



**HAL**  
open science

# Vers une Conception et une Certification d'un Système de Décision Obtenu par Apprentissage par Renforcement Profond

Christophe Bohn, Rezzoug Mehdi, Jaafra Yesmina, Adjed Faouzi, Pelliccia Frédéric, Habib Lydia

## ► To cite this version:

Christophe Bohn, Rezzoug Mehdi, Jaafra Yesmina, Adjed Faouzi, Pelliccia Frédéric, et al.. Vers une Conception et une Certification d'un Système de Décision Obtenu par Apprentissage par Renforcement Profond. Congrès Lambda Mu 23 " Innovations et maîtrise des risques pour un avenir durable " - 23e Congrès de Maîtrise des Risques et de Sûreté de Fonctionnement, Institut pour la Maîtrise des Risques, Oct 2022, Paris Saclay, France. hal-03878440

**HAL Id: hal-03878440**

**<https://hal.science/hal-03878440>**

Submitted on 29 Nov 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



# Vers une Conception et une Certification d'un Système de Décision Obtenu par Apprentissage par Renforcement Profond

Towards Conception and Certification of a Decision System Based on Deep Reinforcement Learning

BOHN Christophe  
IRT-SystemX

2 boulevard Thomas Gobert, 91120  
Palaiseau, France  
christophe.bohn@irt-systemx.fr

JAAFRA Yesmina  
Expleo

3 Avenue des Pres, 78180 Montigny-le-  
Bretonneux, France  
yesmina.jaafra@expleogroup.com

PELLICCIA Frédéric  
Airbus Protect

36 rue Raymond Grimaud, 31700  
Blagnac, France  
frederic.pelliccia@airbus.com

REZZOUG Mehdi  
IRT-SystemX

2 boulevard Thomas Gobert, 91120  
Palaiseau, France  
mehdi.rezzoug@irt-systemx.fr

ADJED Faouzi  
IRT-SystemX

2 boulevard Thomas Gobert, 91120  
Palaiseau, France  
faouzi.adjed@irt-systemx.fr

HABIB Lydia  
Airbus Protect

4 rue marceau, 92130 Issy-les-  
Moulineaux  
lydia.habib@airbus.com

**Résumé** — Nous avons travaillé sur un agent d'intelligence artificielle pour le changement de voie, l'accélération et le freinage dans un environnement autoroutier simulé. Nous avons proposé une approche de conception basée sur les notions de marge de sécurité dans la fonction de récompense de l'agent. Les résultats de l'apprentissage nous ont permis d'améliorer significativement les performances de l'agent d'un facteur 10 tout en maintenant une vitesse supérieure au flux de circulation. Les méthodes mises en œuvre démontrent la complémentarité entre apprentissage et évaluation pour la conception d'un agent de décision sûr. Nous avons complété cette approche par la méthode des attaques adverses afin d'améliorer la robustesse de l'agent. Ces éléments constituent une première étape dans la conception d'un système autonome sûr.

**Mots-clés** — *Conception fonctionnelle sûre, Conduite autonome, Apprentissage profond par renforcement, Evaluation de performances*

**Abstract**— We have worked on an artificial intelligence agent for lane change, acceleration and braking in a simulated highway environment. We proposed a design approach based on the notion of safety margin in the reward function of the agent. The learning results allowed us to significantly improve the agent's performance by a factor of 10 while maintaining a speed above the traffic flow. The implemented methods demonstrate the complementarity between learning and evaluation for the design of a safe decision agent. We have completed this approach with the adversarial attack method in order to guarantee the robustness of the agent. These elements constitute a first step in the design of a safe autonomous system.

**Keywords** — *Safe functional design, Autonomous driving, Deep reinforcement learning, Evaluation of performance*

## I. INTRODUCTION

L'apprentissage par renforcement (« Reinforcement Learning », RL) est un cadre de modélisation général de l'Intelligence Artificielle (IA) qui permet de déplacer l'attention des méthodes d'apprentissage automatique (« Machine Learning », ML) de la reconnaissance de patterns vers la prise de décision séquentielle fondée sur l'expérience [1]. Avec l'avènement des réseaux de neurones profonds (« Deep Neural Networks », DNN) comme approximateurs universels de fonctions, l'apprentissage par renforcement profond (« Deep Reinforcement Learning », DRL) a été appliqué aux systèmes cyber-physiques, où l'espace d'exploration devient insoluble pour les algorithmes tabulaires. Ces entités physiques complexes en interaction avec le monde réel sont progressivement étudiés pour automatiser des services courants, tels que la conduite de véhicules routiers, l'avionique et l'assistance domestique. Compte tenu de l'implication d'équipements coûteux et de la sécurité humaine, le déploiement d'algorithmes DRL nécessite des considérations juridiques, éthiques et de sécurité. Par conséquent, le comportement des politiques DNN doit être rigoureusement évalué avant leur déploiement afin de prévenir les situations à risque.

Dans l'état de l'art, il a été démontré que les DNN ne sont pas robustes à des petites perturbations, nommées attaques adverses [2]. En effet, de telles perturbations dans l'espace d'entrée des DNN peuvent entraîner une variation significative de la sortie prédite. Ce problème, bien connu dans la littérature sur les algorithmes de classification, a récemment été étendu aux politiques de réseau de neurones générées par des algorithmes DRL [3]. Ainsi, le DRL reste vulnérable aux attaques adverses, où une perturbation imperceptible injectée dans l'entrée du réseau provoque des

incohérences dans le comportement de l'algorithme. En effet, lorsqu'un agent RL obtient son état actuel via des observations, ces dernières peuvent contenir des incertitudes dues à l'imprécision des capteurs ou à des perturbations malveillantes. Une politique non robuste à ces incertitudes peut conduire à des décisions aux conséquences catastrophiques.

Les attaques adverses sont pratiquement indétectables avec les métriques standards, soulignant les limites des approches d'évaluation telles que la récompense globale ou le taux de réussite sur l'ensemble des épisodes. Ainsi, le manque de garanties dans les approches DRL développées représente un point bloquant pour l'extension de ce paradigme aux applications du monde réel, en particulier lorsque le contexte légal exige des preuves de sécurité, comme dans le cas des véhicules routiers ou des aéronefs. Dans de telles conditions, les tests seuls via un déploiement réel ou une simulation ne sont plus considérés comme suffisants et doivent être complétés par une forme plus robuste de vérification.

