgan, D., Quan, J., Budden, D., Barth-Maron, G., Hessel, M., Van Hasselt, H. et Silver, D. (2018). Distributed prioritized experience replay. *arXiv preprint arXiv :1803.00933*.

Kapturowski, S., Ostrovski, G., Quan, J., Munos, R. et Dabney, W. (2018). Recurrent experience replay in distributed reinforcement learning. *In International conference on learning representations (ICLR)*.

Karpe, M., Fang, J., Ma, Z. et Wang, C. (2021). Multi-agent reinforcement learning in a realistic limit order book market simulation. *In Proceedings of the first ACM international conference on AI in finance*, ICAIF'20, pages 1–7, New York, NY, USA. Association for Computing Machinery.

Kaushik, M., Singhania, N., S., P. et Krishna, K. M. (2020). Parameter sharing reinforcement learning architecture for multi agent driving. *In Proceedings of the 2019 4th International Conference on Advances in Robotics*, AIR'19, New York, NY, USA. Association for Computing Machinery.

Kesting, A., Treiber, M. et Helbing, D. (2007). General lane-changing model mobil for car-following models. *Transportation Research Record*, 1999(1):86–94.

Kesting, A., Treiber, M. et Helbing, D. (2010). Enhanced intelligent driver model to access the impact of driving strategies on traffic capacity. *Philosophical Transactions of the Royal Society A : Mathematical, Physical and Engineering Sciences*, 368(1928):4585–4605.

Kondoh, T., Yamamura, T., Kitazaki, S., Kuge, N. et Boer, E. R. (2008). Identification of visual cues and quantification of drivers' perception of proximity risk to the lead vehicle in car-following situations. *Journal of Mechanical Systems for Transportation and Logistics*, 1(2):170–180.

Kreps, D. M. (1989). *Nash equilibrium*, pages 167–177. Palgrave Macmillan UK.

Ksontini, F., Mandiau, R., Guessoum, Z. et Espié, S. (2015). Affordance-based agent model for road traffic simulation. *Autonomous Agents and Multi-Agent Systems*, 29(5):821–849.

Kubera, Y., Mathieu, P. et Picault, S. (2010). Everything can be Agent ! Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2010), pages 1547–1548, Toronto, Canada.

Lapan, M. (2020). *Deep Reinforcement Learning Hands-On - Second Edition*. Packt.

Levy, A., Konidaris, G., Platt, R. et Saenko, K. (2017). Learning multi-level hierarchies with hindsight. *arXiv preprint arXiv :1712.00948*.

Li, D., Pan, H., Xiao, Y., Li, B., Chen, L., Li, H. et Lyu, H. (2022). Social-aware decision algorithm for on-ramp merging based on level-k gaming. *In 2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)*, page 1753–1758. IEEE Press.

Lighthill, M. J. et Whitham, G. B. (1955). On kinematic waves i. flood movement in long rivers. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 229(1178):281–316.

Littman, M. L. (1994). Markov games as a framework for multi-agent reinforcement learning. *In Proceedings of the Eleventh International Conference on International Conference on Machine Learning*, ICML'94, page 157–163. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Liu, B., Ding, Z. et Lv, C. (2020). Platoon control of connected autonomous vehicles : A distributed reinforcement learning method by consensus. *IFAC-PapersOnLine*, 53(2):15241–15246.

Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P. et Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. *In Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6382–6393, Red Hook, NY, USA. Curran Associates Inc.

Mandiau, R., Champion, A., Auberlet, J.-M., Espié, S. et Kolski, C. (2008). Behaviour based on decision matrices for a coordination between agents in a urban traffic simulation. *Applied Intelligence*, 28:121–138.

McDowd, J. M. (2007). An overview of attention : behavior and brain. *Journal of Neurologic Physical Therapy*, 31(3):98–103.

McKenzie, M. C. et McDonnell, M. D. (2022). Modern value based reinforcement learning : A chronological review. *IEEE Access*, 10:134704–134725.

Michon, J. A. (1985). *A critical view of driver behavior models : what do we know, what should we do ?*, pages 485–524. Springer US.

