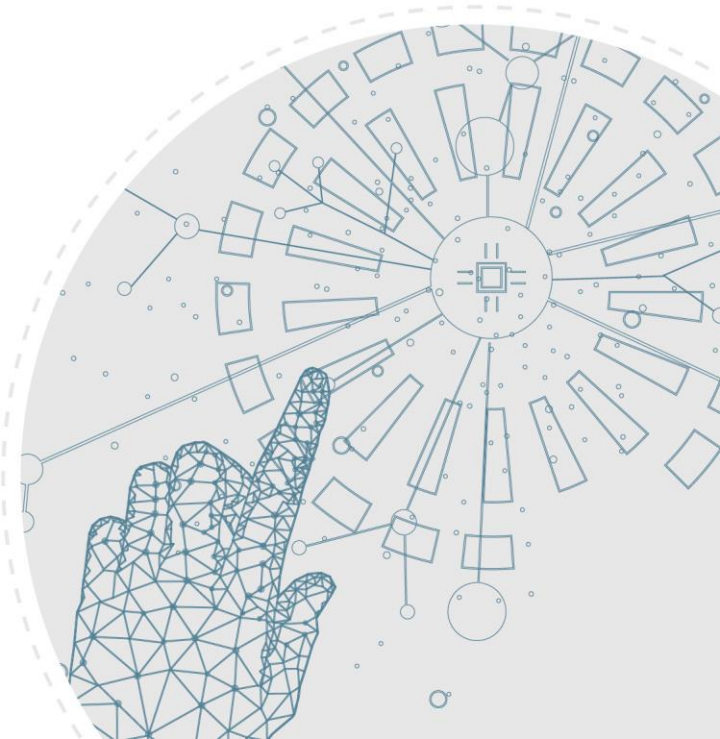
SYS DOCT 001 V1

Doctrine

Intelligence artificielle

Conditions d'autorisation des systèmes comprenant des algorithmes d'apprentissage automatique

23 octobre 2023



Avant-propos

Cette doctrine a été rédigée dans le cadre d'un groupe de travail interne à l'Établissement public de sécurité ferroviaire (EPSF) et reflète uniquement la position de l'EPSF.

Elle constitue une première réflexion sur les conditions dans lesquelles des modèles d'inférence issus d'algorithmes d'apprentissage automatique pourraient être autorisés pour des applications relatives à la sécurité dans le système ferroviaire.

L'EPSF ne possède pas d'expertise dans le domaine de l'apprentissage automatique. Cette doctrine présente donc sa compréhension à date de ce sujet avec l'objectif d'alimenter le débat et de permettre l'échange entre les experts de l'apprentissage automatique et les experts de la sécurité ferroviaire.

Table des matières

1. Introduction

L'augmentation du nombre des données a permis un développement de l'utilisation des algorithmes d'apprentissage automatique (ou *machine learning*) qui sont devenus de plus en plus performants. Ce développement permet d'envisager de nouvelles applications pour ces algorithmes et donc pour les modèles d'inférence qui en sont issus. Dans le secteur ferroviaire, l'utilisation de modèles d'inférence issus d'algorithmes d'apprentissage automatique est envisagée pour des systèmes assurant des fonctions d'assistance à un opérateur humain, par exemple pour réaliser de la maintenance prédictive, ou pour des systèmes visant à remplacer une fonction faite par un opérateur humain, par exemple, lire la signalisation latérale pour permettre la conduite d'un train sans conducteur, de façon autonome ou semi-autonome.

Or ces modèles d'inférence issus d'algorithmes d'apprentissage automatique posent de nouvelles questions vis-à-vis de la démonstration de leur niveau de sécurité. En effet, contrairement aux algorithmes « classiques » déjà autorisés dans les systèmes de transport ferroviaire et de transport guidé urbain, le niveau de sécurité de ces modèles d'inférence issus d'algorithmes d'apprentissage automatique ne peut pas être démontré uniquement en garantissant que les règles décrites par l'algorithme sont complètes et correctement codées. Les algorithmes « classiques », qui font partie des « systèmes experts », s'appuient sur la vérification d'un ensemble de règles élaborées par des êtres humains. Si ces règles sont complètes, correctement codées et correctement exécutées, le résultat obtenu sera le résultat attendu.

Pour ces modèles d'inférence issus d'algorithmes d'apprentissage automatique, démontrer que l'apprentissage a été correctement réalisé et que l'algorithme a pu déterminer la « bonne » valeur pour chaque paramètre est donc fondamental.

Pour les modèles d'inférence issus d'algorithmes d'apprentissage automatique, la bonne application des règles seules ne permet pas d'obtenir le résultat attendu. Dans ces modèles d'inférence issus d'algorithmes d'apprentissage automatique, les règles font appel à des paramètres qui permettront d'obtenir ou non le bon résultat selon la valeur qu'ils prennent. La valeur de chacun de ces paramètres n'est pas fixée par le concepteur humain de l'algorithme mais déterminée automatiquement par l'algorithme lors de la phase d'apprentissage. Pour ces modèles d'inférence issus d'algorithmes d'apprentissage automatique, démontrer que l'apprentissage a été correctement réalisé et que l'algorithme a pu déterminer la « bonne » valeur pour chaque paramètre est donc fondamental.

Les figures 1 et 2 ci-dessous schématisent ces différences dans la conception et l'exploitation des algorithmes « classiques » (en vert) et des modèles d'inférence issus d'algorithmes d'apprentissage automatique (en gris).

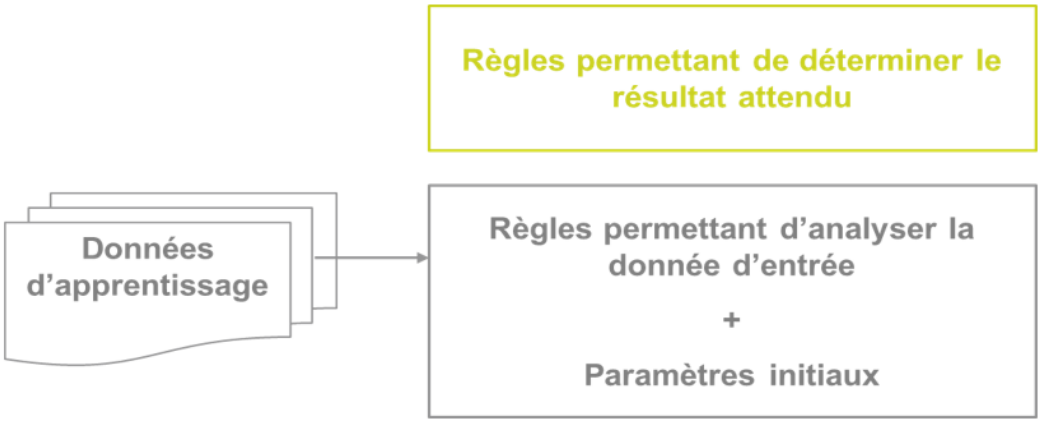


Figure 1 - Algorithme « classique » et algorithme d'apprentissage automatique en phase de conception

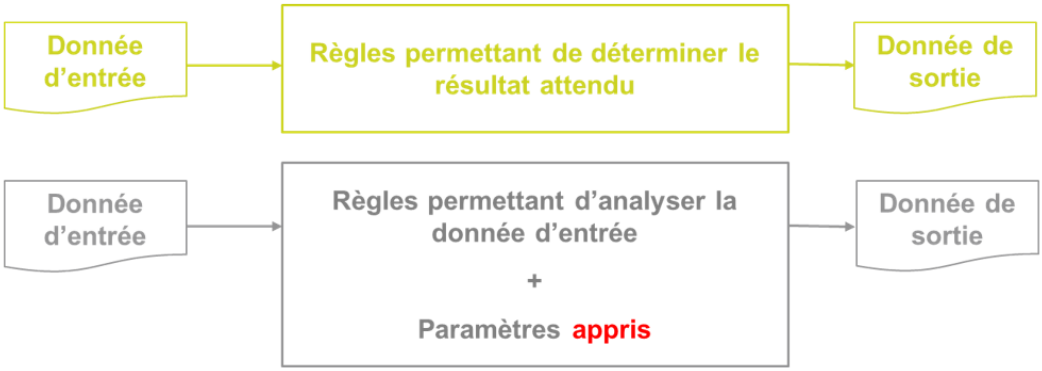


Figure 2 - Algorithme « classique » et algorithme d'apprentissage automatique en phase d'utilisation