Conscients de ces contraintes, nous considérons dans cet article le problème de la robustesse des agents DRL et de l'évaluation formelle de leur sûreté dans des applications de contrôle de systèmes autonomes. A cette fin, nous proposons les contributions suivantes :

- Le développement d'une fonction de récompense sûre (safe) pour façonner et guider l'entraînement des politiques RL. Les expressions de récompense ainsi améliorées prennent en compte les situations dangereuses de circulation.
- L'entraînement d'un agent via un algorithme DRL en présence d'attaques adverses générées par des techniques de l'état de l'art adaptées aux systèmes RL.
- L'évaluation de ces agents dans le cadre d'une étude de cas sur des véhicules autonomes équipés de politiques DNN qui traitent les observations reçues de l'environnement pour produire des actions de contrôle.

Nous décrivons tout d'abord l'environnement dans lequel nous avons réalisé nos travaux dans la section II. Puis nous présentons les approches développées sur la conception de la fonction de récompense dans la section III. Nous présentons ensuite nos travaux sur les attaques adverses dans la section IV. Enfin la discussion et la conclusion suivent dans les sections V et VI.

## II. ENVIRONNEMENT DE SIMULATION

L'environnement de simulation en 2D open-source pour conduite autonome « HighwayEnv-v0 » [4] est utilisé (Fig. 1). Dans cet environnement, un agent conduit un véhicule (i.e. ego-véhicule) sur une autoroute unidirectionnelle infinie à 4 voies. Ce dernier est inséré dans un flux de circulation comprenant d'autres véhicules (i.e. exo-véhicules) qui suivent un algorithme de conduite intelligent [5] avec une vitesse maximale de 110 km/h.

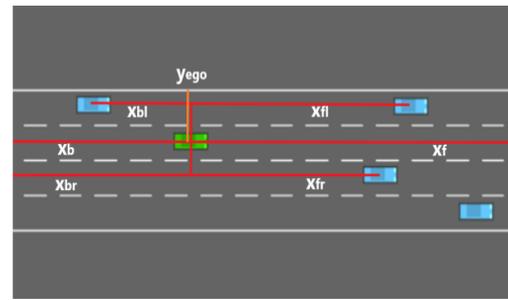


Fig. 1. Représentation de l'environnement HighwayEnv-v0 avec l'ego-véhicule (vert) et des exo-véhicules (bleus).

La mission de l'agent consiste à conduire l'ego-véhicule dans le flux de circulation aussi vite et aussi longtemps que possible, à la vitesse maximum de 130 km/h. Pour obtenir une récompense maximale (cf. section III) l'agent doit dépasser pour maintenir sa vitesse ou freiner afin d'éviter les accidents. L'épisode se termine lorsque l'agent entre en collision avec un autre véhicule ou atteint une limite de temps prédéfinie. L'ego-véhicule peut être contrôlé à l'aide d'un ensemble discret de cinq décisions tactiques mises en œuvre par des contrôleurs de bas niveau : ne rien faire (rester sur la même voie, à la même vitesse), changement de voie à droite ou à gauche, accélération ou décélération sur la même voie.

Le paramètre utilisé pour collecter les états de l'environnement comprend la position transversale de l'agent «  $y_{ego}$  » et de sa vitesse «  $v_{ego}$  ». Afin de tenir compte des interactions avec les autres véhicules, nous utilisons des données relatives (distance et vitesse) des exo-véhicules autour de l'agent. Seuls les véhicules les plus proches de l'ego-véhicule devant et derrière, dans sa voie ou dans les voies directement adjacentes sont pris en compte.

L'agent de conduite utilisé pour ces travaux a un comportement discret : il observe son environnement et prend une décision à chaque pas de temps d'une seconde. Il faut attendre le pas de temps suivant pour prendre la décision suivante en fonction des nouvelles observations.

La configuration initiale de l'environnement est définie par une faible densité de trafic autour du l'ego-véhicule. Dans des travaux antérieurs [6], il a été démontré que l'apprentissage était plus performant dans un environnement plus complexe en termes de densité de trafic. Ici, nous avons donc choisi un environnement avec une densité de trafic élevée pour l'apprentissage et trois niveaux de densité de trafic pour les tests : la même densité que celle prise pour l'apprentissage, une densité plus faible et une densité plus élevée.

Plus spécifiquement, la densité du trafic reflète l'espacement entre les exo-véhicules.  $N$  unités de densité revient à établir un espacement équivalent à la distance parcourue chaque seconde par l'exo-véhicule avec sa vitesse maximale d'initialisation, divisée par  $N$ .

Nous avons implémenté trois agents de contrôle de l'ego-véhicule chacun étant entraîné par une fonction de récompense différente. Ces fonctions de récompenses sont caractérisées respectivement en termes de vitesse ( $V$ ), vitesse-accident ( $VA$ ) et sécurité collisions ( $SC$ ). Les entraînements de l'agent avec chacune de ces fonctions sont réalisés avec trois densités de trafic d'exo-véhicules différentes : une densité complexe basse ( $DC^- = 2.5$ ), une densité complexe ( $DC = 3.0$ ) et une densité complexe haute ( $DC^+ = 3.3$ ). Les fonctions de récompense et les résultats des entraînements sont présentés dans les sections suivantes. Avec la définition

du paragraphe précédent, la densité de trafic DC- correspond à un espacement des exo-véhicules de 10 mètres, DC de 8,3 mètres et DC+ de 7,6 mètres.

### III. NOUVELLE CONCEPTION DE LA FONCTION DE RECOMPENSE (REWARD SHAPING)

Nous basons notre étude sur le RL qui est une approche fondamentale de l'optimisation orientée objectifs, inspirée de la psychologie comportementaliste [7]. L'élément de base du RL est un agent qui apprend par interaction avec son environnement guidé par un signal d'impact appelé récompense. Le retour de l'environnement conduit l'agent à choisir une meilleure action améliorant le processus d'apprentissage, d'où le nom d'apprentissage par renforcement. La formulation mathématique du RL est dérivée du processus de décision de Markov (Markov Decision Process, MDP) en termes d'état, d'action, de récompense et de dynamique du système. A chaque pas de temps, l'agent observe l'état actuel  $s_t$  et exécute une action  $a_t$  basée sur sa politique courante  $\pi$ . Une fois l'action exécutée, l'agent observe sa récompense  $r_t$  et son état suivant  $s_{t+1}$ . Plus formellement, une tâche RL  $T_i$  est définie selon le n-uplet  $(S, A, p, r, \gamma, \rho_0, H)$  où  $S$  est l'ensemble des états,  $A$  est l'ensemble des actions,  $p(s_{t+1}|s_t, a_t)$  est la distribution de transition d'état prédisant la probabilité d'atteindre un état  $s_{t+1}$  au prochain pas de temps compte tenu de l'état et de l'action actuels ;  $r$  est une fonction de récompense,  $\gamma$  est le facteur d'actualisation,  $\rho_0$  est la distribution de l'état initial et  $H$  l'horizon. Considérons la somme des récompenses attendues (retour) d'une trajectoire  $\tau_{(t, t+H-1)} = (s_t, a_t, \dots, s_{t+H-1}, a_{t+H-1}, s_{t+H})$ . Un framework RL vise à apprendre une politique  $\pi$  de paramètres  $\theta$  qui associe chaque état  $s$  à une action optimale  $a$  maximisant le retour  $R_t$  de la trajectoire (cf. équation (1)).