Miller, T. (2019). Explanation in artificial intelligence : Insights from the social sciences. *Artificial intelligence*, 267:1–38.

Mitchell, T. M. (1980). The need for biases in learning generalizations. *Rutgers University*.

Mitroi, I.-S., Ciobîcă, A.-M. et Popa, M. (2016). Car-following models comparison between models used by vissim and aimsun. *UPB Scientific Bulletin Series D*, 78:71–82.

Moerland, T. M., Broekens, J., Plaat, A., Jonker, C. M. *et al.* (2023). Model-based reinforcement learning : A survey. *Foundations and Trends in Machine Learning*, 16(1):1–118.

Moridpour, S., Sarvi, M. et Rose, G. (2010). Lane changing models : a critical review. *Transportation letters*, 2(3):157–173.

Munduteguy, C. (2001). *Reconnaissance d'intention et prédiction d'action pour la gestion des interactions en environnement dynamique*. Thèse de doctorat, Paris, CNAM.

Munduteguy, C. et Darses, F. (2007). Perception et anticipation du comportement d'autrui en situation simulée de conduite automobile. *Le travail humain*, 70(1):1–32.

Munos, R., Stepleton, T., Harutyunyan, A. et Bellemare, M. G. (2016). Safe and efficient off-policy reinforcement learning. *In Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 1054–1062, Red Hook, NY, USA. Curran Associates Inc.

Murphy, R. O., Ackermann, K. A. et Handgraaf, M. J. (2011). Measuring social value orientation. *Judgment and Decision making*, 6(8):771–781.

Näätänen, R. et Summala, H. (1974). A model for the role of motivational factors in drivers' decision-making. *Accident Analysis & Prevention*, 6(3-4):243–261.

Nachum, O., Gu, S., Lee, H. et Levine, S. (2018). Data-efficient hierarchical reinforcement learning. *In Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 3307–3317, Red Hook, NY, USA. Curran Associates Inc.

Nagel, K. et Schreckenberg, M. (1992). A cellular automaton model for freeway traffic. *Journal de physique I*, 2(12):2221–2229.

Naik, A., Shariff, R., Yasui, N., Yao, H. et Sutton, R. S. (2019). Discounted reinforcement learning is not an optimization problem. *arXiv preprint arXiv :1910.02140*.

Narvekar, S., Peng, B., Leonetti, M., Sinapov, J., Taylor, M. E. et Stone, P. (2020). Curriculum learning for reinforcement learning domains : A framework and survey. *The Journal of Machine Learning Research*, 21(1):7382–7431.

Nash Jr, J. F. (1950). Equilibrium points in n-person games. *Proceedings of the national academy of sciences*, 36(1):48–49.

Nguyen, T. T., Nguyen, N. D. et Nahavandi, S. (2020). Deep reinforcement learning for multiagent systems : A review of challenges, solutions, and applications. *IEEE transactions on cybernetics*, 50(9):3826–3839.

Nordfjærn, T., Jørgensen, S. et Rundmo, T. (2011). A cross-cultural comparison of road traffic risk perceptions, attitudes towards traffic safety and driver behaviour. *Journal of Risk Research*, 14(6):657–684.

Onieva, E., Hernandez-Jayo, U., Osaba, E., Perallos, A. et Zhang, X. (2015). A multi-objective evolutionary algorithm for the tuning of fuzzy rule bases for uncoordinated intersections in autonomous driving. *Inf. Sci.*, 321(C):14–30.

Oroojlooy, A. et Hajinezhad, D. (2023). A review of cooperative multi-agent deep reinforcement learning. *Applied Intelligence*, 53(11):13677–13722.

Özkan, T., Lajunen, T., Chliaoutakis, J. E., Parker, D. et Summala, H. (2006). Cross-cultural differences in driving behaviours : A comparison of six countries. *Transportation research part F : traffic psychology and behaviour*, 9(3):227–242.

Pateria, S., Subagdja, B., Tan, A.-h. et Quek, C. (2021). Hierarchical reinforcement learning : A comprehensive survey. *ACM Computing Surveys (CSUR)*, 54(5):1–35.

