L'IA explicable appliquée à la détection de ceintures et de téléphones au volant

M. Gornet[1], W. Maxwell[1]

[1] Télécom Paris, IP Paris - Institut Polytechnique de Paris
NOS - Numérique, Organisation et Société
i3 - institut interdisciplinaire de l'innovation

melanie.gornet@telecom-paris.fr

Résumé

Une nouvelle génération de dispositifs de détection d'infractions routières à base d'IA pour le contrôle du port de la ceinture, ou du téléphone au volant, est en cours de développement. Ces systèmes s'appuient sur des réseaux de neurones profonds pour classifier des images d'éventuelles infractions et envoyer la décision de l'algorithme à un opérateur humain pour vérification. Un cadre réglementaire existe pour les radars d'excès de vitesse classiques mais ce cadre doit être aménagé pour s'adapter aux particularités des nouveaux systèmes, notamment l'impossibilité d'un audit par essais et l'introduction de l'humain dans la boucle de contrôle des sanctions. L'objectif de cette étude est de comparer le fonctionnement et l'encadrement des radars classiques avec le fonctionnement des nouveaux systèmes, afin d'identifier les besoins en explicabilité de ces dispositifs à base d'IA. Nos conclusions pour ce cas d'usage peuvent s'appliquer à d'autres cas d'usages où des images sont proposées à des opérateurs humains, tels que les systèmes de vidéo intelligente.

Mots-clés

radar, explicabilité, humain dans la boucle, vidéo intelligente, infractions routières

Abstract

A new generation of AI-based tools is being developed to help identify road safety violations such as failure to wear a seatbelt, or use of a phone while driving. An existing regulatory framework applies to traditional speed radars, but this framework must be tailored to accommodate the unique features of new systems, like the inability to conduct audits through testing and the involvement of humans in the enforcement of sanctions. The objective of this study is to compare the functioning and regulatory framework of classic speed radar systems with the functioning of new image recognition systems in order to identify needs for explainability of the new AI-based systems. Our conclusions in this use case can be generalized to other "smart video" use cases where images are proposed to human operators.

Keywords

radar, explainability, human in the loop, smart video, traffic violations

1 Introduction

Depuis quelques années les systèmes dits d'intelligence artificielle [1], ou IA, se sont multipliés. Notamment, les systèmes d'apprentissage profond [2] sont utilisés pour les tâches de reconnaissance d'images : ils sont alors appelés réseaux de neurones convolutionnels [3].

D'autre part, de nombreux textes, réalisés par des experts internationaux comme le Groupe d'Experts de Haut Niveau sur l'Intelligence Artificielle (GEHN IA [4]) de la Commission Européenne, appellent à prendre en compte un certain nombre de principes éthiques lors du cycle de vie des systèmes d'IA. Parmi ces principes, l'explicabilité et la transparence sont des préoccupations principales, elles apparaissent dans la quasi intégralité des textes sur l'éthique de l'IA [22], ainsi que dans le projet de règlement européen sur l'intelligence artificielle, aussi appelé *AI Act* [17]. En effet, les systèmes d'apprentissage, et notamment les réseaux neuronaux, sont très opaques et le nombre considérable de paramètres rend la décision d'autant plus compliquée à interpréter.

Pour caractériser l'explicabilité et la transparence, nous considérerons ici les définitions données par le GEHN IA qui, dans ses lignes directrices, définit l'explicabilité comme *« la capacité d'expliquer à la fois les processus techniques d'un système d'IA et les décisions humaines qui s'y rapportent »* et la transparence comme *« l'exigence selon laquelle les systèmes d'IA doivent être conçus et mis en œuvre de manière à en permettre leur supervision / leur suivi »*. La transparence apparaît alors plus générale, comprenant l'explicabilité des systèmes et des décisions, mais aussi la traçabilité et la communication [33]. Elle est davantage liée à l'accès à l'information, alors que l'explicabilité

1. Le terme « intelligence artificielle » est très controversé, notamment pour son caractère anthropomorphe, c'est-à-dire rappelant des traits humains [42].
2. Ou *deep learning* en anglais.
3. Ou *convolutional neural networks* en anglais.
4. Plus connu sous le nom de *HLEG* en anglais.

est lié au contenu de l'information, notamment sa compréhensibilité et sa justesse [28]. Par exemple, là où la transparence exigerait de disposer des poids d'un réseau de neurones, l'explicabilité permet d'interpréter ces chiffres et le résultat qui en découle.

Nous nous intéressons dans cette étude au cas d'usage des dispositifs de détection d'infractions routières. En effet, si les premiers systèmes de radar fonctionnaient par effet Doppler, les plus récents permettent de détecter si une voiture franchit la ligne d'un feu rouge [5] ou de discriminer entre une voiture et un camion dans le cas de limites de vitesse différentes [6]. Aujourd'hui, une troisième génération de dispositif est en cours de développement par les entreprises françaises, qui permettraient de détecter si un conducteur porte bien sa ceinture de sécurité et ne téléphone pas au volant [21] à l'aide d'apprentissage profond. Toutefois, les performances et la chaîne de décision de ces nouveaux systèmes sont différents : ils sont moins précis et un opérateur humain est en charge de vérifier la sortie du système et de sanctionner ou non l'automobiliste. Le constat de l'infraction est effectué par l'humain, même si celui-ci s'appuie sur des images recommandées au préalable par le système. Les dispositifs de vidéos intelligentes adoptent généralement ce schéma : le système propose, et l'humain décide.