De façon schématique, un modèle d'inférence issus d'algorithmes d'apprentissage automatique doit donc apprendre « seul » à réaliser la tâche qui lui est assignée avant de pouvoir être utilisé de façon opérationnelle, dans le sens où l'algorithme d'apprentissage doit déterminer la valeur optimale pour chacun des paramètres du modèle. L'algorithme s'appuie pour cela sur une base de données d'apprentissage qui doit être adaptée à la tâche qu'il doit réaliser (par exemple pour apprendre à reconnaître des signaux ferroviaires, il faut que la base de données contienne des images de signaux ferroviaires). À partir des données de cette base, l'algorithme va déterminer une valeur pour chaque paramètre du modèle de façon à obtenir le résultat souhaité. Parmi les grands types d'apprentissage automatique, on peut trouver :

- l'apprentissage supervisé ;
- l'apprentissage non supervisé ;
- l'apprentissage par renforcement.

La phase d'apprentissage doit comprendre une stratégie englobant toutes les actions pour calibrer tous les paramètres du modèle pouvant faire appel à un ou plusieurs de ces grands types d'apprentissage.

Dans le cas d'un apprentissage supervisé, chaque donnée de la base de données d'apprentissage est libellée avec le résultat attendu. L'algorithme d'apprentissage automatique va donc appliquer son modèle interne sur chaque donnée d'entrée présente dans la base d'apprentissage, comparer le résultat

qu'il obtient avec le résultat attendu et, si le résultat qu'il a obtenu est différent du résultat attendu, modifier ses paramètres internes. C'est par exemple ce type d'apprentissage qui est utilisé pour la classification d'image.

Dans le cas d'un apprentissage non supervisé, l'algorithme d'apprentissage va lui-même déterminer les caractéristiques correspondant aux différentes classes à partir des données de la base de données d'apprentissage. L'idée est que l'algorithme va découvrir les structures sous-jacentes à ces données non étiquetées. Ce type d'apprentissage est donc utilisé pour constituer des groupes d'éléments avec des caractéristiques communes (*clustering*).

Dans le cas d'un apprentissage par renforcement, le principe de fonctionnement consiste à identifier les actions à faire de façon à optimiser une récompense quantitative au cours du temps. L'algorithme ne sait s'il a atteint l'objectif (optimiser la récompense) qu'après plusieurs actions. Pour pouvoir déterminer ses paramètres internes, il va effectuer plusieurs expériences en calculant à chaque fois la récompense. C'est ce type d'apprentissage qui a, par exemple, été utilisé pour apprendre à un algorithme à jouer au jeu de Go.

La constitution de cette base de données d'apprentissage pour les deux premiers types d'apprentissage et la forme de la récompense revêtent donc un caractère particulièrement important pour que le modèle d'inférence issue de l'algorithme d'apprentissage automatique puisse être efficace. Pour l'apprentissage supervisé et l'apprentissage non-supervisé, la base de données d'apprentissage doit être bien représentative du problème. Il faut notamment être vigilant par rapport aux biais qui auraient pu apparaître lors de sa constitution et qui pourraient être reproduits par le modèle d'inférence issu de l'algorithme.

La présente note a pour objectif de présenter l'état des réflexions de l'EPSF sur les conditions dans lesquelles des modèles d'inférence issus d'algorithmes d'apprentissage automatique pourraient être utilisés pour des applications ferroviaires. Dans un premier temps (chap. 2.), elle rappelle les principes mis en place dans le monde ferroviaire pour garantir un haut niveau de sécurité dans le temps. Dans un second temps (chap. 3.), elle présente les principaux cas d'usage retenus ainsi que les grands enjeux associés à l'utilisation de ces modèles d'inférence issus d'algorithmes d'apprentissage automatique pour des applications ferroviaires. Dans un troisième temps (chap. 4. et chap. 5.), elle indique les éléments de description qui seraient attendus dans un dossier d'autorisation pour une application ferroviaire et questionne les exigences à remplir dans le cas d'un modèle d'inférence issus d'un algorithme d'apprentissage automatique supervisé utilisé pour une fonction de perception.

2. Contexte : les autorisations, le suivi de la sécurité et les méthodes de démonstration dans le secteur ferroviaire

Le principe fondamental de la sécurité ferroviaire est la non-régression et le maintien dans le temps du niveau global de sécurité du système ferroviaire. Ce principe fondamental implique notamment de démontrer que l'introduction d'un nouveau sous-système ou la modification d'un système existant ne dégrade pas le niveau de sécurité global du système. Ce principe de non-régression du niveau de sécurité est inscrit dans la réglementation européenne ainsi que dans le

décret n° 2019-525 relatif aux voies ferrées interopérables et dans le décret n° 2022-664 relatif aux voies ferrées locales. Ce principe est parfois appelé « Globalement au moins équivalent » (GAME). Au niveau de la conception, dans le cadre d'une démonstration explicite, la non-régression est évaluée par chaque situation dangereuse dont le taux d'apparition doit être inférieur à une valeur seuil déterminée en fonction de la gravité de l'accident associé à la situation dangereuse (matrice occurrence – gravité).

De plus, le règlement (UE) n° 402/2013 de la Commission du 30 avril 2013 *concernant la méthode de sécurité commune relative à l'évaluation et à l'appréciation des risques* fixe des objectifs de conception harmonisés pour la conception des systèmes techniques électriques, électroniques et électroniques programmables de la façon suivante au point 2.5.5 de son annexe :

a) lorsqu'une défaillance présente un potentiel crédible d'être directement à l'origine d'un accident catastrophique, il n'est pas nécessaire de réduire davantage le risque associé s'il a été établi que la défaillance de la fonction est hautement improbable (c'est-à-dire 10^{-9}/h) ;
b) lorsqu'une défaillance présente un potentiel crédible d'être directement à l'origine d'un accident critique, il n'est pas nécessaire de réduire davantage le risque associé s'il a été établi que la défaillance de la fonction est improbable (c'est-à-dire 10^{-7}/h).

Le point 2.5.11 de cette même annexe précise toutefois que *« […] si le proposant peut démontrer, pour un danger donné, que le niveau de sécurité existant dans l'État membre où le système est appliqué peut être maintenu avec un objectif de conception moins strict que l'objectif de conception harmonisé, cet objectif moins strict peut être utilisé à la place de l'objectif de conception harmonisé ».* L'objectif de non-régression est donc bien le minimum qui doit être atteint.

Pour le système ferroviaire interopérable et local, le respect de ce principe de non-régression et de maintien dans le temps du niveau de sécurité s'appuie sur deux piliers : d'une part, l'autorisation des installations fixes, des véhicules et des exploitants ferroviaires, d'autre part, le contrôle et la prise en compte du retour d'expérience.

2.1. Les autorisations

Aux fins d'autorisation, l'analyse des risques constitue le dénominateur commun entre les demandeurs d'autorisation et les autorités de sécurité. C'est cette analyse qui va permettre d'identifier les risques, puis d'y opposer des barrières de sécurité permettant de les couvrir.

Rappelons qu'un risque est défini par le règlement (UE) n° 402/2013 comme « *la fréquence d'occurrence d'accidents et d'incidents causant un dommage (dû à un danger) et le degré de gravité de ce dommage* », et que l'analyse des risques est définie comme « l'utilisation systématique de toutes les informations disponibles pour identifier les dangers et estimer le risque ».