$$R_t = \sum_{k=t}^{t+H-1} \gamma^{k-t} r_{k+1} \quad (1)$$

Les fonctions de récompense jouent un rôle important dans la spécification des problèmes d'apprentissage et la construction de politiques de conduite pour des applications à grande échelle. Cependant, l'apprentissage dans ces contextes caractérisés par des retours clairsemés est généralement lent, nécessitant la spécification de récompenses denses et élaborées. Une technique supportant le traitement des tâches complexes par les approches RL consiste à transformer des connaissances sur des domaines sous-jacents en récompenses complémentaires. La combinaison des récompenses de différentes origines est connue sous le nom de *reward shaping* [8] et permet aux agents RL d'apprendre plus efficacement en renforçant le signal de récompense naturel par des retours intermédiaires cohérents avec certaines connaissances préalables.

#### A. Apprentissage avec une fonction de récompense basée sur la vitesse

L'agent RL, noté « agent\_V », est entraîné avec une fonction de récompense qui l'encourage à atteindre une vitesse élevée. L'agent sera incité à éviter les collisions avec les exo-véhicules par le seul fait qu'en cas d'accident, il n'a plus de récompense (cf. ; équation (2)).

$$R_v = \max \left( -1, \frac{V_{ego} - \frac{\sum_{i=0}^n V_{exo_i}}{n}}{V_{max} - \frac{\sum_{i=0}^n V_{exo_i}}{n}} \right) \quad (2)$$

Où  $V_{ego}$  et  $V_{exo_i}$  représentent les vitesses de l'ego-véhicule et le  $i^{ième}$  exo-véhicule, respectivement.  $V_{max}$  est la vitesse maximale que les véhicules peuvent atteindre dans le simulateur et  $n$  est le nombre d'exo-véhicules qui entourent l'ego-véhicule.

La durée des épisodes et la distance parcourue obtenues pour chaque entraînement sont présentées dans Fig. 2 avec les trois densités de trafic (DC-, DC et DC+). En comparant les moyennes, les résultats montrent que les performances de l'agent\_V se dégradent avec l'augmentation de la densité du trafic (Fig. 2).

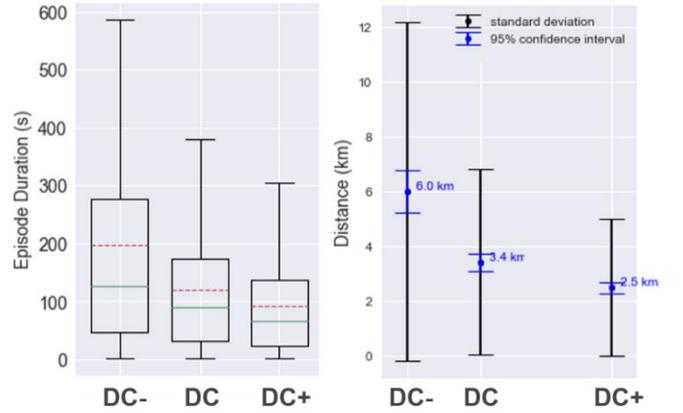


Fig. 2. Représentation de la durée des épisodes (en sec.) et des distances parcourues (en km) par l'Agent\_V pour les trois densités de trafic.

La Fig. 3 montre également que la vitesse de l'ego-véhicule et la vitesse du flux d'exo-véhicules diminuent avec la densité de trafic. Toutefois, l'agent\_V a maintenu des vitesses plus élevées que les exo-véhicules, ce qui satisfait l'un des objectifs de la tâche.

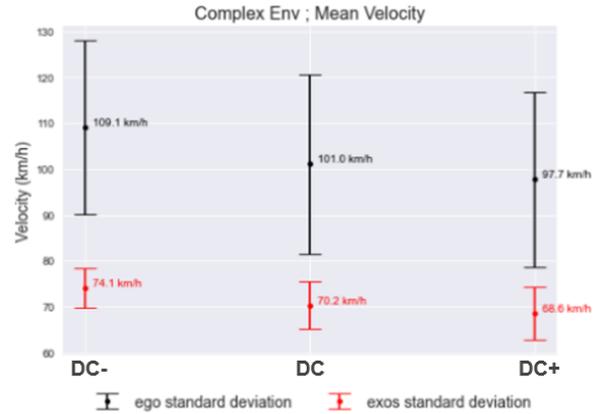


Fig. 3. Vitesse moyenne (en km/h) de l'agent\_V et des exo-véhicules pour les trois densités de trafic.

#### B. Apprentissage avec une fonction de récompense basée sur la vitesse et les accidents

L'agent RL noté « Agent\_VA » est entraîné avec une fonction de récompense en deux parties. Pour la vitesse, la fonction de récompense est identique à celle de l'« agent\_V » (fonction  $R_v$  donnée par l'équation (2)). Pour inciter l'agent à éviter les accidents, la fonction de récompense intègre une forte pénalité en cas de collision. Cette fonction de

récompense notée  $R_{va}$  est donnée par l'équation (3), (où  $c = -10$  dans notre cas).

$$R_{va} = \begin{cases} c, & \text{si collision} \\ R_v, & \text{sinon} \end{cases} \quad (3)$$

Les résultats présentés dans Fig. 4 et Fig. 5 montrent les mêmes tendances que les résultats précédents, c'est-à-dire que plus la complexité de l'environnement augmente, plus la vitesse de l'agent\_VA diminue et plus la distance parcourue et le temps écoulé sans accident sont élevés. Nous constatons également qu'indépendamment de la densité du trafic, l'agent\_VA garde toujours une vitesse supérieure à celle des exo-véhicules (cf. Fig. 5).

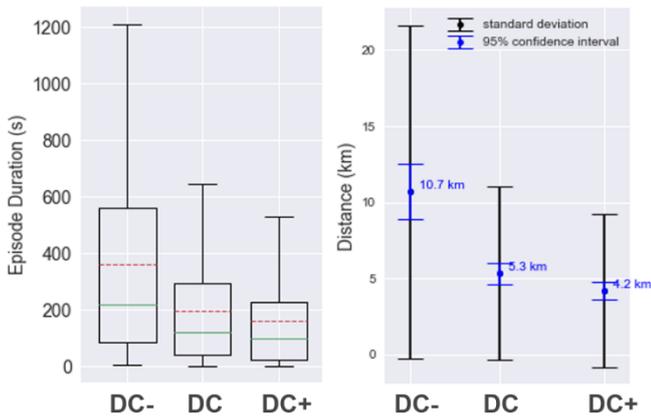


Fig. 4. Représentation de la durée des épisodes (en sec.) et des distances parcourues (en km) par l'Agent\_VA pour les trois densités de trafic.