Dans cette étude, nous examinons les besoins en explicabilité de ces nouveaux dispositifs à base d'apprentissage profond. Nous étudions le cas des radars routiers classiques afin de soulever les différences avec le nouveau dispositif.

Les différents types de radars actuels, ainsi que leurs exigences techniques et réglementaires sont exposés en Section 2. En Section 3, nous présentons les dispositifs de nouvelle génération à base d'IA et montrons que les anciennes méthodes d'évaluation, fondées sur des essais et calculs de performance, ne sont pas adaptées à ces nouveaux systèmes. Nous expliquons également comment le changement dans la chaîne de décision, avec l'introduction du contrôle humain, fait passer le besoin d'explicabilité de la machine à l'homme. En Section 4, nous analysons les textes réglementaires à la recherche d'exigences d'explicabilité pour les systèmes d'IA et montrons qu'aujourd'hui elles sont davantage semblables à des exigences de transparence. En Section 5, nous revenons au cas d'usage des dispositifs de détection d'infractions, et discutons du rôle que pourrait jouer l'explicabilité dans cette chaîne de commande. Enfin, nous discutons en Section 6 de ce qui peut être retenu de l'étude de ce cas d'usage et donnons quelques recommandations, utilisables pour ces dispositifs mais aussi pour d'autres systèmes tels que les dispositifs de vidéos intelligentes qui seront expérimentés lors des Jeux Olympiques de Paris en 2024.

5. Voir les radars feu rouge, sur le site de la sécurité routière : https://www.securite-routiere.gouv.fr/radars/differents-types-de-radars/radars-fixes/radars-de-franchissement

6. Voir les radars discriminants, sur le site de la sécurité routière : https://www.securite-routiere.gouv.fr/radars/differents-types-de-radars/radars-fixes/radars-de-controle-de-la-vitesse-fixes

2 Le fonctionnement et l'encadrement des radars routiers actuels

2.1 Une diversité de radars qui utilisent parfois des techniques d'apprentissage machine

2.1.1 Types de radars actuels

Nous connaissons bien aujourd'hui les radars routiers qui flashent l'automobiliste imprudent lorsque la vitesse de son véhicule dépasse un certain seuil. Mais cet appareil est loin d'être le seul type de radar existant. Nous allons ici nous intéresser au cas des radars fixes, bien que l'on notera tout de même l'existence de radars déplaçables, destinés à être disposés par exemple sur des chantiers, pour couvrir les zones de danger, ainsi que de radars mobiles, débarqués au bord de la route ou embarqués dans une voiture radar. Les radars fixes comprennent : les radars de contrôle de la vitesse fixes, aussi appelés cinémomètres, les radars de franchissement, et les radars pédagogiques [7]. Ces derniers, visent à inciter les usagers à ralentir sans les verbaliser, en affichant par exemple un symbole « danger » ou leur vitesse en rouge si elle est trop élevée. Nous les écartons pour notre discussion puisqu'ils ne mènent pas à un processus de contravention. Il reste alors deux types de radars : les radars vitesse et les radars de franchissement.

Les plus anciens radars routiers sont les radars de vitesse fixes qui calculent instantanément la vitesse d'un véhicule à son passage et sanctionnent lorsque la vitesse dépasse la limite autorisée. Plusieurs variantes de ces systèmes existent comme les radars qui permettent de différencier entre plusieurs catégories de véhicules, par exemple entre poids lourds et véhicules légers, pour contrôler des limites de vitesses différentes. Les radars de vitesse moyenne, communément appelés radars tronçon, permettent, quant à eux, de calculer la vitesse moyenne d'un véhicule sur une portion de route [8]. Le dernier type de radar, les radars de franchissement, comprennent les radars feux rouges, positionnés au niveau des carrefours routiers et prenant une photo si un véhicule franchit la ligne du feux ou s'il poursuit sa route au delà du feu, et les radars de passage à niveau [9].

2.1.2 Certains radars intègrent des modèles d'apprentissage machine

S'il est impossible de connaître les détails du fonctionnement des systèmes de radars actuels pour des raisons de propriété intellectuelle des entreprises, nous savons néanmoins que les techniques de détection ont bien évoluées depuis les radars à effet Doppler.

7. Voir la liste des différents types de radars sur le site de la sécurité routière : https://www.securite-routiere.gouv.fr/radars/differents-types-de-radars

8. Voir la liste des différents radars de contrôle de la vitesse fixes sur le site de la sécurité routière : https://www.securite-routiere.gouv.fr/radars/differents-types-de-radars/radars-fixes/radars-de-controle-de-la-vitesse-fixes

9. Voir la liste des différents radars de franchissement sur le site de la sécurité routière : https://www.securite-routiere.gouv.fr/radars/differents-types-de-radars/radars-fixes/radars-de-franchissement

Les techniques employées aujourd'hui relèvent de l'apprentissage machine « classique », notamment l'utilisation de *features extractor* pour détecter des objets dans une image. Une des méthodes les plus répandue se nomme *Histogram of Oriented Gradients* (HOG). Elle consiste à créer un histogramme en utilisant les gradients et les orientations des valeurs des pixels de l'image : l'amplitude est plus élevée lorsqu'il y a un changement brusque d'intensité, comme sur les bords, ce qui permet de détecter l'objet - ici la voiture. Ces techniques permettent aux radars actuels d'atteindre une très forte précision : leur taux d'erreur est d'environ 10^{-6}.