Le cadre réglementaire européen relatif à l'analyse de risques est décrit dans le règlement n° 402/2013 applicable dans le secteur ferroviaire. Il indique notamment trois principes d'acceptation des risques :

- l'application de règles de l'art, en premier lieu les spécifications réglementaires et les normes, dont il est admis que leur respect garantit un niveau de sécurité acceptable ;
- une comparaison avec un système similaire, utilisé dans les mêmes conditions, dans la mesure où ce système a démontré au travers de son fonctionnement qu'il garantit un niveau de sécurité acceptable ;

- une estimation explicite des risques, appelée lorsque les deux premiers principes ne peuvent pas être utilisés, et qui s'appuie sur les techniques de sûreté de fonctionnement. Ce principe est particulièrement utilisé dans le cadre d'innovations disruptives, pour lesquelles aucune règle de l'art n'est définie et qu'il n'existe aucun système similaire. L'introduction croissante de nouvelles technologies tend à augmenter l'utilisation de ce principe.

Il est à noter que, au-delà de l'analyse de risque, certaines règles de l'art sont d'application obligatoire afin de garantir l'interopérabilité du système ferroviaire. Ce sont les spécifications techniques d'interopérabilité et les règles nationales. Le demandeur d'autorisation doit donc s'assurer, à la fois de la conformité de son projet aux règles d'interopérabilité et de sécurité et de la couverture de l'ensemble des risques associés à son projet. À ce jour, il n'existe aucune exigence relative à l'apprentissage automatique dans les spécifications d'interopérabilité ou les règles nationales françaises.

2.2. Les contrôles et le suivi de la sécurité

Une fois les véhicules, les installations fixes et les exploitants ferroviaires autorisés, le second pilier de la sécurité ferroviaire est le retour d'expérience au sens large, qui comprend aussi bien l'analyse des événements de sécurité que les contrôles réalisés par l'exploitant [entreprise ferroviaire (EF) ou gestionnaire de l'infrastructure (GI)] ou par l'EPSF, qui permet de vérifier, dans le temps, l'efficacité des mesures de couvertures des risques ainsi que leur application effective.

À cette fin, notamment, l'arrêté du 4 janvier 2016 *relatif à la nomenclature de classification des événements de sécurité ferroviaire* impose à l'ensemble des exploitants ferroviaires de notifier à l'EPSF les événements de sécurité survenant dans le cadre de l'exploitation de leurs services, ainsi que les éléments d'analyse adaptés à la gravité des événements. Cela représente une base d'informations abondée de plus de 20 000 évènements chaque année, classés selon une taxonomie permettant de structurer en partie les données recueillies.

Les évènements identifiés font l'objet d'une analyse dont la profondeur peut varier en fonction du potentiel d'apprentissage. Cette analyse peut être réalisée par plusieurs entités : exploitants ferroviaires, EPSF, BEA-TT.

Les systèmes et sous-systèmes mis en service doivent donc permettre de réaliser ces analyses.

Pour un système comprenant modèle d'inférence issu d'un algorithme d'apprentissage automatique, le chapitre suivant va présenter pour différents cas d'usage les enjeux associés aux deux grands piliers mentionnés ci-dessus (analyse de risques pré-opérationnelle et retour d'expérience en phase opérationnelle)

3. Les différents cas et les enjeux à prendre en compte

Aux fins d'autorisation, l'analyse des risques constitue le dénominateur

L'utilisation de modèles d'inférence issus d'algorithme d'apprentissage automatique pour des applications ferroviaires pose des enjeux différents selon l'usage de cet algorithme. Dans la suite du document, nous distinguerons donc les deux types d'usage suivant :

- Le cas où le modèle d'inférence issu d'un algorithme d'apprentissage automatique (qui sera désigné par « le modèle d'inférence » par la suite) est **directement impliqué dans une fonction de sécurité** (c'est-à-dire qu'il n'y a pas d'intervention humaine systématique qui permettrait de porter un regard critique sur ce qui est produit par le modèle d'inférence). Dans ce cas, une décision impliquant la sécurité est prise sans intervention humaine soit par modèle d'inférence issu d'un algorithme d'apprentissage automatique soit par un algorithme « classique » qui a déjà fait l'objet d'une démonstration de sécurité mais qui s'appuie sur des informations provenant d'un modèle d'inférence issu d'un algorithme d'apprentissage automatique. L'interprétation des informations portées par la signalisation latérale par un train sans conducteur rentre par exemple dans ce cas ;

- Le cas où le modèle d'inférence issu d'un algorithme d'apprentissage automatique **fournit une information à un opérateur humain** qui prendra lui la décision. Dans ce cas, ce qui est produit par le modèle d'inférence ne va pas directement aboutir à telle ou telle action de sécurité mais les informations qu'il fournit vont orienter la décision de l'opérateur humain. De plus, une absence, à tort, d'information transmise par le modèle d'inférence ne permettra pas à l'opérateur humain de porter un regard critique et de réagir en conséquence. Un système d'analyse des rails pour détecter des fissures en formation et proposer de déclencher une action de maintenance préventive rentre par exemple dans ce cas.

Les enjeux associés à ces deux types d'usage sont précisés dans les chapitres 3.1. et 3.2.

De plus, un enjeu lié au besoin de réaliser un retour d'expérience une fois le système autorisé et mis en service sera commun aux deux cas présentés ci-dessus. Cet enjeu de reproductibilité et d'explicabilité sera précisé au chapitre 3.3.

3.1. Modèle d'inférence issu d'un algorithme d'apprentissage automatique directement impliqué dans une action de sécurité

Dans le cas où le modèle d'inférence est directement impliqué dans une action de sécurité, le système comprenant ce modèle doit avoir un niveau de sécurité compatible avec la gravité de l'accident ou de l'incident qui est couvert. Ceci signifie que le modèle d'inférence ne doit pas contribuer à faire apparaître la situation dangereuse considérée au-delà d'une fréquence donnée. La démonstration de l'atteinte de ce niveau de sécurité portera, toutefois, bien sur le système dans son ensemble et non sur le seul modèle d'inférence.

De façon schématique, deux cas de figure principaux peuvent être rencontrés, chacun avec des enjeux différents :

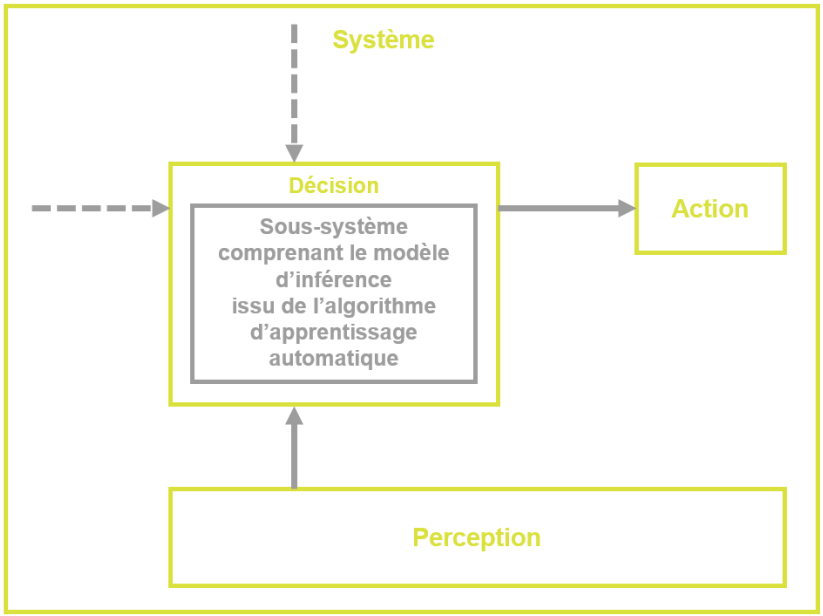
1. le sous-système comprenant le modèle d'inférence fournit une donnée de sécurité indispensable à la décision de l'automate et il est le seul à fournir cette donnée ;
2. le sous-système comprenant le modèle d'inférence analyse les données qui lui sont transmises et prend la décision de sécurité seul.