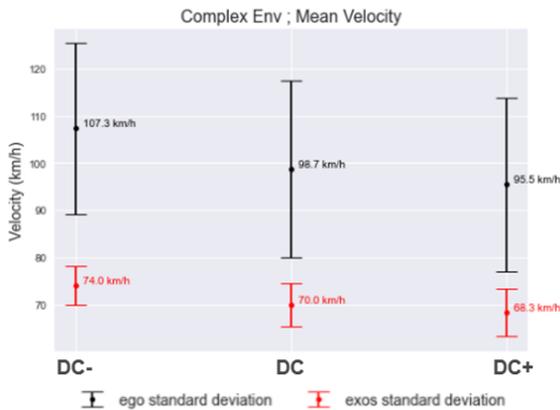


Fig. 5. Vitesse moyenne (en km/h) de l'agent\_VA et des exo-véhicules pour les trois densités de trafic.

Nous notons une amélioration de la survie (durée des épisodes) de l'agent\_VA par rapport à l'agent\_V de l'ordre de 80%. Cette amélioration s'accompagne d'une réduction de vitesse négligeable. Cette première amélioration de la fonction de récompense est donc tout à fait satisfaisante.

### C. Apprentissage avec récompense basée sur la sécurité vis-à-vis des collisions

Dans le but d'améliorer davantage les performances de l'agent vis-à-vis des collisions, une autre fonction de récompense est conçue en intégrant des mesures de sécurité ; il en résulte un nouvel agent RL noté « Agent\_SC ».

La démarche se base sur une approche analytique des risques encourus lors de la manœuvre :

- Pour un changement de voie, nous allons nous intéresser au véhicule devant l'égo-véhicule dans sa voie et aux véhicules devant et derrière dans la voie cible.
- Pour le freinage et l'accélération, nous regardons uniquement le véhicule devant l'égo-véhicule dans sa voie.

Pour chaque cas, nous définissons une fonction locale qui permet de quantifier le risque pour chaque exo-véhicule et nous laissons l'agent\_SC apprendre la meilleure stratégie afin de traiter le scénario dans sa globalité. La quantification repose sur des notions d'espacement des véhicules issues du champ applicatif des automates de conduite :

- Lorsque l'écart de vitesse ego-exo est faible, nous utilisons le temps inter véhicules « TIV » (le rapport de la distance entre l'égo-véhicule et l'exo-véhicule qui le précède sur la vitesse de l'égo-véhicule).
- Lorsque l'écart de vitesse ego-exo est plus important, nous utilisons le temps avant collision « TTC » (le rapport de la distance entre l'égo-véhicule et l'exo-véhicule qui le précède sur la différence de leurs vitesses respectives).
- Il est ensuite possible de comparer ce temps avant collision avec le temps de freinage « BT » (le rapport de la différence de vitesse entre l'égo-véhicule et l'exo-véhicule qui le précède sur la décélération moyenne de freinage).

Les fonctions de récompense locales qui utilisent les critères ci-dessus sont définies de façon continue. Cela conduit l'agent à arbitrer de façon relative entre deux comportements plutôt que de répondre à un comportement attendu. Il apprend ainsi le meilleur compromis pour chaque situation à risque.

Pour l'apprentissage, nous retenons comme note globale le minimum entre les fonctions de récompense locales (qui correspondent au pire cas) et la récompense de vitesse. La formulation de cette fonction de récompense est donnée par l'équation (4) suivante :

$$R_{SC} = \min(R_v, R_o) \quad (4)$$

Où  $R_v$  représente la récompense basée sur la vitesse donnée par l'équation (2) et  $R_o$  représente la récompense opérationnelle prenant en compte les critères définis précédemment (TIV, TTC et BT). Le détail de la fonction récompense  $R_o$  est donné par l'équation (5).

$$R_o = \begin{cases} \min(r_f, r_{ft}, r_{bt}), & \text{si changement de voie} \\ r_{fb}, & \text{si freinage} \\ c, & \text{si collision} \end{cases} \quad (5)$$

Où  $r_f$ ,  $r_{ft}$ ,  $r_{bt}$  et  $r_{fb}$  estiment respectivement les risques par rapport à l'exo-véhicule devant dans la même voie ( $r_f$ ), devant dans la voie cible ( $r_{ft}$ ), derrière dans la voie cible, ( $r_{bt}$ ) et le risque lié au temps de freinage dans la même voie ( $r_{fb}$ ). La pénalité en cas de collision  $c$  a pour valeur de  $-10$ .

La Fig. 6 présente les résultats de l'agent\_SC en termes de durée des épisodes et de distance parcourue, pour les trois densités de trafic. La Fig. 7 montre la vitesse moyenne de l'agent\_SC (ego) par rapport aux exo-véhicules dans les trois densités de trafic.

Nous constatons un gain important pour l'agent\_SC. La durée des épisodes et les distances parcourues sont

multipliées par 10 par rapport à l'agent\_V et plus de 5 par rapport à l'agent\_VA. Il est également à noter que l'augmentation de la densité de trafic n'a pas d'impact significatif sur les performances de l'agent\_SC pour les valeurs testées.

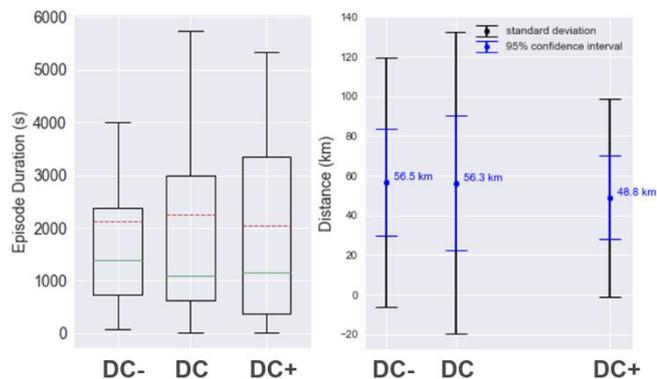


Fig. 6. Représentation de la durée des épisodes (en sec.) et des distances parcourues (en km) par l'Agent\_SC pour les trois densités de trafic.