2.1.3 La décision des radars est automatique et est présumée juste

En France, ces radars sont encadrés par la loi n° 2003-495 du 12 juin 2003, renforçant la lutte contre la violence routière [6], qui autorise la mise en place de cinémomètres fixes et l'édition des procès-verbaux de façon automatique à partir de photographies numériques. La gestion du traitement automatisé des infractions routières est confié depuis 2011 à l'Agence Nationale de Traitement Automatisé des Infractions (ANTAI) [10]. Ainsi, les radars actuels détectent automatiquement, c'est-à-dire sans contrôle humain, que les véhicules dépassent la limite de vitesse et envoient l'infraction directement au poste de police [35]. Le message d'infraction contient notamment la vitesse estimée, l'heure, le numéro du radar, ainsi que la plaque d'immatriculation lue par le radar.

Le Code de la route précise dans son Article L.130-9 que *« Lorsqu'elles sont effectuées par ou à partir des appareils de contrôle automatique ayant fait l'objet d'une homologation, les constatations relatives aux infractions dont la liste est fixée par décret en Conseil d'État font foi jusqu'à preuve du contraire »*. Cela signifie qu'il y a présomption de justesse du résultat : en cas de contestation, il revient à l'utilisateur de prouver que le résultat est faux et que la sanction est infondée. Cela ne laisse place qu'à quelques possibles cas de contestation : par exemple si le conducteur peut démontrer que son véhicule ne pouvait pas atteindre la vitesse qui lui est reprochée, si le véhicule a été volé, cédé ou détruit, ou en cas d'usurpation de plaque d'immatriculation [35].

Le fonctionnement des radars actuels est résumé en Figure 1.

2.2 L'encadrement des radars actuels

2.2.1 Une homologation obligatoire par le LNE

Selon le décret n° 2001-387 du 3 mai 2001 relatif au contrôle des instruments de mesure [5], les radars routiers sont soumis à plusieurs examens, listés en article 4. La première vérification, appelée examen de type ou homologation, permet l'approbation du modèle du système qui sert ensuite de référence à la production des autres appareils. La seconde, la vérification primitive, consiste en un contrôle de chaque appareil avant la mise en service. Une vérification de l'installation sur site est ensuite effectuée. Enfin, le contrôle en service, ou vérification périodique, est réalisé à intervalles réguliers pendant toute la durée de vie de l'appareil [11].

Ces examens sont principalement [12] effectués par le Laboratoire National de Métrologie et d'Essais (LNE) [13], et, dans le cas de l'homologation, aboutissent à la délivrance d'un certificat. Ces certificats sont trouvable sur le site du LNE, souvent accompagnés d'annexes décrivant brièvement le fonctionnement du système.

Les radars routiers sont soumis à une série de vérifications pour garantir l'exactitude de leurs mesures. Les processus de vérification varient selon les pays d'Europe. Par exemple, l'Allemagne se base sur de l'audit logiciel. En France, ces vérifications sont effectuées dans le cadre de l'examen d'homologation et reposent sur des essais réalisés sur le système final. Le but général des essais est de garantir le bon fonctionnement des systèmes et le respect des exigences réglementaires. Nous nous concentrons dans cette analyse sur le cas français.

2.2.2 Les marges d'erreur définies par arrêté

L'arrêté du 4 juin 2009 relatif aux cinémomètres de contrôle routier [3], précise les modalités des examens dans le cas des radars de vitesse. Il contient un descriptif de l'examen d'homologation, qui comporte : un examen de conformité du dossier, des essais en laboratoire, dans les conditions de fonctionnement, des essais en fonctionnement réel dans des conditions normales d'utilisation sur route, ainsi que *« des analyses de simulation pour les situations dangereuses qui ne peuvent être reproduites lors des essais sur route »* (art. 9). La liste des essais minimaux à réaliser en laboratoire est également fournie. Parmi eux se trouvent notamment : la courbe d'erreurs en fonction de la vitesse, l'exactitude de la valeur des vitesses, des exigences sur les affichages, sur les tolérances en termes de température, d'humidité ou de choc (annexe III).

L'arrêté fourni, en outre, les exigences essentielles de construction des cinémomètres, comme les erreurs maximales tolérées sur les mesures de vitesses : 5 km/h pour des vitesses inférieures à 100 km/h et 5% de la vitesse pour des vitesses supérieures (art. 6). Ce 5 (km/h et %) est remplacé par un 3 pour les appareils neufs fixes, 7 pour les appareils neufs dans un véhicule en mouvement et 10 pour les appareils en service dans un véhicule en mouvement (art.5 et 6).

Les radars de franchissement, détectant si un véhicule ne respecte pas l'arrêt à un feux rouge, sont soumis à une réglementation différente : celle de l'arrêté du 18 janvier 2012 relatif à l'homologation des systèmes de contrôle au-

10. https://www.antai.gouv.fr/

11. Voir les différents examens des radars sur le site de la sécurité routière : https://www.securite-routiere.gouv.fr/radars/fonctionnement-des-radars/verification-des-radars-de-vitesse

12. Pour les radars de franchissements, seule l'homologation est réalisée par le LNE, les autres vérifications sont menées par des agents habilités des Centres d'Etudes Techniques de l'Equipement : https://www.securite-routiere.gouv.fr/radars/fonctionnement-des-radars/verification-des-radars-de-franchissement-de-feux

13. https://www.lne.fr/fr

tomatisé de franchissement d'une signalisation lumineuse fixe ou clignotante [1]. Les examens nécessaires sont semblables à ceux des cinémomètres et détaillés dans l'article 3.