Nota : Dans la suite de la présente note, nous utiliserons le terme « justesse » (dans le sens du terme « *accuracy* » en anglais) pour qualifier la donnée de sortie issue de l'algorithme d'apprentissage automatique. Une donnée de sortie sera considérée comme juste si : i) elle correspond à ce qui est attendu ; ii) elle est transmise dans un délai compatible avec son utilisation. Pour prendre un exemple

en dehors du champ ferroviaire, la donnée de sortie d'un algorithme de classification d'image sera considérée comme juste si, lorsqu'une image de chat est présentée en entrée, l'algorithme indique que la classe la plus probable de l'image est « chat » dans le délai imparti. Le choix de ce terme a été fait de façon à éviter toute confusion avec les termes habituellement utilisés dans la sûreté de fonctionnement.

Dans le premier cas de figure, la donnée doit être fournie avec un niveau de justesse adéquat. Le sous-système comprenant le modèle d'inférence doit donc faire l'objet d'une analyse permettant d'évaluer la justesse de l'information transmise. Ce niveau de justesse devra être intégré dans la démonstration de sécurité (dans le sens de mener à son terme le processus de gestion des risques conformément au règlement (UE) n° 402/2013) du système permettant de garantir la fréquence d'apparition des situations dangereuses. Cette démonstration de sécurité prendra aussi en compte les autres données sur lesquels s'appuient le sous-système prenant la décision (notamment en cas de fusion de capteurs)

Dans le deuxième cas de figure, la décision prise par le sous-système comprenant le modèle d'inférence est directement une décision ayant un impact sur la sécurité et ce sous-système doit donc faire l'objet d'une analyse permettant d'évaluer la justesse de la décision prise. Comme dans le cas précédent, ce niveau de justesse devra être intégré dans la démonstration de sécurité du système permettant de garantir la fréquence d'apparition des situations dangereuses.

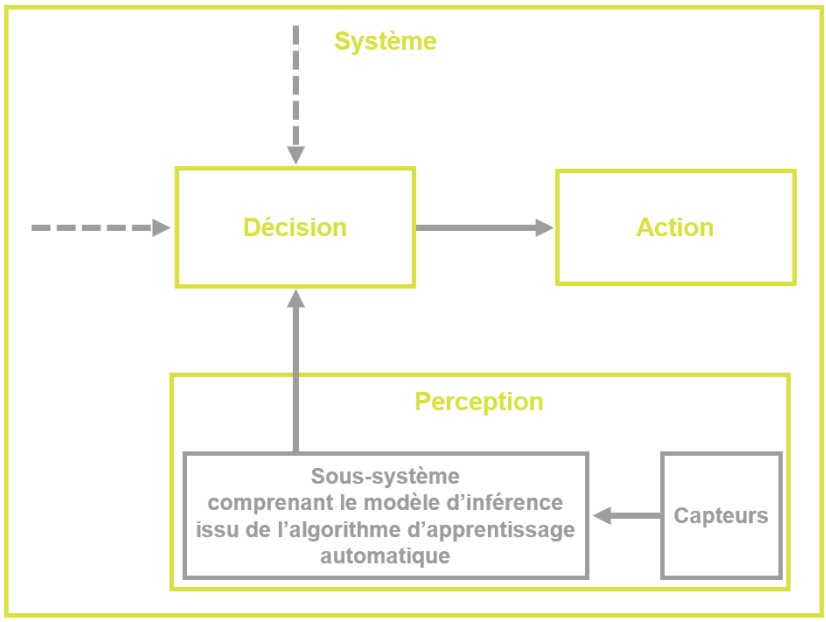


Pour les deux cas présentés ci-dessus, la péremption possible des données d'apprentissage devra être prise en compte. En effet, si les paramètres du modèle ont été fixés par rapport à une base de données d'apprentissage dont les éléments ne correspondent plus aux données qui seront fournis en entrée du modèle, l'apprentissage ne sera plus valable. Cela pourrait, par exemple, être le cas pour la lecture de la signalisation latérale si les châssis des signaux utilisés dans la base de données d'apprentissage ne correspondaient plus aux châssis utilisés sur les lignes de chemin de fer.

POUR RÉSUMER :

➔ **Pour ces deux cas, l'enjeu est donc de déterminer le niveau de justesse du sous-système comprenant le modèle d'inférence de façon à l'intégrer dans la démonstration de sécurité complète du système. Ceci implique de pouvoir déterminer, dans le domaine d'utilisation du système, une probabilité maximale pour chaque apparition d'une donnée de sortie erronée susceptible d'aboutir à une situation dangereuse. Ce niveau de justesse sera garanti tant qu'il n'y a pas péremption de la base de données d'apprentissage.**

Nota : La suite de cette note ne traitera pas plus en détail du 2e cas dont les spécificités nécessitent des approfondissements supplémentaires. La détermination des conditions d'autorisation en sécurité de modèles d'inférence utilisés dans le premier cas de figure nous semble être, déjà, une première étape intéressante à franchir.



3.2. Modèle d'inférence issu d'un algorithme d'apprentissage automatique assistant un opérateur humain dans la prise de décision

Dans le cas où le modèle d'inférence fournit une donnée qui sera analysée par un opérateur humain, l'enjeu est double dans la mesure où il porte à la fois sur le sous-système comprenant le modèle d'inférence et sur l'opérateur humain. Pour ce cas de figure, il est considéré dans ce document que l'opérateur humain a des moyens indépendants du sous-système comprenant le modèle d'inférence pour porter un jugement critique sur la donnée transmise. Il est, toutefois, considéré qu'il n'est pas possible à l'opérateur humain de porter un jugement critique dans le cas d'un faux négatif, c'est-à-dire qu'une donnée aurait dû être transmise à l'opérateur humain et qu'elle ne l'a pas été.

Pour ce qui concerne le sous-système comprenant le modèle d'inférence, pour certaines situations dangereuses, l'objectif pourrait être que la donnée qu'il va transmettre puisse être comprise et analysée avec un regard critique par l'opérateur humain. Ceci signifierait, a minima, que la justesse attendue (et

non garantie comme dans le chapitre 3.1.) de l'algorithme d'apprentissage automatique ainsi que son domaine d'utilisation devraient être précisés. Selon les modèles utilisés, la donnée transmise pourrait être accompagnée d'un taux de fiabilité ou d'un intervalle de confiance. Ceci signifie aussi que les résultats donnés doivent pouvoir être explicable pour l'opérateur humain (explicabilité locale[1]).

Pour ce qui concerne l'opérateur humain, dans le cas des situations dangereuses mentionnées ci-dessus, l'objectif est qu'il soit en mesure de porter un jugement critique sur la donnée transmise par le sous-système comprenant le modèle d'inférence. Ceci implique qu'il soit formé de façon à pouvoir comprendre les résultats issus de l'algorithme d'apprentissage automatique. Cette formation pourrait comprendre une formation sur les grands principes de fonctionnement et les limites de l'algorithme en lui-même mais aussi une formation sur des outils lui donnant des éléments d'explicabilité sur les données issues de l'algorithme. Cela implique aussi qu'il ait à sa disposition d'autres moyens pour confirmer ou infirmer une analyse issue de l'algorithme d'apprentissage automatique. Enfin cela implique la mise en place de mesures permettant de rappeler à l'opérateur humain que le système n'est pas infaillible même s'il est performant.