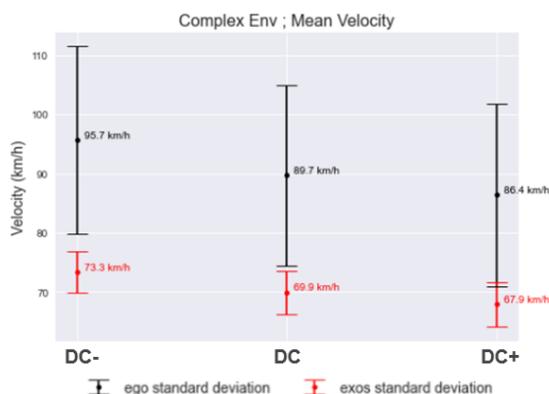


Fig. 7. Vitesse moyenne (en km/h) de l'agent\_SC et des exo-véhicules pour les trois densités de trafic.

La Fig. 7 montre qu'en contrepartie des meilleures performances, la vitesse moyenne de l'agent\_SC est plus faible que celle l'agent\_V et l'agent\_VA. Nous constatons une baisse de vitesse d'environ 10%. La vitesse de l'agent\_SC reste toujours plus élevée que celle du flux d'exo-véhicules.

#### D. Comparaison des agents

Les trois agents avec des fonctions de récompense présentés dans les sections précédentes montrent une évolution dans le comportement de la conduite. En effet l'agent\_V qui est récompensé uniquement sur la vitesse a des performances limitées en termes de durée d'épisodes et de distance parcourue.

L'agent\_VA qui prend en compte les collisions améliore davantage ses performances par rapport à l'agent\_V. Le fait de sanctionner les accidents lui permet de converger plus facilement vers un comportement plus sûr. De façon assez cohérente, les performances des deux agents se dégradent au fur et à mesure que le trafic se densifie car la complexité de la tâche de conduite augmente.

L'Agent\_SC est encore mieux guidé dans son apprentissage pour prendre en compte les situations à risque avant l'accident. Il améliore ainsi considérablement ses performances de conduite. La conduite de l'Agent\_SC qui anticipe les accidents potentiels peut être qualifiée de plus

prudente. Cette différence de comportement est clairement visible sur les vidéos de la simulation. Il en résulte que cette conduite plus prudente est également plus lente que celle des agents précédents. Elle reste toutefois nettement supérieure à celle du flux de circulation (de l'ordre de 20 km/h). Une autre preuve de cette anticipation et cette prudence apparaît en constatant que l'agent\_SC est moins sensible à l'augmentation de la densité de trafic que l'agent\_V et agent\_VA.

En complément des résultats des entraînements présentés ci-dessus, certains indicateurs clé de performance (KPIs) ont été mises en place afin de pouvoir évaluer un agent sur des indicateurs de performance et de sûreté. Les indicateurs présentés dans le tableau 1 ne représentent qu'une partie de ceux mis en place dans le projet pour les trois agents dans un environnement DC.

TABEAU I. INDICATEURS SERVANT A L'EVALUATION DES AGENTS DANS UNE DENSITE COMPLEXE (DC = 3)

KPIs	Agent_V	Agent_VA	Agent_SC
Durée moyenne des Scenarios (s)	121,09	195,02	2257,77
Vitesse moyenne de l'EGO (Km/h)	101,04	98,74	89,71
Vitesse moyenne des EXO (Km/h)	70,22	69,95	69,99
Nombre de dépassement	19494	19181	14175
Vitesse moyenne lors du dépassement (km/h)	100,83	96,79	86,66
Nombre d'accident	412	256	22
TIV Min. (s)	0,15	0,18	0,23
TTC Min. (s) (TTC lors des dépassements et freinages sans accident)	0,63	0,78	1,08
Nombre d'accident lors d'un dépassement	361	197	19
Vitesse moyenne de l'EGO lors des accidents (km/h)	103,12	97,86	76,49
Vitesse maximale de l'EGO lors des accidents (km/h)	125,97	125,52	90,00

Le tableau 1 confirme que l'agent\_SC est impliqué dans beaucoup moins d'accidents, d'où la durée plus longue des scénarios, dans un environnement identique et un flux d'exo-véhicules circulant à une vitesse moyenne d'environ 70 km/h. L'amélioration de la sûreté de l'agent est clairement visible dans les indicateurs TIV et TTC minimums, ainsi que dans sa prudence constatée au niveau du nombre de dépassements et de sa vitesse lors ceux-ci.

#### IV. APPRENTISSAGE ADVERSE

Notre deuxième contribution pour améliorer la robustesse des politiques RL consiste à adopter des approches d'entraînement antagonistes (adversarial training). Dans cet article, nous implémentons un algorithme d'optimisation de politique proximale (PPO) [9], cependant, toute autre méthode de type « acteur-critique » pourrait être retenue. La composante critique donne une estimation de la fonction de valeur  $V^\pi(s)$  correspondant à la récompense actualisée à long terme de l'état  $s$  représenté par l'observation  $x$  et suivant la politique  $\pi$ . La plupart des méthodes de perturbation adverses développées pour attaquer les DNN sont applicables aux politiques RL dans le contexte d'un espace d'action discret. De plus, il a été montré que l'introduction d'observations perturbées lors de l'entraînement aide le modèle RL à devenir

plus robuste aux entrées modifiées [10]. Des revues intéressantes des différentes méthodes d'attaques et de leurs contremesures dans le contexte de DRL sont disponibles dans les travaux de [3], [11]–[13].

Dans ce travail, nous considérons une méthode white-box à base de gradient inspirée de l'approche JSMA [14]. Elle vise à attaquer l'agent RL en générant des exemples adverses. La méthode consiste à produire des observations perturbées  $\hat{x}$  pour remplacer les observations réelles  $x$  renvoyées par l'environnement, puis à permettre à l'agent de décider de l'action  $a = \pi(\hat{x})$ . Dans le cas des espaces d'action discrets, une attaque est efficace si l'agent modifie sa décision  $\pi(x) \neq \pi(\hat{x})$ . L'agent est ainsi amené à prendre de mauvaises décisions compliquant la manière dont il peut atteindre son objectif. Dans le cadre d'une politique DNN, nous opérons en deux temps. Premièrement, nous proposons d'utiliser le gradient de la fonction de perte par rapport à chaque composant des données d'entrée (c'est-à-dire la matrice jacobienne) pour extraire la direction de la sensibilité. Ensuite, une carte de saillance est calculée pour sélectionner la dimension qui génère l'erreur maximale en utilisant le moins de perturbations possible.

Dans cette partie, nous étudions l'impact d'attaques adverses de l'observation sur l'agent. Nous testons tout d'abord les agents précédemment appris dans un environnement de test attaqué. Nous cherchons ensuite à rendre les agents plus robustes en intégrant les attaques adverses de l'observation lors de la phase d'entraînement.

Pour cette étape, nous avons retenu uniquement les deux agents les plus performants : l'agent\_VA et l'agent\_SC. Nous comparons les résultats dans les trois densités de trafic (DC-, DC et DC+).