L'arrêté fourni la liste des spécifications techniques nécessaires à l'homologation de ces systèmes. Notamment, l'article 30 stipule qu'*« aucune fausse détection n'est autorisée, ce qui signifie qu'aucun véhicule doit être contrôlé alors qu'il n'aurait pas dû l'être »*, et l'article 31 de rajouter que *« le pourcentage maximum de non détection doit être inférieur ou égal à 10% »*. De plus, *« le pourcentage de plaques lisibles [...] doit être supérieur ou égal à 95% hors cas de masquage »* (art. 35).

Ces exigences reflètent le choix de ne tolérer aucun faux positif, à savoir une infraction signalée à tort, même si cela conduirait à baisser le nombre de « vraies » infractions détectées. Toutefois, ces taux *« s'applique[ent] à toute série [...] consécutive d'au moins 100 véhicules contrôlés par le système »* (art. 31 et 35), ce qui ne suffit pas à garantir qu'aucune fausse détection n'est envoyée sur un grand nombre d'essais.

2.2.3 Une documentation complète sur la conception et le fonctionnement du système

L'arrêté de 2009 détaille la liste des pièces justificatives nécessaires à la demande d'homologation des cinémomètres, qui s'apparentent à des exigences de transparence. Ainsi, il est exigé de fournir : un projet de manuel d'utilisation, un carnet métrologique, le logiciel et ses documents ainsi que *« le détail de la détermination d'un résultat et le calcul d'incertitude associé, les facteurs d'incertitude pris en compte et les limites imposées à certains paramètres »* (art. 8). La liste des informations minimales que doit contenir le carnet métrologique est donnée en annexe II.

D'autres exigences de transparence sont stipulées par l'arrêté du 31 décembre 2001 [2], complétant le décret n° 2001-387 relatif au contrôle des instruments de mesure, qui stipule que la demande d'examen de type doit être accompagnée d'un dossier contenant : une notice explicative sur le fonctionnement de l'appareil, ses caractéristiques métrologiques, des plans de conception et de fabrication avec descriptions, les résultats des calculs de conception et des contrôles, la plaque d'identification et de poinçonnage, le plan de scellement (art. 5)

Ainsi, les exigences définies dans le cadre de l'homologation portent principalement sur la transparence et l'accès à l'information, notamment vis-à-vis de la communication aux laboratoires d'essais. Il n'est en effet pas vraiment question d'explicabilité. Les performances de l'appareil sont simplement vérifiés en terme de précision sur la vitesse par le biais du calcul d'incertitude de mesure. Si l'incertitude est suffisamment basse, l'homologation est approuvée sans explication supplémentaire. Ces exigences de transparence, pour chacun des acteurs, sont illustrées par les flèches vertes de la Figure 1.

Les exigences de transparence pour les radars classiques se rapprochent de la documentation technique et la notice d'utilisation exigées par le projet de règlement européen *AI Act*.

3 Nouveaux dispositifs ceinture/ téléphone

3.1 Fonctionnement des nouveaux dispositifs

Aujourd'hui de nouveaux systèmes sont en cours de développement en France, et déjà déployés en Australie. Ils permettent, une fois placés sur le bord de la route, de détecter si un conducteur ne porte pas de ceinture de sécurité ou s'il téléphone au volant [32], des obligations explicitées par les articles R.412-1 et R.412-6-1 du Code de la route [**?**]. En effet, un dixième des accidents de la route sont dû à l'utilisation d'un téléphone au volant [21].

Ces nouveaux dispositifs ceinture/ téléphone utilisent des techniques d'apprentissage profond sur des réseaux neuronaux entraînés à partir d'images récoltées par des dispositifs tests aux alentours de Paris. Le système, une fois déployé, analyse les images des véhicules à partir du flux vidéo pour détecter les infractions. Si le système détecte une infraction, l'image est alors envoyée au Centre National de Traitement des infractions routières (CNT) à Rennes, où des officiers de police du Centre Automatisé de Constatation des Infractions Routières (CACIR) vérifient l'image et constatent l'infraction le cas échéant. La vérification humaine sert donc à valider l'existence de l'infraction à partir de l'image. Pour les radars d'excès de vitesse, cette validation n'est pas possible, car l'humain ne peut pas indépendamment vérifier la mesure de la vitesse. Il doit se fier entièrement à l'appareil. C'est l'appareil, non l'humain, qui constate l'excès de vitesse. Dans le cas des dispositifs ceinture/ téléphone, le système propose des candidats, mais c'est bien l'agent de police qui décidera *in fine* si une infraction est effectivement constatée. Le paramétrage du dispositif permettra d'ajuster la qualité des candidats envoyés selon des seuils de confiance. Plus le seuil de confiance est élevé, plus l'examen du vérificateur humain deviendra une pure formalité, soulevant des risques de biais d'automatisation.

L'article R.130-11 du Code de la route stipule que : *« Font foi jusqu'à preuve du contraire les constatations, effectuées par ou à partir des appareils de contrôle automatique ayant fait l'objet d'une homologation, relatives aux infractions sur : 1° Le port d'une ceinture de sécurité homologuée dès lors que le siège qu'il occupe en est équipé prévu [...] ; 2° L'usage du téléphone tenu en main [...] »*. Par conséquent, la constatation de l'officier de police du non port de la ceinture ou de l'utilisation du téléphone au volant à partir des images délivrées par le système est présumée juste. Dès lors, les conducteurs sanctionnés devront être en mesure de prouver qu'ils respectaient bien la législation s'ils veulent contester la décision. Cette présomption de justesse est similaire à celle utilisée pour les radars automatiques actuels, à la différence que c'est la décision humaine qui est supposée être juste, et non la décision automatisée du radar.