POUR RÉSUMER :

➔ **Pour ce cas de figure, pour certaines situations dangereuses, l'enjeu est que l'opérateur humain ait conscience du caractère faillible du sous-système comprenant le modèle d'inférence, soit en mesure de porter un jugement critique sur la donnée issue du modèle d'inférence (parce que les sorties de l'algorithme sont intelligibles pour lui et parce qu'en cas de doute, il a d'autres outils à sa disposition pour confirmer ou infirmer l'analyse du modèle d'inférence).**

3.3. Permettre le retour d'expérience

Le retour d'expérience, au sens large, qui comprend aussi bien l'analyse des événements de sécurité que les contrôles réalisés par l'exploitant (EF ou GI) ou par l'EPSF, joue un rôle important dans le maintien et l'amélioration du niveau de sécurité du système ferroviaire. L'analyse des événements permettant le retour d'expérience est réalisée par les différents acteurs du système ferroviaire sur des périmètres et des événements différents : les EF et les GI sur les événements qui les concernent directement, l'EPSF sur les événements qui lui sont remontés et avec un rôle d'agrégateur au niveau national notamment pour partager le retour d'expérience au profit de tous, le bureau enquête accident pour les transports terrestres (BEA-TT) sur les événements les plus graves .

Pour effectuer ce retour d'expérience, il est nécessaire que ces acteurs puissent déterminer les causes des différents incidents et accidents. Chaque événement doit pouvoir être analysé en profondeur de façon à en déterminer les causes premières. L'identification de ces causes premières a pour objectif de déterminer les barrières de sécurité qui n'ont pas été efficaces et les raisons de leur inefficacité ainsi que les éventuelles barrières de sécurité manquantes.

[1] Au sens de l'intelligibilité des sorties in Maël Pégny, Mohamed Issam Ibnouhsein. Quelle transparence pour les algorithmes d'apprentissage machine ? 2018. Hal-01791021

La reproductibilité sera nécessaire pour pouvoir vérifier si le sous-système comprenant un algorithme d'apprentissage automatique a connu une défaillance ou non. Ceci signifie que l'état dans lequel se trouvait ce sous-système au moment de l'évènement doit pouvoir être retrouvé et que les données d'entrée de ce système doivent pouvoir être disponibles afin de pouvoir « rejouer » l'événement. Sur ce second point, l'enregistrement des données d'entrée peut être réalisé à différents moments : en sortie des capteurs (données brutes), pendant les pré-traitements éventuels, juste avant leur traitement par le modèle d'inférence. Le moment où ces données sont enregistrées devra faire l'objet d'une réflexion et d'une justification en fonction notamment des traitements effectués sur les données en amont du modèle d'inférence.

Si la reproduction de l'événement conclut à une défaillance du sous-système comprenant un algorithme d'apprentissage automatique, il sera nécessaire de pouvoir expliquer cette défaillance, notamment vis-à-vis du modèle d'inférence. Ceci signifie que les personnes en charge du retour d'expérience (au sein des exploitants ferroviaires mais aussi au sein de l'EPSF et du BEA-TT) devront être en mesure de comprendre le choix réalisé par le modèle d'inférence dans le cas précis de l'événement. Il est donc, a minima, attendu une explicabilité locale[2] du modèle d'inférence. Cette explicabilité locale pour les personnes en charge du retour d'expérience peut nécessiter des outils spécifiques.

Dans le cadre d'un système comprenant un algorithme d'apprentissage automatique, ce besoin de pouvoir réaliser un retour d'expérience pose la question de la reproductibilité d'une part et de l'explicabilité d'autre part.

POUR RÉSUMER :

- ➔ **Compte tenu du rôle du retour d'expérience dans le maintien et l'amélioration du niveau de sécurité du système ferroviaire, notamment avec l'analyse d'événements précurseurs, la reproductibilité des événements de sécurité est un enjeu important. Cette reproductibilité implique, à ce stade, que l'apprentissage soit figé à la mise en service du modèle d'inférence et que les données d'entrée de l'algorithme soient enregistrées sur une période suffisante avec des dispositifs robustes de type « boîte noire ».**
- ➔ **De plus, pour réaliser le retour d'expérience, les algorithmes d'apprentissage automatique devront pouvoir être explicables localement, c'est-à-dire qu'un résultat donné par le modèle d'inférence doit pouvoir être expliqué, par les personnes en charge du retour d'expérience (au sein des exploitants ferroviaires mais aussi de l'EPSF et du BEA-TT) ce qui peut nécessiter des connaissances et des outils spécifiques.**

Le présent chapitre se limitera aux cas d'un modèle d'inférence directement impliqué dans une action de sécurité présent dans sous-système de perception (cf. chap. 3.1.) et à celui d'un modèle d'inférence assistant un opérateur humain (cf. chap. 3.2.). Comme indiqué ci-avant, le cas d'un modèle d'inférence présent dans un sous-système de décision ne sera pas détaillé dans le cadre de cette note.

[2] Cette notion est notamment développée in Maël Pégny, Mohamed Issam Ibnouhsein. Quelle transparence pour les algorithmes d'apprentissage machine ? 2018. Hal-01791021

4. Description attendu

4.1. Le système (véhicule ou installations fixes)

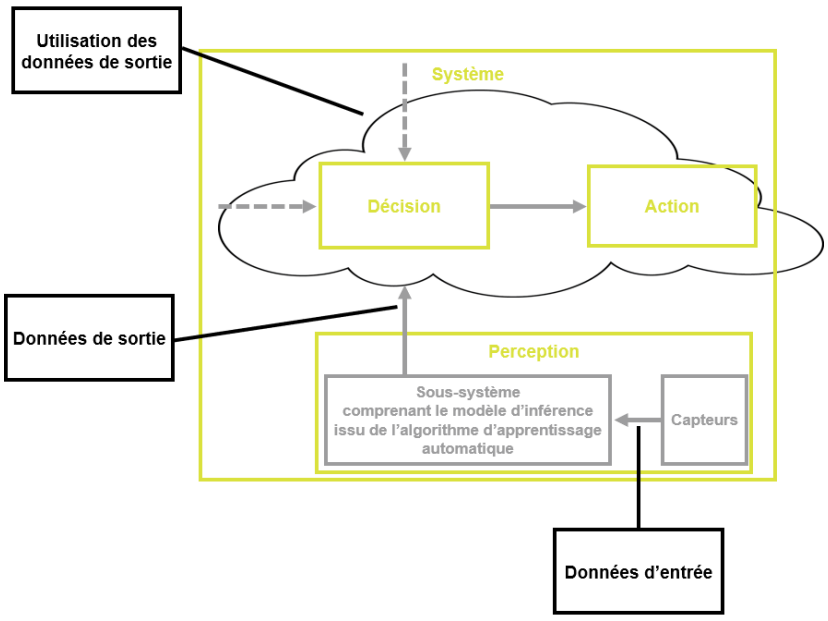
La logique d'autorisation dans le secteur ferroviaire qui porte soit sur des véhicules soit sur des installations fixes a été rappelée au début de cette note. Il n'y aura donc pas d'autorisation d'un équipement comprenant un algorithme d'apprentissage en tant que tel. L'autorisation portera sur l'ensemble du système « véhicule » ou l'ensemble du système « installations fixes ». Il est à noter que, dans le cadre de cette note, le terme système se rapporte au périmètre de l'autorisation (véhicule ou installations fixes). Pour mémoire, l'autorisation prend bien aussi en compte l'intégration en sécurité (« *safe integration* ») de ce système dans le système ferroviaire au sens de l'annexe I de la directive (UE) 2016/797 du Parlement européen et du Conseil du 11 mai 2016 *relative à l'interopérabilité du système ferroviaire au sein de l'Union européenne*.

La description de ce système « véhicule » ou « installations fixes » est donc primordiale. Elle doit donner une vue d'ensemble du fonctionnement du système et décrire comment le sous-système comprenant le modèle d'inférence s'inscrit dans ce système et comment il contribue à la réalisation des fonctions de ce système. Elle doit aussi expliciter les conditions d'exploitation et de maintenance de ce système avec une attention particulière aux facteurs organisationnels et humains.