#### A. Apprentissage adverse avec une fonction de récompense basée sur la vitesse et les accidents

Nous observons ici l'agent avec la fonction de récompense vitesse et accident (agent\_VA). Nous avons retenu l'indicateur de durée d'épisode et l'indicateur de vitesse moyenne (cf. Fig. 8 et Fig. 9).

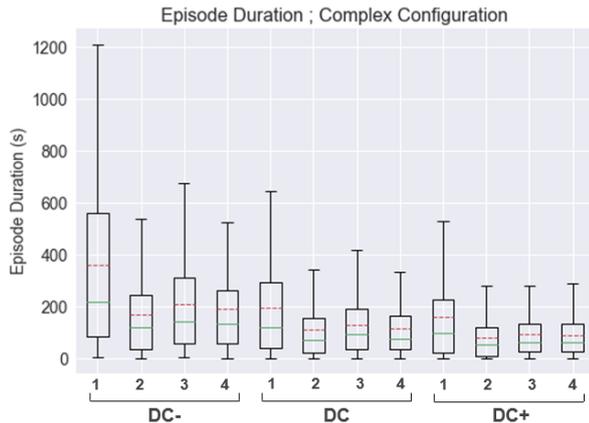


Fig. 8. Représentation de la durée des épisodes (en sec.) pour l'Agent\_VA pour les trois densités de trafic. Les boîtes à moustache : 1 correspond à l'apprentissage sans attaque (ASA) et test sans attaque (TSA), 2 correspond à ASA et test avec attaques (TA), 3 correspond à l'apprentissage avec attaques (AA) et TA et 4 correspond à AA et TSA.

Pour les 3 densités de trafic, nous observons sur la Fig. 8 les mêmes types de comportements de l'agent\_VA :

- L'agent\_VA originel voit ses performances divisées par 2 en présence d'attaques de l'environnement.
- L'utilisation des attaques adverses lors de l'apprentissage limite la perte de performances dans un environnement attaqué.
- L'utilisation des attaques adverses lors de l'apprentissage dégrade les performances dans un environnement qui n'est pas attaqué.

La tendance de dégradation de performance avec l'augmentation de la densité de trafic subsiste dans tous les cas.

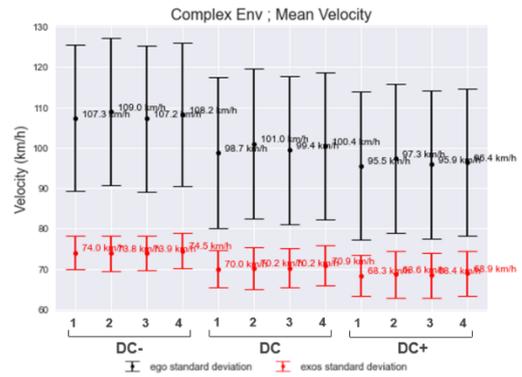


Fig. 9. Vitesse moyenne de l'agent\_VA et des exo-véhicules pour les trois densités de trafic. Les colonnes de barre d'erreur : 1 correspond à ASA et TSA, 2 correspond à ASA et TA, 3 correspond à AA et TA et 4 correspond à AA et TSA.

Nous constatons sur la Fig. 9 qu'à densité de trafic constante, les attaques adverses de l'environnement n'ont qu'un impact mineur sur les vitesses de l'égo, que son apprentissage ait bénéficié ou non des attaques adverses de l'observation.

#### B. Apprentissage adverse avec récompense basée sur la sécurité vis-à-vis des collisions

Nous observons ici l'agent\_SC avec la fonction de récompense sécurité collisions. Nous avons retenu également l'indicateur de durée d'épisode et l'indicateur de vitesse moyenne.

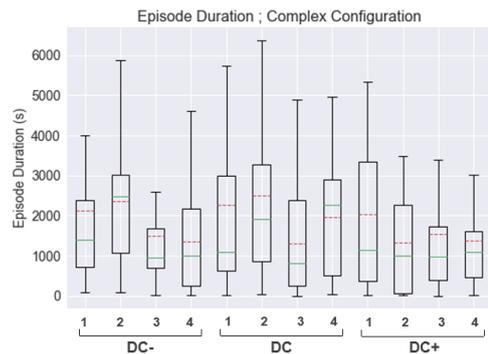


Fig. 10. Représentation de la durée des épisodes (en sec.) pour l'Agent\_SC pour les trois densités de trafic. Les boîtes à moustache : 1 correspond à l'apprentissage sans attaque (ASA) et test sans attaque (TSA), 2 correspond à ASA et test avec attaques (TA), 3 correspond à l'apprentissage avec attaques (AA) et TA et 4 correspond à AA et TSA.

Les résultats observés sur la Fig. 10 sont très différents de ceux de l'agent\_VA. Nous ne sommes pas encore en mesure de les expliquer en totalité. En effet, l'agent\_SC appris sans attaques adverses présente de meilleures performances lorsqu'il est attaqué lors des tests dans les densités de trafic DC- et DC. Pour la densité de trafic DC+, nous retrouvons un

comportement plus conforme à l'attendu avec une dégradation de l'agent\_SC lors d'un test avec attaques.

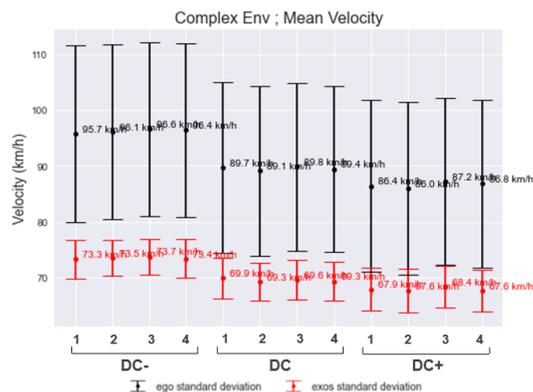


Fig. 11. Vitesse moyenne (en km/h) de l'agent\_SC et des exo-véhicules pour les trois densités de trafic. Les colonnes de barre d'erreur : 1 correspond à ASA et TSA, 2 correspond à AA et TA, 3 correspond à AA et TA et 4 correspond à AA et TSA.

Nous constatons à nouveau qu'à densité de trafic constante, les attaques adverses de l'environnement n'ont qu'un impact mineur sur les vitesses de l'égo, que son apprentissage ait bénéficié des attaques ou non.

### C. Comparaison des agents

Les résultats obtenus avec les attaques adverses diffèrent sensiblement de ce que nous avons observé dans des travaux préalables. A ce titre une comparaison complète n'est pas possible à ce stade.