Le fonctionnement des dispositifs ceinture/ téléphone est illustré en Figure 2.

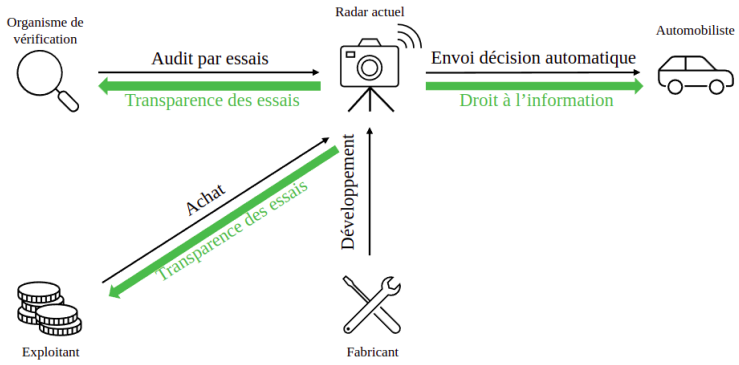

FIGURE 1 – Fonctionnement des radars actuels (en noir) et exigences de transparence (en vert)

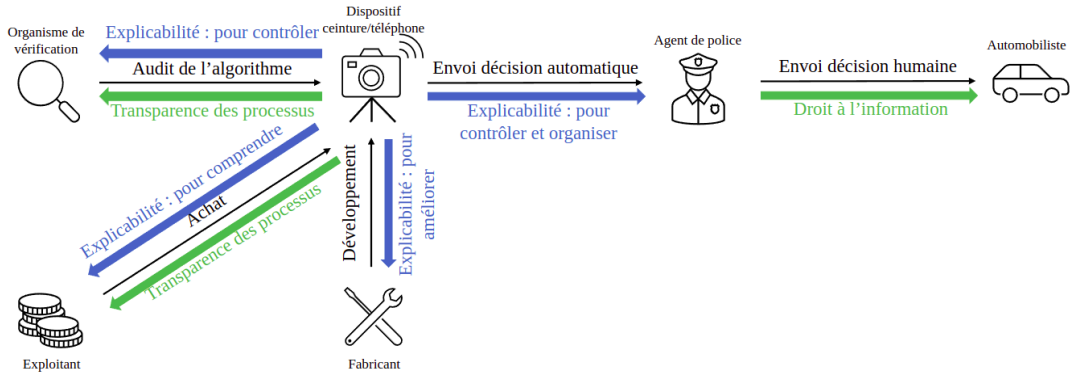

FIGURE 2 – Fonctionnement des dispositifs ceinture/ téléphone (en noir) et exigences de transparence (en vert) et d'explicabilité (en bleu)

3.2 Les nouveaux dispositifs soulèvent des questions nouvelles, notamment par rapport à l'explicabilité

3.2.1 Des modèles opaques d'apprentissage profond

La première différence concerne la technologie utilisée. Ici l'apprentissage profond renforce le nombre de paramètres et donc l'opacité du système. Alors que l'homologation des radars actuels ne requiert qu'une simple transparence du processus [14], l'homologation des futurs dispositifs devra évoluer vers l'explicabilité.

En effet, la transparence totale d'un réseaux de neurones est peu utile : qui aimerait recevoir une liste de millions ou milliards de paramètres ? Une transparence partielle est toutefois possible et utile, par exemple en documentant la planification du processus de conception, les données et annotations utilisées, le type d'algorithme choisi, le processus d'apprentissage, les domaines d'usages, le protocole d'évaluation du modèle, les méthodes de gestion des cas d'erreurs, le maintien en condition opérationnelles... Tous ces éléments se retrouve dans le « Référentiel de certification de processus pour l'IA » [15] développé récemment par le LNE. Des exigences similaires figurent dans le projet de règlement européen *AI Act*, dans son annexe IV. Toutefois, le référentiel du LNE discute de la certification d'un processus et non du système en lui même. Or connaître les données d'entraînement et de test ne donne pas nécessairement d'information sur la performance globale du système, par exemple. On pourrait notamment souhaiter savoir si le dispositif de détection d'infractions est sujet à des biais qui pourraient compromettre son équité, tels que la détection plus facile de certaines infractions sur des types spécifiques de véhicules ou de morphologies.

Le problème pour les nouveaux systèmes est également plus complexe, car le même système doit gérer à la fois la détection des ceintures et des téléphones. Or, le problème ceinture/ téléphone est asymétrique : dans un cas, l'infraction est envoyée si un objet est détecté (le téléphone), dans l'autre, s'il ne l'est pas (la ceinture). Il est légitime de se demander si ces deux problèmes opposés ne requièrent pas des règles adaptées à chacun. Nous avons vu avec l'exemple des radars de vitesse et de franchissement que les exigences de précision visent à éviter les erreurs conduisant à accuser à tort le conducteur de dépasser la limite de vitesse ou de brûler un feu rouge (des faux positifs). Dans le cas ceinture/ téléphone, deux types d'erreurs peuvent être envisagés : les faux positifs, où l'objet est détecté alors qu'il n'est en réalité pas présent, et les faux négatifs où l'objet n'est pas détecté alors qu'il est présent. Dans le cas de la ceinture, l'erreur problématique où le conducteur est faussement accusé correspond à un faux négatif. En revanche, pour le téléphone, cette erreur se traduit par un faux positif. Les autres types d'erreurs, à savoir les faux positifs pour la

14. Voir Section 2.2.3

ceinture et les faux négatifs pour le téléphone, peuvent être considérés comme importants pour le policier qui manque une infraction, mais semblent moins problématiques vis-à-vis des droits de la personne accusée. Compte tenu de la présomption de l'existence de l'infraction créée par l'article R.130-11 du Code de la route, on peut imaginer que le système sera paramétré pour éviter tout risque de fausse accusation, et que les opérateurs humains ne constateront l'infraction que si l'image signalée par le système ne laisse aucun doute. Mais si le système n'envoie que des images d'infractions manifestes, la vigilance du contrôleur humain diminuera, celui-ci devenant complaisant devant la performance quasi-infaillible de l'algorithme [15].