Payne, H. J. (1973). Freeway traffic control and surveillance model. *Transportation Engineering Journal of ASCE*, 99(4):767–783.

Perrusquía, A., Yu, W. et Li, X. (2021). Multi-agent reinforcement learning for redundant robot control in task-space. *International Journal of Machine Learning and Cybernetics*, 12:231–241.

Petrović, Đ., Mijailović, R. et Pešić, D. (2020). Traffic accidents with autonomous vehicles : type of collisions, manoeuvres and errors of conventional vehicles' drivers. *Transportation research procedia*, 45:161–168.

Picard, G., Hübner, J. F., Boissier, O. et Gleizes, M.-P. (2009). Reorganisation and self-organisation in multi-agent systems. *In 1st International Workshop on Organizational Modeling, ORGMOD*, pages 66–80.

Pipes, L. A. (1953). An operational analysis of traffic dynamics. *Journal of applied physics*, 24(3):274–281.

Popescu, M.-C., Balas, V. E., Perescu-Popescu, L. et Mastorakis, N. (2009). Multilayer perceptron and neural networks. *WSEAS Transactions on Circuits and Systems*, 8(7):579–588.

Puterman, M. L. (1994). *Markov Decision Processes : Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., USA, 1st édition.

Regan, M. A., Hallett, C. et Gordon, C. P. (2011). Driver distraction and driver inattention : Definition, relationship and taxonomy. *Accident Analysis & Prevention*, 43(5):1771–1781.

Saad, F. (2006). Some critical issues when studying behavioural adaptations to new driver support systems. *Cognition, Technology & Work*, 8(3):175–181.

Saad, F. et Mundutéguy, C. (2002). *Interaction management between car drivers and new driver support system*, pages 841–848. ASCE Library.

Sadigh, D., Landolfi, N., Sastry, S. S., Seshia, S. A. et Dragan, A. D. (2018). Planning for cars that coordinate with people : leveraging effects on human actions for planning and active information gathering over human internal state. *Auton. Robots*, 42(7):1405–1426.

Salehinejad, H., Sankar, S., Barfett, J., Colak, E. et Valaee, S. (2017). Recent advances in recurrent neural networks. *arXiv preprint arXiv :1801.01078*.

Schaul, T., Horgan, D., Gregor, K. et Silver, D. (2015). Universal value function approximators. *In Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 1312–1320. JMLR.org.

Schmidt, L. M., Brosig, J., Plinge, A., Eskofier, B. M. et Mutschler, C. (2022). An introduction to multi-agent reinforcement learning and review of its application to autonomous mobility. *In 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1342–1349. IEEE.

Schulze, T. et Fliess, T. (1997). Urban traffic simulation with psycho-physical vehicle-following models. *In Proceedings of the 29th Conference on Winter Simulation*, WSC'97, page 1222–1229, USA. IEEE Computer Society.

Schwarting, W., Pierson, A., Alonso-Mora, J., Karaman, S. et Rus, D. (2019). Social behavior for autonomous vehicles. *Proceedings of the National Academy of Sciences*, 116(50):24972–24978.

Sewak, M. (2019). Policy-based reinforcement learning approaches. *In Deep Reinforcement Learning*, pages 127–140. Springer.

Shapley, L. S. (1953). Stochastic games. *Proceedings of the national academy of sciences*, 39(10):1095–1100.

Shoham, Y. et Leyton-Brown, K. (2008). *Multiagent Systems : Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press, USA.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A. *et al.* (2017). Mastering the game of go without human knowledge. *nature*, 550(7676):354–359.

Sinha, S., Bharadhwaj, H., Srinivas, A. et Garg, A. (2020). D2rl : Deep dense architectures in reinforcement learning. *arXiv preprint arXiv :2010.09163*.

Stiglitz, J. E. (1981). Pareto optimality and competition. *The Journal of Finance*, 36(2):235–251.