Nous pouvons simplement constater que l'Agent\_VA a des performances fortement dégradées en présence d'attaques de l'environnement. L'utilisation des attaques adverses lors de l'apprentissage permet de rendre plus robuste l'agent à de telles attaques mais dégrade les performances dans un environnement sans attaques ce qui n'est pas souhaitable.

Nous avons une nouvelle indication que l'Agent\_SC est nettement plus robuste que l'Agent\_VA. Non seulement il conserve le gain initial mais il est moins dégradé lorsqu'il subit un environnement attaqué. A ce stade, nous ne pouvons pas conclure sur l'impact de l'apprentissage adverse sur l'Agent\_SC car ils sont difficilement interprétables sans travaux complémentaires.

Comme précédemment, les KPIs ont été générés afin de pouvoir évaluer et comparer les différents agents\_SC. Les résultats se trouvent dans les tableaux II, III et IV.

TABLEAU II. INDICATEURS SERVANT A L'EVALUATION DE L'AGENT SC DANS UNE DENSITE COMPLEXE BASSE (DC- = 2,5)

KPIs	Agent_SC_ DC_ ASA_TSA	Agent_SC_ DC_ ASA_TA	Agent_SC_ DC_ AA_TA	Agent_SC_ DC_ AA_TSA
Durée moyenne des Scenarios (s)	2126,30	2370,40	1493,18	1343,15
Vitesse moyenne de l'EGO (Km/h)	95,69	96,13	96,57	96,43
Vitesse moyenne des EXO (Km/h)	73,28	73,57	73,70	73,35
Nombre de dépassement	13100	12872	9100	9414
Vitesse moyenne lors du dépassement (km/h)	91,39	91,64	90,83	90,77
Nombre d'accident	23	20	28	32
TIV Min. (s)	0,22	0,25	0,22	0,24
TTC Min. (s) (TTC lors des dépassements et freinages sans accident)	1,01	1,19	1,01	1,04
Nombre d'accident lors d'un dépassement	22	20	12	16
Vitesse moyenne de l'EGO lors des accidents (km/h)	78,39	78,75	69,74	75,70
Vitesse maximale de l'EGO lors des accidents (km/h)	90,24	92,06	90,00	90,00

TABLEAU III. INDICATEURS SERVANT A L'EVALUATION DE L'AGENT SC DANS UNE DENSITE COMPLEXE (DC = 3)

KPIs	Agent_SC_ DC_ ASA_TSA	Agent_SC_ DC_ ASA_TA	Agent_SC_ DC_ AA_TA	Agent_SC_ DC_ AA_TSA
Durée moyenne des Scenarios (s)	2257,77	2498,7	1306,94	1960,45
Vitesse moyenne de l'EGO (Km/h)	89,71	89,10	89,82	89,39
Vitesse moyenne des EXO (Km/h)	69,99	69,27	69,65	69,30
Nombre de dépassement	14175	14272	11056	11404
Vitesse moyenne lors du dépassement (km/h)	86,66	86,00	85,71	85,72
Nombre d'accident	22	20	35	24
TIV Min. (s)	0,23	0,23	0,20	0,24
TTC Min. (s) (TTC lors des dépassements et freinages sans accident)	1,08	1,08	0,92	0,84
Nombre d'accident lors d'un dépassement	19	20	21	14
Vitesse moyenne de l'EGO lors des accidents (km/h)	76,49	79,99	78,34	66,80
Vitesse maximale de l'EGO lors des accidents (km/h)	90,00	90,03	90,00	87,70

TABLEAU IV. INDICATEURS SERVANT A L'EVALUATION DE L'AGENT SC DANS UNE DENSITE COMPLEXE HAUTE (DC+ = 3,3)

KPIs	Agent_SC_ DC+_ ASA_TSA	Agent_SC_ DC+_ ASA_TA	Agent_SC_ DC+_ AA_TA	Agent_SC_ DC+_ AA_TSA
Durée moyenne des Scenarios (s)	2033,08	1320,16	1529,65	1370,63
Vitesse moyenne de l'EGO (Km/h)	86,40	85,98	87,21	86,81
Vitesse moyenne des EXO (Km/h)	67,87	67,61	68,35	67,59
Nombre de dépassement	13744	13921	11737	12147
Vitesse moyenne lors du dépassement (km/h)	84,01	83,76	83,97	83,70
Nombre d'accident	24	37	32	36
TIV Min. (s)	0,24	0,26	0,24	0,23
TTC Min. (s) (TTC lors des dépassements et freinages sans accident)	1,06	1,00	1,00	0,81
Nombre d'accident lors d'un dépassement	24	32	19	21
Vitesse moyenne de l'EGO lors des accidents (km/h)	73,38	77,68	59,18	64,93
Vitesse maximale de l'EGO lors des accidents (km/h)	90,00	92,06	88,16	90,00

Nous confirmons que l'agent ayant le moins d'accidents dans les trois densités de véhicules est l'agent\_SC appris sans attaques et testé avec attaques. Ce résultat demeure surprenant et les KPIs de performance et de sûreté n'apportent pas de piste d'explication.

Les résultats montrent également que l'égo-véhicule est beaucoup plus prudent dans une densité plus élevée d'exo-véhicules même s'il a plus d'accidents (cf. tableau IV). Nous constatons également une tendance sur un agent appris avec attaques d'une diminution de 20% à 30% de son nombre de dépassements (se référer aux deux dernières colonnes de chaque tableau).

## V. DISCUSSION

La conception d'une fonction de récompense est un sujet incontournable d'un modèle RL. En effet, un agent RL est conditionné par les récompenses prédéfinies par la fonction lors de son apprentissage. L'approche implémentée dans ce travail illustre l'impact de l'ajustement de la fonction de récompense. La question sous-jacente concerne la part de l'apprentissage apportée par la fonction de récompense et la part d'apprentissage apportée par l'environnement. Dans un environnement donné, les résultats obtenus ont montré que la fonction basée sur le « reward shaping » le plus évolué, - issue des critères de risque liés à la conduite, est la plus performante. En complément, cette approche rend l'agent plus robuste à des variations de son environnement (densité de trafic ou attaques adverses des observations).

L'approche de conception de la fonction de récompense « reward shaping » montre également un intérêt vis-à-vis de la sûreté de fonctionnement. A l'ordre un, l'augmentation de performance est une condition nécessaire avant de traiter les défaillances du système ou la sécurité de la fonction attendue. Ensuite, nous avons pu constater que l'approche analytique de la fonction de récompense permettait une meilleure interprétabilité de l'agent. En effet, dans les phases de mise au

Identify applicable funding agency here. If none, delete this text box.

point, l'analyse des accidents en simulation nous a permis d'identifier facilement les éléments de la fonction de récompense qui avait conduit à la collision afin de les améliorer. Il semble donc envisageable de pouvoir allouer des exigences de sécurité à la fonction de récompense et de garantir leur traçabilité lors des essais.