3.2.2 Performance du système

L'homologation est le processus soulevant le plus de questions pour les nouveaux systèmes de détection de ceinture/ téléphone. En effet, elle reposait jusqu'à présent sur des essais réalisables grâce à la précision élevée des systèmes radars. Or, en moyenne, la précision est bien moindre sur les nouveaux dispositifs à base d'IA que sur les anciens : le taux d'erreur peut atteindre l'ordre du pourcentage, voire de la dizaine de pourcent. Ces différences rendent ces nouveaux produits plus difficilement certifiables par essais qui requièrent de passer plusieurs centaines de tests sans échouer.

Un autre problème technique concerne le calcul d'incertitude. Pour les anciens radars, cette incertitude est une incertitude de mesure : elle est lié à l'instrument de mesure qu'est le radar. Dans le cas des dispositifs ceinture/ téléphone, l'incertitude est liée au réseau de neurones. Pour calculer l'incertitude du modèle d'apprentissage, des statistiques sur la distribution prédictive sont souvent utilisées. Dans le cas de la classification, cette distribution est composée de probabilités de classe qui représentent intuitivement le degré de confiance dans un résultat [12].

De plus, les résultats des tests sont considérablement influencés par les conditions d'essai, ce qui peut entraîner des variations importantes de performance en fonction de facteurs tels que la météo, l'éclairage et l'angle de vue, là où avant ces écarts étaient négligeables.

La possibilité de paramétrer les dispositifs pour qu'ils n'envoient des images qu'en cas de confiance élevée sera déterminante pour mesurer la qualité du système. Les opérateurs humains devraient recevoir des images seulement de cas manifestes, où l'infraction sautent aux yeux.

Aujourd'hui, le LNE se penche sur un nouveau processus d'homologation pour ces dispositifs. Si les exigences exactes ne sont pas encore publiques, elles risquent fort de s'inspirer de leur « Référentiel de certification de processus pour l'IA » [15]. L'approche envisagée relèverait alors plus de la certification d'un algorithme, que de celle d'un système complet en conditions de déploiement. Or, l'évaluation d'un système doit tenir compte non seulement des données et paramètres d'entraînement, mais également du bon fonctionnement du contrôle humain [29].

15. Voir Section 3.2.3

3.2.3 Assurer la qualité du contrôle humain

Alors que les radars actuels sont entièrement automatisés, le nouveau système ceinture/ téléphone ne fait qu'envoyer une photo lorsque le modèle détecte une infraction mais laisse à l'officier de police la responsabilité de décider si oui ou non l'image représente bien une infraction. Ce processus d'humain dans la boucle [16] est fondamental d'un point de vue légal aujourd'hui puisqu'il entraîne la responsabilité de l'individu qui prend la décision.

Néanmoins, les systèmes d'aide à la décision peuvent engendrer des biais chez les utilisateurs humains, surtout lorsque les recommandations générées par le système sont de très grande qualité. Si le système n'envoie que des images d'infractions manifestes, les humains deviendront moins vigilants, réduisant ainsi les bienfaits du contrôle humain. Ce phénomène conduit à des décisions qui ne sont humaines qu'en apparence et dissimulent une décision entièrement algorithmique. Pour autant, ce contrôle reste une garantie nécessaire selon la grande majorité des textes sur l'éthique algorithmique [22]. La Cour de Justice de l'Union Européenne exige un contrôle humain de chaque résultat algorithmique notamment dans le cadre de systèmes qui détectent des risques de terrorisme [14]. De plus, l'article 47 de la loi Informatique et Liberté (LIL) du 6 janvier 1978 [7] exclut l'utilisation de l'apprentissage machine pour des décisions entièrement automatisées de l'administration, ce qui rend indispensable un contrôle humain effectif de chaque décision.

4 Exigences de transparence et d'explicabilité pour l'IA

Nous avons montré que les nouveaux dispositifs ceinture/ téléphone sont très différents des radars classiques, ce qui peut nécessiter des exigences différentes de transparence et d'explicabilité. Mais qu'est-il aujourd'hui prévu dans la loi pour les systèmes d'IA ?