Dans le cadre de la description du système, le sous-système comprenant le modèle d'inférence peut être considéré comme une « boîte noire ». L'objectif de cette description du système vis-à-vis du sous-système comprenant le modèle d'inférence est d'identifier :

- les données d'entrée du sous-système comprenant le modèle d'inférence ;
- les données de sortie du sous-système comprenant le modèle d'inférence ;
- la façon dont ces données de sorties sont utilisées par le système considéré et, le cas échéant, par l'opérateur humain pour remplir les fonctions attendues.

Le schéma ci-dessous représente ces éléments.



Afin d'illustrer cela, nous allons considérer le cas simplifié d'un sous-système permettant de détecter et de reconnaître les obstacles éventuels qui pourraient se situer au niveau de la voie dans le gabarit du train. Le sous-système de détection des obstacles comprenant le modèle d'inférence a comme données d'entrée, dans cet exemple, les images d'une caméra filmant ce qui se trouve devant le train. En sortie, le sous-système transmet une présomption de catégorie de ce qu'il y a devant le train : pas d'obstacle, humain, animal, arbre, rocher, fumée. En fonction de cette vraisemblance d'information, le train va réagir de la façon suivante (en considérant que tous les autres paramètres restent inchangés) :

- pas d'obstacle : pas de changement de la consigne de traction ;
- humain, animal, arbre, rocher : déclenchement d'un freinage d'urgence ;
- fumée (due à un feu à proximité des voies qui pourrait se propager au train s'il s'arrêtait à côté) : consigne de traction pour atteindre la vitesse maximale permise.

Dans cet exemple, on peut voir que ce n'est pas le niveau de justesse de toutes les données de sortie du sous-système comprenant le modèle d'inférence qui va être utile pour la démonstration de sécurité mais plutôt le niveau de justesse de sa capacité à détecter qu'il n'y a pas d'obstacle ainsi que le niveau de justesse de sa capacité à détecter que l'obstacle devant lui est de la fumée. Ces niveaux de justesse seront intégrés dans la démonstration de sécurité pour garantir que le risque de collision et le risque d'incendie sont couverts.

De plus, cette description du fonctionnement du système doit être accompagnée de son domaine d'emploi en mode nominal et dégradé qui précisera les limites d'utilisation du système et donc du sous-système comprenant le modèle d'inférence. Ce domaine d'emploi indiquera notamment la vitesse maximale de fonctionnement, les conditions de luminosité limites (nuit/ensoleillement important), les conditions climatiques limites (neige, brouillard, etc.), les éventuelles contraintes exportées vers l'exploitant et/ou le mainteneur. La démonstration de sécurité devra apporter la garantie que le système n'est pas utilisé en dehors de son domaine d'emploi.

Dans le cas d'un algorithme d'apprentissage automatique assistant un opérateur humain, la description du système devra aussi comprendre l'articulation entre le système et l'opérateur humain. Cette description devra permettre de cartographier et qualifier exhaustivement les éléments mis à la disposition de l'opérateur humain pour qu'il puisse comprendre et porter un regard critique sur la donnée, transmise par le système, associé à un niveau de « confiance ». Elle devra aussi permettre de comprendre dans quelles conditions l'opérateur humain sera amené à interagir avec le système en situation nominale et dégradée d'exploitation. Une attention particulière sera apportée sur la prise en compte des facteurs organisationnels et humains.

4.2. Le sous-système comprenant le modèle d'inférence

Une fois le système décrit et le rôle du sous-système comprenant le modèle d'inférence précisé, les éléments permettant de déterminer le niveau de justesse des données de sortie de ce sous-système devront être justifiés et détaillés.

À ce stade, les éléments qui ont été identifiés comme devant faire l'objet d'une attention particulière compte tenu de leur impact sur le niveau de justesse des données de sortie du sous-système comprenant le modèle d'inférence sont :

- l'architecture du modèle d'inférence ;
- la phase d'apprentissage ;

- la phase de validation du modèle d'inférence
- la configuration matérielle utilisée pour la phase de validation et d'exploitation.

L'architecture du modèle d'inférence devra être décrite ainsi que les raisons qui ont mené à ce choix. L'objectif de cette description est double : d'une part, donner les arguments permettant d'expliquer en quoi l'architecture retenue est adaptée à la fonction que doit remplir le modèle d'inférence ; d'autre part, permettre la traçabilité nécessaire dans le cadre du retour d'expérience.

La façon dont l'apprentissage est réalisé sera décrite. L'objectif de cette description est double : d'une part indiquer ce qui a été réalisé pour atteindre l'optimum d'apprentissage pour la fonction considérée (c'est-à-dire, par exemple, le minimum de la fonction de coût utilisée) ; d'autre part démontrer la représentativité des données d'apprentissage par rapport au domaine d'emploi envisagé.

Le processus de validation du modèle d'inférence sera décrit. Des éléments de justification devront notamment être apportés sur les trois points suivants :
- comment le processus de conception de l'algorithme contribue à sa validation ?
- comment la base de données utilisée pour la validation a été réalisée, notamment, vis-à-vis de la base de données d'apprentissage, et sa représentativité de l'ensemble des situations réelles rencontrées est-elle garantie ?
- quelle est la métrique utilisée pour l'évaluation et quelles sont les raisons du choix de cette métrique ?

La gestion de la péremption de l'apprentissage devra être abordée. Les mesures à mettre en œuvre pour vérifier que l'apprentissage est toujours valable par rapport aux situations réelles rencontrées devront être décrites.

La configuration matérielle qui sera utilisée lors de l'utilisation du sous-système comprenant le modèle d'inférence et ses conditions d'utilisation seront décrites avec l'objectif de démontrer que les calculs réalisés sont bien ceux qui sont attendus, que le temps de calcul est bien compatible avec l'utilisation du sous-système et que l'erreur lié au calcul est maîtrisée. Il est à noter que la configuration matérielle lors de l'utilisation du sous-système d'apprentissage automatique peut-être différente de celle de l'apprentissage. Cela doit, cependant, être la même configuration qui est utilisée pour les tests de validation.

Par rapport à l'ensemble des éléments qui sont mentionnés ci-dessus, le processus de second regard interne au projet qui est mis en place ainsi que le processus de second regard par un tiers indépendant conformément au règlement (UE) n° 402/2013 seront décrits.

5. Exigences attendues et questions à traiter

Compte tenu des éléments évoqués dans la présente note, l'autorisation de système comprenant des algorithmes d'apprentissage automatique ne pourra être possible que si certaines exigences sont remplies. Certaines exigences semblent atteignables dès à présent, d'autres devraient nécessiter des recherches et des développements complémentaires.

La représentativité de la base de données d'apprentissage fera l'objet d'une attention particulière. Elle devra être liée au domaine d'emploi tel que précisé dans la description du système.

À ce stade des réflexions, les exigences identifiées et les actions à suivre associées à certaines de ces exigences sont décrites ci-dessous. Ces exigences doivent permettre de répondre aux trois exigences générales suivantes :

1. le sous-système comprenant le modèle d'inférence doit être certifiable de façon à pouvoir être pris en compte dans la démonstration de sécurité du système (cela comprend les phases d'exploitation et de maintenance pour garantir le maintien du niveau de sécurité dans le temps) ;
2. le sous-système comprenant le modèle d'inférence doit pouvoir être auditable ;
3. une fois mis en service, l'action du sous-système comprenant le modèle d'inférence doit pouvoir être reproductible.