Toutefois, il est important de limiter la complexité de la fonction de récompense pour ne pas générer de biais. La conception doit rester dans une logique d'apprentissage de l'agent ; l'objectif de la fonction de récompense est de le guider dans son apprentissage et non pas de spécifier son comportement.

Enfin, les problématiques liées à un environnement ouvert dans lequel la répétabilité de l'agent n'est pas garantie reste entière en termes de démonstration de la sécurité.

## VI. CONCLUSION

Dans ces travaux, nous avons proposé une fonction de récompense pour un agent de conduite obtenu par apprentissage par renforcement profond qui améliore significativement ses performances en se basant sur une connaissance applicative et l'analyse de risque associée. Dans le même temps nous avons montré un gain significatif en termes de robustesse.

Nous avons cherché à cumuler ce bénéfice lié à l'ajout de connaissances opérationnelles dans l'apprentissage avec une méthode propre à l'intelligence artificielle : les attaques adverses. A ce stade, les résultats obtenus ne nous permettent pas de conclure sur l'apport croisé des deux méthodes.

L'environnement de simulation choisi nous a permis de mener ces travaux méthodologiques pour améliorer les performances de l'agent de conduite. Toutefois, le nombre d'accidents rencontré demeure incompatible d'un usage réel. Il semble que certaines caractéristiques de l'environnement de simulation contribuent à limiter les performances atteignables. La limitation majeure concerne le mode discret. Dans l'état actuel, l'agent n'observe son environnement pour prendre une décision qu'une fois par seconde. Dans de prochains travaux, nous envisageons de nous rapprocher d'un modèle « continu » avec un pas de temps plus cohérent du besoin de la conduite autonome (100 ms voire 10 ms). En complément, ce changement permettrait de rendre les décisions modifiables avant leurs réalisations complètes pour mieux s'adapter à l'évolution continue de l'environnement. Cela nécessite la gestion par l'agent de la direction du véhicule et plus seulement de l'information binaire de changement de voie.

Nous avons également des travaux en cours pour évaluer la robustesse des agents et la couverture de l'espace des scénarios rencontrés avec des méthodes formelles telles que l'interprétation abstraite et l'analyse topologique des données [15]. L'application de ces méthodes aux agents évoqués dans l'article devrait nous permettre de mieux interpréter les résultats de l'Agent\_SC dans un environnement attaqué.

## ACKNOWLEDGMENT

Ce travail est mené par l'institut de recherche Technologique SystemX, Paris Saclay, France. Il est financé et supporté par le Programme Investissements d'Avenir (PIA). Ce travail fait partie du projet collaboratif EPI (Evaluation des Performances de l'Intelligence artificielle) supervisé par l'IRT SystemX et ses partenaires Airbus Protect, Expleo France, Naval Group et Stellantis.

## REFERENCES

- [1] S. J. Russell, P. Norvig, et E. Davis, *Artificial intelligence: a modern approach*, 3rd ed. Upper Saddle River: Prentice Hall, 2010.
- [2] I. J. Goodfellow, J. Shlens, et C. Szegedy, « Explaining and Harnessing Adversarial Examples ». arXiv, 20 mars 2015. Consulté le: 8 juin 2022. [En ligne]. Disponible sur: <http://arxiv.org/abs/1412.6572>
- [3] S. Huang, N. Papernot, I. Goodfellow, Y. Duan, et P. Abbeel, « Adversarial Attacks on Neural Network Policies ». arXiv, 7 février 2017. Consulté le: 8 juin 2022. [En ligne]. Disponible sur: <http://arxiv.org/abs/1702.02284>
- [4] E. Leurent, « An Environment for Autonomous Driving Decision-Making ». mai 2018. Consulté le: 8 juin 2022. [En ligne]. Disponible sur: <https://github.com/eleurent/highway-env>
- [5] A. Kesting, M. Treiber, et D. Helbing, « General Lane-Changing Model MOBIL for Car-Following Models », *Transp. Res. Rec. J. Transp. Res. Board*, vol. 1999, n° 1, p. 86-94, janv. 2007, doi: 10.3141/1999-10.
- [6] L. Schott, H. Hajri, et S. Lamprier, « Improving Robustness of Deep Reinforcement Learning Agents: Environment Attack based on the Critic Network ». arXiv, 17 février 2022. Consulté le: 9 juin 2022. [En ligne]. Disponible sur: <http://arxiv.org/abs/2104.03154>
- [7] R. S. Sutton et A. G. Barto, *Reinforcement Learning : An Introduction*, MIT Press. 2018.
- [8] Y. Hu *et al.*, « Learning to Utilize Shaping Rewards: A New Approach of Reward Shaping », p. 11.
- [9] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, et O. Klimov, « Proximal Policy Optimization Algorithms ». arXiv, 28 août 2017. Consulté le: 10 juin 2022. [En ligne]. Disponible sur: <http://arxiv.org/abs/1707.06347>
- [10] V. Behzadan et A. Munir, « Vulnerability of Deep Reinforcement Learning to Policy Induction Attacks ». arXiv, 15 janvier 2017. Consulté le: 10 juin 2022. [En ligne]. Disponible sur: <http://arxiv.org/abs/1701.04143>
- [11] Y.-C. Lin, Z.-W. Hong, Y.-H. Liao, M.-L. Shih, M.-Y. Liu, et M. Sun, « Tactics of Adversarial Attack on Deep Reinforcement Learning Agents ». arXiv, 12 novembre 2019. Consulté le: 10 juin 2022. [En ligne]. Disponible sur: <http://arxiv.org/abs/1703.06748>
- [12] J. Kos et D. Song, « Delving into adversarial attacks on deep policies ». arXiv, 18 mai 2017. Consulté le: 10 juin 2022. [En ligne]. Disponible sur: <http://arxiv.org/abs/1705.06452>
- [13] L. Pinto, J. Davidson, R. Sukthankar, et A. Gupta, « Robust Adversarial Reinforcement Learning », p. 10.
- [14] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, et A. Swami, « The Limitations of Deep Learning in Adversarial Settings ». arXiv, 23 novembre 2015. Consulté le: 10 juin 2022. [En ligne]. Disponible sur: <http://arxiv.org/abs/1511.07528>
- [15] F. Adjed, M. Mziou-Sallami, F. Pelliccia, M. Rezzoug, L. Schott, C. Bohn, Y. Jaafra, « Coupling algebraic topology theory, formal methods and safety requirements toward a new coverage metric for artificial intelligence models », *Neural Comput. Appl.*, mai 2022, doi: 10.1007/s00521-022-07363-6.