Sun, H., Ge, Y. et Qu, W. (2024). Greater prosociality toward other human drivers than autonomous vehicles : Human drivers' discriminatory behavior in mixed traffic. *Accident Analysis & Prevention*, 203:107623.

Sutton, R. S. (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. *In* Porter, B. et Mooney, R., éditeurs : *Machine learning proceedings 1990*, pages 216–224. Morgan Kaufmann, San Francisco (CA).

Sutton, R. S. et Barto, A. G. (2018). *Reinforcement learning : An introduction*. A Bradford Book, Cambridge, MA, USA.

Takahashi, M. (1964). Equilibrium points of stochastic non-cooperative n-person games. *Journal of Science of the Hiroshima University, Series AI (Mathematics)*, 28(1):95–99.

Toghi, B., Valiente, R., Sadigh, D., Pedarsani, R. et Fallah, Y. P. (2021a). Altruistic maneuver planning for cooperative autonomous vehicles using multi-agent advantage actor-critic. *arXiv preprint arXiv :2107.05664*.

Toghi, B., Valiente, R., Sadigh, D., Pedarsani, R. et Fallah, Y. P. (2021b). Cooperative autonomous vehicles that sympathize with human drivers. *In 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, page 4517–4524. IEEE Press.

Toghi, B., Valiente, R., Sadigh, D., Pedarsani, R. et Fallah, Y. P. (2022). Social coordination and altruism in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 23(12):24791–24804.

Trannois, H., Lebrun, A. et Deleage, J.-L. (1998). *Utilisation de l'intelligence artificielle distribuée pour la simulation microscopique d'un carrefour*. Thèse de doctorat.

Treiber, M., Hennecke, A. et Helbing, D. (2000). Congested traffic states in empirical observations and microscopic simulations. *Physical review E*, 62(2): 1805.

Treuil, J.-P., Drogoul, A. et Zucker, J.-D. (2008). *Modélisation et simulation à base d'agents*. Dunod.

Trommer, S., Kolarova, V., Fraedrich, E., Kröger, L., Kickhöfer, B., Kuhnimhof, T., Lenz, B. et Phleps, P. (2016). The impact of vehicle automation on mobility behaviour. *Autonomous Driving*, 94.

Valiente, R., Toghi, B., Pedarsani, R. et Fallah, Y. P. (2022). Robustness and adaptability of reinforcement learning-based cooperative autonomous driving in mixed-autonomy traffic. *IEEE Open Journal of Intelligent Transportation Systems*, 3:397–410.

Van Seijen, H., Fatemi, M., Romoff, J., Laroche, R., Barnes, T. et Tsang, J. (2017). Hybrid reward architecture for reinforcement learning. *In Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 5398–5408, Red Hook, NY, USA. Curran Associates Inc.

Vander Werf, J., Shladover, S. E., Miller, M. A. et Kourjanskaia, N. (2002). Effects of adaptive cruise control systems on highway traffic flow capacity. *Transportation Research Record*, 1800(1):78–84.

Vasirani, M. et Ossowski, S. (2012). A market-inspired approach for intersection management in urban road traffic networks. *J. Artif. Int. Res.*, 43(1):621–659.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P. *et al.* (2019). Grandmaster level in starcraft 2 using multi-agent reinforcement learning. *Nature*, 575(7782):350–354.

Warner, H. W., Özkan, T. et Lajunen, T. (2009). Cross-cultural differences in drivers' speed choice. *Accident Analysis & Prevention*, 41(4):816–819.

Yang, N., Ding, B., Shi, P. et Feng, D. (2022). Improving scalability of multi-agent reinforcement learning with parameters sharing. *In 2022 IEEE International Conference on Joint Cloud Computing*, JCC'22, pages 37–42. IEEE.

Young, M. S., Mahfoud, J. M., Stanton, N. A., Salmon, P. M., Jenkins, D. P. et Walker, G. H. (2009). Conflicts of interest : The implications of roadside advertising for driver attention. *Transportation research part F : traffic psychology and behaviour*, 12(5):381–388.

Zhang, K., Yang, Z. et Başar, T. (2021). Multi-agent reinforcement learning : A selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, pages 321–384.