4.1 Code des Relations entre le Public et l'Administration

Le Code des Relations entre le Public et l'Administration (CRPA) impose un certain nombre d'obligations en matière de transparence lorsque l'administration prend une décision individuelle sur le fondement d'un traitement algorithmique [17]. D'une part, l'administration doit publier en ligne les règles définissant les principaux traitements utilisés dans l'accomplissement de ses missions lorsqu'ils fondent des décisions individuelles. D'autre part, l'administration doit communiquer à l'intéressé, à sa demande, les informations suivantes *« 1° Le degré et le mode de contribution du traitement algorithmique à la prise de décision ; 2° Les données traitées et leurs sources ; 3° Les paramètres de traitement et, le cas échéant, leur pondération, appliqués à la situation de l'intéressé ; 4° Les opérations effectuées par le*

16. Ou *human-in-the-loop* en anglais.

17. Pour le détail des exigences de transparence applicables aux algorithmes publics, voir le guide d'Etalab : `https://etalab.github.io/algorithmes-publics/guide.html`

traitement ». Ces informations doivent être communiquées *« sous une forme intelligible »*.
Ces obligations soulèvent des questions particulières lorsqu'il s'agit des systèmes de détection ceinture/ téléphone. En premier lieu, la communication de l'image fondant l'infraction est prévue par le CRPA qui permet un accès aux documents administratifs. Mais l'exigence forte concerne surtout la délivrance des *« paramètres de traitements et, le cas échéant, leur pondération, appliqués à la situation de l'intéressé »*. Dans le cadre de modèles d'apprentissage profond ces paramètres peuvent correspondre par exemple aux valeurs des poids du réseau de neurones ou encore aux hyperparamètres d'apprentissage. Les paramètres de traitement doivent correspondre à ceux appliqués à la situation de l'intéressé, donc les paramètres et les pondération appliqués pour classifier l'image particulière de l'infraction. Cette exigence nécessiterait de préserver, pour chaque image conduisant à une décision, une copie du modèle au moment de la décision algorithmique, ainsi qu'une copie de l'image, afin de permettre une réplicabilité de la décision. La mention de « forme intelligible » suggère que les paramètres et leur pondération doivent être compréhensibles, ce qui est loin d'être évident pour un modèle complexe.

4.2 Directive Police-Justice et la loi Informatique et Liberté

Le dispositif ceinture/ téléphone constitue un traitement de données à caractère personnel à des fins de détection d'infractions pénales, ce qui relève des dispositions de la LIL qui transposent la Directive européenne « Police-Justice » 2016/680 du 27 avril 2016 [4]. S'agissant d'une nouvelle technologie, le dispositif ceinture/ téléphone nécessiterait probablement une analyse d'impact au titre de l'article 90 de la LIL, une analyse qui serait transmise à la CNIL. Dans le cadre de cette analyse, des obligations particulières de transparence et d'explicabilité du dispositif pourraient être identifiées comme étant nécessaires pour la protection des droits des personnes.

4.3 Projet de Règlement européen sur l'intelligence artificielle *AI Act*

Le projet de règlement *AI Act* impose des obligations particulières de transparence aux fournisseurs de systèmes d'IA à haut risque. La liste des applications à haut risque (Annexe III du règlement) n'inclut pas explicitement des dispositifs de reconnaissance d'image non-biométriques, même pour des applications répressives. Ainsi, en l'état actuel du projet, le dispositif ceinture/ téléphone ne semble pas être considéré comme étant une application à haut risque au regard du futur règlement [18]. Ce dispositif ne tomberait pas non plus dans la liste des applications interdites par le projet de règlement, ou dans la liste des applications d'IA nécessitant des mesures de transparence particulières. Mais les dispositions du projet sont susceptibles d'évoluer en fonction des négociations du texte.

18. L'Annexe III listant les systèmes à haut risque mentionne en 6.a *« les systèmes d'IA destinés à être utilisés par les autorités répressives pour mener des évaluations individuelles des risques visant à déterminer la probabilité qu'une personne physique commette une infraction ou récidive, ou le risque encouru par les victimes potentielles d'infractions pénales »*. Toutefois, nous considérons que notre cas d'usage ceinture/ téléphone ne rentre pas dans cette catégorie, car il ne constitue pas une évaluation des risques et ne calcule pas la probabilité d'un individu à commettre une infraction.

5 Définir les besoins d'explicabilité pour les dispositifs ceinture/ téléphone

Les dispositifs ceinture/ téléphone nécessitent des exigences d'explicabilité supplémentaires qui viennent s'ajouter aux exigences de transparence prévues pour les systèmes de radars actuels. Elles sont illustrées sur la Figure 2 par les flèches bleues. Dans cette section, nous détaillons ces besoins en explicabilité.

5.1 Les destinataires et les finalités des explications

Les mesures d'explicabilité doivent correspondre aux besoins de chaque public visé par l'explication. En effet, l'explicabilité doit être adaptée à la catégorie d'acteurs à laquelle elle s'adresse [26]. Il est possible de distinguer notamment les experts des non-experts [36] car le niveau d'expertise de l'acteur peut influer sur sa capacité à comprendre l'explication donnée [45]. Il faut alors trouver la juste mesure et adapter le niveau d'explicabilité à la situation donnée et au niveau d'expertise de la personne concernée [10]. Un expert aura par exemple besoin d'une explication plus en profondeur, avec un examen exhaustif du modèle qui pourra être plus complexe. Au contraire, pour des non-experts, l'explication pourra être plus brève, plus visuelle et peu complexe, pour être facilement compréhensible.
Il est possible d'identifier plusieurs catégories d'acteurs, telles que les créateurs, les opérateurs, les exécutants, les personnes ciblées, et les examinateurs/ régulateurs [43]. De même, il existe quatre raisons principales pour demander une explication : pour justifier, pour contrôler, pour améliorer, pour découvrir [8].
Nous souhaitons évaluer les besoins d'explicabilité pour chacun des publics, pour les dispositifs ceinture/ téléphone. Le premier public est constitué des personnes ciblées par la décision. Celles-ci auront besoin d'informations pour éventuellement contester la décision, et vérifier s'il ne s'agit pas d'une erreur. Comme pour les radars classiques, la communication de l'image pourrait suffire à démontrer que la plaque d'immatriculation est fausse, ou à démontrer que l'image est complètement floue et ne pourrait pas fonder une constatation d'infraction. Au-delà de l'image, il est difficile d'identifier l'utilité d'autres explications, et notamment les paramètres et les pondérations prévues par le CRPA. L'humain est capable de classifier une image issue de la vie courante aussi bien qu'un ordinateur, indépendamment de la compréhension des calculs de l'ordinateur. Pour détecter une erreur de classification, l'image suffit donc. Il ne s'agit alors plus d'une exigence d'explicabilité, mais d'une simple vérification des données d'entrée.