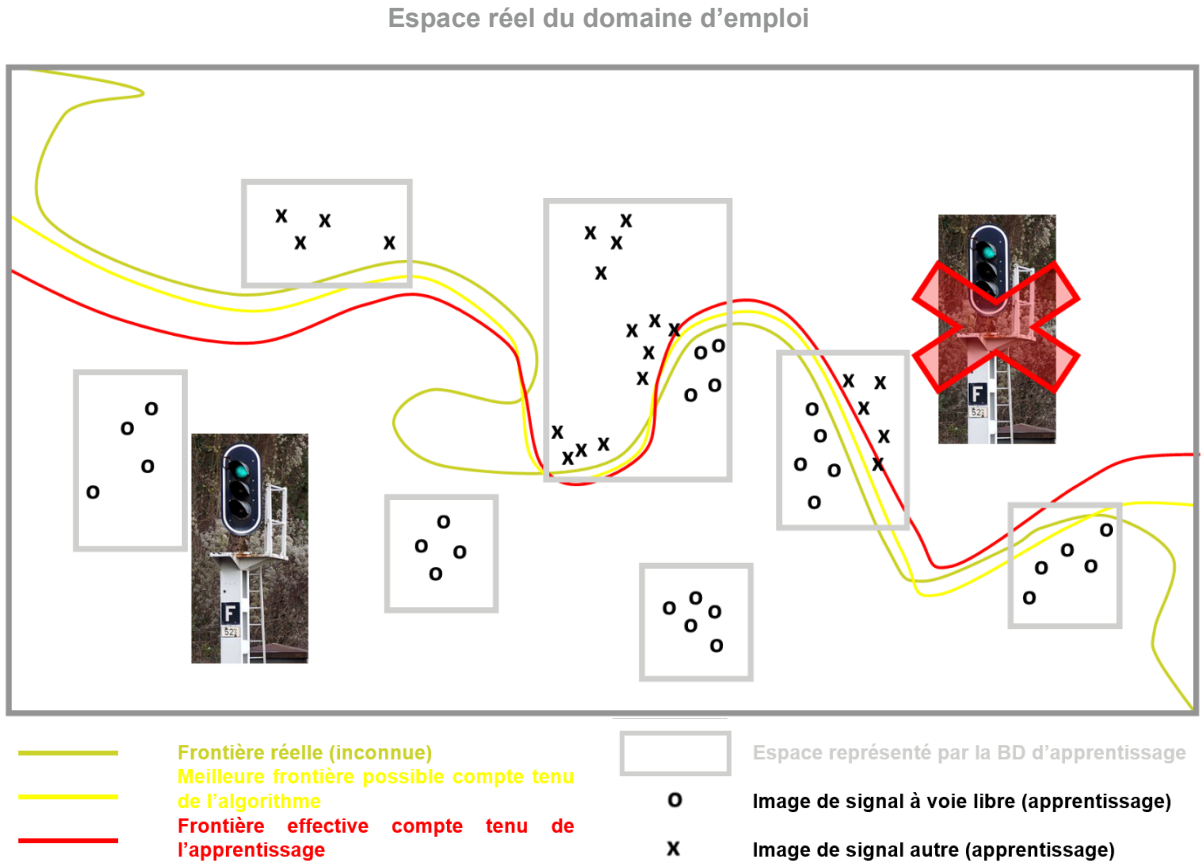
5.1. Le sous-système comprenant le modèle d'inférence doit être certifiable de façon à pouvoir être pris en compte dans la démonstration de sécurité du système

Exigence E1 : Un niveau de justesse pour chaque donnée de sortie d'un sous-système comprenant un modèle d'inférence devra pouvoir être déterminé et démontré pour un domaine d'emploi donné.

Pour répondre à cette exigence, deux grandes sources d'erreur semblent devoir être distinguées :

- la première source d'erreur est propre aux algorithmes d'apprentissage automatique. Elle est précisée ci-après ;
- la seconde source d'erreur concerne les erreurs de calcul dues à une erreur du matériel utilisé. Cette seconde source est commune à tous les logiciels. Elle n'est pas détaillée ci-après mais fait l'objet de questions à traiter.

Le schéma ci-dessous décompose, de façon théorique, certaines courbes n'étant pas connues, les erreurs propres aux modèles d'inférence issus d'algorithmes d'apprentissage automatique. Il prend pour exemple le cas d'un algorithme de détection d'un signal ferroviaire à voie libre.



Un résultat de sortie erroné du sous-système comprenant le modèle d'inférence pourrait donc être dû :

- au choix d'un modèle d'inférence qui ne permet pas d'épouser complétement la courbe faisant office de frontière entre tous les signaux qui sont à « voie libre » et tous ceux qui ne le sont pas. Ceci est représenté sur le schéma par l'écart entre la courbe verte et la courbe jaune ;
- à une base de données d'apprentissage qui n'est pas représentative de toutes les situations qui peuvent être rencontrées (l'espace réel des situations), ce qui ne permet pas au sous-système d'être juste sur certains cas de figure ;
- à la mise en œuvre de l'apprentissage qui ne permet pas d'atteindre la meilleure courbe possible compte-tenu de la structure du modèle d'inférence choisi. Ceci concerne notamment les réseaux de neurones profonds pour lesquels la fonction de coût n'est pas convexe et donc seul un minimum local pourrait être atteint lors de l'apprentissage.

Ces deux derniers points composent l'écart entre la courbe jaune et la courbe rouge.

Compte tenu de ces éléments, cette première exigence amène les questions à traiter suivantes :

Question Q1.1 : comment démontrer que l'espace représenté par la base de données d'apprentissage est représentatif de l'espace réel dans lequel le sous-système va fonctionner ?

Question Q1.2 : comment concevoir le modèle d'inférence de façon à démontrer qu'il est adapté à la fonction souhaitée ?

Question Q1.3 : comment réaliser l'apprentissage de façon à démontrer que l'optimum d'apprentissage a été atteint ? Cette démonstration est-elle nécessaire à la certification ?

Question Q1.4 : comment valider la justesse du sous-système comprenant le modèle d'inférence ? En particulier, quel doit être le volume de tests réalisés pour être statistiquement significatifs en fonction de l'objectif de sécurité à atteindre (erreur de généralisation par rapport à l'erreur réellement constaté lors des tests) ?

Question Q1.5 : les langages de programmation ont-ils un impact sur le niveau de sécurité du modèle d'inférence issu de l'algorithme d'apprentissage automatique ? Si oui, les langages actuellement utilisés sont-ils suffisamment robustes ?

Question Q1.6 : la configuration matérielle utilisée pour les tests et l'exploitation du sous-système doit-elle être de sécurité (par exemple avec une architecture 2 sur 3) ?

Question Q1.7 : Est-il possible de contrôler la pertinence de l'utilisation du sous-système durant son exploitation ?

***

> **Exigence E2 :** Pour garantir le maintien dans le temps de ce niveau de justesse pour chaque donnée de sortie d'un sous-système comprenant un algorithme d'intelligence artificielle, un processus de suivi devra être mis en place.

Au cours du temps, le niveau de justesse du modèle d'inférence pourrait diminuer soit parce que la qualité des données transmises pourrait évoluer (vieillissement des capteurs, nouveaux capteurs avec une sensibilité différente, évolution de l'environnement « technique » du capteur – par exemple, la teinte du pare-brise derrière lequel se situe une caméra change) ou parce que « l'espace réel » évolue au-delà des capacités de généralisation de l'algorithme (par exemple avec la mise en place d'un nouveau châssis pour les feux ferroviaires).

Compte tenu de ces éléments, cette deuxième exigence amène les questions à traiter suivantes :

Question Q2.1 : Quel suivi doit être mis en place pour surveiller l'environnement technique en amont du modèle d'inférence ? Le principe utilisé pour qualifier une modification du système ferroviaire est-il nécessaire et suffisant ?

Question Q2.2 : Est-il possible de surveiller ou maîtriser l'évolution de « l'espace réel » en continue ?

Question Q2.3 : Faut-il une revue périodique de la certification du sous-système comprenant le modèle d'inférence ?

5.2. Une fois mis en service, l'action du sous-système comprenant le modèle d'inférence doit pouvoir être reproductible

Exigences E3 : En cas d'incident ou d'accident, l'action du sous-système comprenant le modèle d'inférence devra pouvoir être reproduite

L'objectif principal de cette exigence est de pouvoir reproduire ce qui a été réalisé par le sous-système comprenant le modèle d'inférence afin de pouvoir déterminer s'il est responsable de l'apparition de la situation dangereuse en déterminant si la donnée issue de ce sous-système était adéquate ou non. Si cette donnée de sortie n'était pas adéquate, cela permettrait, avec le caractère auditable du sous-système de comprendre à quoi est due l'erreur.

Pour pouvoir reproduire une action du sous-système comprenant le modèle d'inférence après un événement, deux conditions semblent devoir être remplies :

i) être en mesure d'utiliser le sous-système dans le même état que celui dans lequel il était au moment de l'événement ;
ii) disposer des données d'entrées connues du sous-système au moment de l'évènement.

Compte tenu de ces éléments, cette deuxième exigence amène les questions à traiter suivantes :

Question Q3.1 : le fait de figer l'apprentissage avant la mise en service est-il nécessaire et suffisant pour répondre à cette exigence ?

Question Q3.2 : Quelles sont les données pertinentes qui devraient être enregistrées en amont du traitement par le modèle d'inférence pour pouvoir reproduire un événement ? (données d'entrée du modèle d'inférence, données brutes issues des capteurs, données prétraitées en partie, etc.)

Question Q3.3 : À l'image du JRU (*Juridical Recording Unit*), comment ces données/décisions pertinentes pourraient être stockées de façon sûre ?

5.3. Le sous-système comprenant le modèle d'inférence doit pouvoir être auditable

Exigence E4 : Les opérateurs humains devront avoir les compétences et les outils nécessaires pour porter un regard critique sur les données transmises par le sous-système comprenant le modèle d'inférence. Cette exigence doit être déclinée dans le cadre de l'analyse d'un événement ainsi que dans le cadre d'un sous-système assistant la prise de décision d'un opérateur humain.

Pour le respect de cette exigence, trois cas semblent devoir être distingués :

- le cas de l'analyse d'un événement réalisé par une équipe spécialisée dans le retour d'expérience ou une équipe spécialisée de maintenance. La temporalité de cette analyse n'est pas immédiate et certains aspects de l'analyse peuvent demander des approfondissements qui pourraient être sous-

traités. Ce cas correspond aussi à des analyses réalisées dans le cadre du contrôle continu de l'exploitant ou des audits de l'EPSF ;
- le cas de l'analyse immédiate d'un accident ou d'un incident afin de savoir si le sous-système comprenant le modèle d'inférence est impliqué et pourrait faire courir un risque grave et imminent, ce qui imposerait de suspendre l'utilisation des sous-systèmes identiques ;
- le cas d'un sous-système assistant un opérateur humain pendant l'exploitation du système ferroviaire.

Dans le premier cas, l'équipe qui devra être capable de comprendre pourquoi le sous-système comprenant le modèle d'inférence a fourni un résultat donné, pourrait avoir plusieurs compétences en son sein dont une dans les algorithmes d'apprentissage automatique. Elle aura de plus le temps de faire appel à des compétences et des outils externes. Dans le deuxième cas, l'équipe en charge de l'analyse devra être capable de déterminer dans un temps court si l'algorithme est en cause ou non, ou a minima, s'il y a un doute concernant son implication. Dans le troisième cas, chaque opérateur qui utilise le sous-système comprenant un modèle d'inférence devra être en mesure de porter un jugement critique sur la donnée de sortie transmise par le sous-système dans un temps court.

Compte tenu de ces éléments, pour chacun de ces trois cas, cette troisième exigence amène les questions à traiter suivantes :

Question Q4.1 : Quelle formation doit être dispensée aux équipes d'analyse d'une part et aux opérateurs d'autre part pour leur permettre de comprendre la donnée de sortie du sous-système comprenant le modèle d'inférence, dans un cas donné (explicabilité locale) et donc de porter un jugement critique sur cette donnée de sortie ?

Question Q4.2 : Est-il envisageable que le sous-système comprenant le modèle d'inférence puisse évaluer de façon autonome et sure s'il est bien utilisé dans ses conditions de fonctionnement nominales (par exemple, avec un niveau de luminosité suffisant ou un niveau de brouillard par trop important).

Question Q4.3 : Des outils indépendants du sous-système comprenant le modèle d'inférence et sûrs sont-ils nécessaires pour permettre de porter ce jugement critique ? Si oui, quels seraient les outils utiles disponibles pour chacun des cas ?

6. Remerciements

Pour l'élaboration de cette note, le groupe de travail a pu échanger avec les entités suivantes : :
- Le laboratoire Heudiasyc de l'UTC (Université Technologique de Compiègne), associé au CNRS, et en particulier Sébastien Destercke et Mohamed Sallak ;
- La mission certification du projet DEEL (*Dependable and Explainable Machine Learning*, www.deel.ai) et en particulier Hugues Bonnin, Florence De Grancey, Sébastien Gerchinovitz, Franck Mamalet ; la direction de la recherche de la SNCF SA, le CIM (Centre d'ingénierie du matériel) de SNCF Voyageurs ainsi que la société Numalis et en particulier Cyril Cappi, Cédrick Lelionnais et Arnaud Ioualalen.

Les membres du groupe de travail les remercient vivement pour le temps qu'ils ont pu leur consacrer.

Le groupe de travail souhaite aussi remercier le groupement Anavid et Elghazel Conseil pour l'état des lieux réalisé sur l'utilisation d'algorithme d'intelligence artificielle dans différents secteurs industriels.

7. Bibliographie

Ian Goodfellow and Yoshua Bengio and Aaron Courville (2016). *Deep Learning*. MIT Press, Cambridge. http://www.deeplearningbook.org

Hervé Delseny, Christophe Gabreau, Adrien Gauffriau, Bernard Beaudouin, Ludovic Ponsolle, Lucian Alecu, Hugues Bonnin, Brice Beltran, Didier Duchel, Jean-Brice Ginestet, Alexandre Hervieu, Ghilaine Martinez, Sylvain Pasquet, Kevin Delmas, Claire Pagetti, Jean-Marc Gabriel, Camille Chapdelaine, Sylvaine Picard, Mathieu Damour, Cyril Cappi, Laurent Gardès, Florence De Grancey, Eric Jenn, Baptiste Lefevre, Gregory Flandin, Sébastien Gerchinovitz, Franck Mamalet, Alexandre Albore (2021). *White Paper Machine Learning in Certified Systems.* DEEL project. arXiv:2103.10529.

Jason Jo, Yoshua Bengio (2017). *Measuring the tendency of CNNs to Learn Surface Statistical Regularities.* arXiv:1711.11561.

Maël Pégny, Issam Ibnouhsein (2018). *Quelle transparence pour les algorithmes d'apprentissage machine ?* hal-01877760

Sitou Afanou, Cédrick Lelionnais (2021). *Les systèmes de « deep learning » pour l'embarqué ferroviaire*. In RGCF janvier 2021

Lucian Alecu, Hugues Bonnin, Thomas Fel, Laurent Gardes, Sébastien Gerchinovitz, et al.. Can we reconcile safety objectives with machine learning performances? ERTS 2022, Jun 2022, TOULOUSE, France. ⟨hal-03765471⟩

LNE. Certification standard of processes for AI – Design, development, evaluation and maintenance in operational conditions. Revision N° .2.0 – 12/07/2021

EASA – EASA Concept Paper: First usable guidance for level 1 machine learning applications. December 2021, issue 01.

Établissement public de sécurité ferroviaire
60, rue de la Vallée – CS 11758 – 80017 Amiens Cedex 1