Si la contestation concerne la légalité du système dans son ensemble, la personne ciblée, ou bien une autorité de régulation, pourrait souhaiter comprendre si le système souffre de biais, par exemple un système qui privilégierait certains types ou couleurs de voitures. Dans ce cas, le dispositif pourrait violer le principe d'égalité devant la loi, conduisant à son illégalité. De même, une explication pourrait permettre de détecter certaines intentions malveillantes lors de la conception de l'algorithme, par exemple des signes distinctifs qui permettraient d'éviter la détection de l'infraction.

Les agents de police qui constatent les infractions auront besoin également de l'image pour valider ou rejeter la recommandation. L'image parlera d'elle-même la plupart du temps, l'agent n'ayant pas non plus besoin de comprendre les calculs du système pour savoir si l'image montre une infraction. Un explication visuelle pourrait consister à montrer la partie de l'image qui, selon l'ordinateur, caractérise l'infraction. Mais si l'image n'est pas claire quant à l'existence de l'infraction, l'agent ne pourra la constater.

Les images envoyées à tort à un agent humain seront en principe détectées par l'agent. Il y aura donc une trace de ces erreurs de classification, et leurs causes pourront être étudiées. Le cas des infractions qui échappent à la détection sera plus difficile. Ces cas de non-détection ne seront pas connus, et ne pourront pas être analysés avec des outils d'explicabilité. Or, les systèmes de reconnaissance d'images peuvent être trompés par l'ajout d'auto-collants ou films destinés justement à changer la classification [18]. On peut donc s'attendre à l'émergence de stratagèmes pour éviter des détections automatiques, à l'instar des dispositifs anti-radars. Un modèle pourra évidemment apprendre à reconnaître, et a éviter, ces stratagèmes anti-détection, s'ils sont connus.

Les régulateurs, l'exploitant, les organismes de certification auront besoin d'informations sur la conception du système, les données d'entraînement, de test, et de validation. Les informations sur les seuils de confiance, le niveau toléré de faux positifs et de faux négatifs, les tests de biais, et les choix de mesures de réduction de biais, seront nécessaires pour apprécier la performance du système.

Même si les dispositions du futur règlement *AI Act* ne s'appliquaient pas au dispositif, les informations listées dans l'Annexe IV du projet de règlement seraient utiles voire nécessaires pour que les régulateurs, l'exploitant du système, et les organismes de certification puissent apprécier les forces et les faiblesses du dispositif.

5.2 Adapter la forme des explications au public visé

Il existe une multitude de solutions possibles pour expliquer une image, mais ces dernières doivent être adaptées au public visé. En effet il y a un « bon » niveau d'explication à trouver pour chaque situation donnée [30]. De nombreuses solutions techniques existent mais elles sont surtout adressées à des spécialistes en IA, et pourraient détériorer la décision humaine au lieu d'aider si elles sont utilisées par des non-experts [44].

Les cas d'erreurs peuvent être étudiés notamment par l'utilisation d'outils d'explicabilité post-hoc, tels que SHAP [27] ou LIME [38] qui permettent une vérification rapide du point de focus de l'algorithme par la mise en avant des pixels les plus importants pour la décision. Toutefois, ces explications sont surtout destinées à des spécialistes en IA, elles seront plus difficiles à interpréter par des non-experts comme l'exploitant ou l'agent de police. De la même façon, pour aider à l'intelligibilité de la décision algorithmique, une carte de chaleur [19] peut montrer les parties de l'image qui ont le plus contribué à la classification. Différentes techniques peuvent être utilisées pour générer de telles explications : la descente de gradient comme la rétropropagation guidée [41] et la méthode Grad-CAM [40], ou encore les études de sensibilité comme les méthodes d'occlusion [46]. Les cartes de chaleur ont été critiquées car elles échouent aux tests de randomisation [20] et les explications sont donc indépendantes des données d'entraînement et des paramètres du modèle [9]. Elles ont également tendance à être similaires pour différentes prédictions [39]. Or, ce que l'on recherche c'est une explication qui soit propre à la fois à notre système et notre donnée d'entrée, pour refléter la décision prise, plutôt qu'une explication générique. En effet, on veut éviter d'avoir la même carte de chaleur lorsqu'il y a une ceinture et lorsqu'il n'y en a pas. Les explications sous forme de carte de chaleur n'aident donc pas forcément à prendre une décision. Elles peuvent toutefois servir à détecter des erreurs évidentes, comme un focus de l'algorithme ailleurs que sur la ceinture ou le téléphone. Elles sont déjà couramment utilisées pour la détection de biais dans les données, comme la présence de filigranes (aussi appelés *watermarks*) sur les images [11].

Pour aider l'agent à détecter la présence de la ceinture ou du téléphone, une simple boîte de délimitation (*bounding box*) pourrait être utilisée. Mais ces méthodes, basées sur la détection d'objets [37], seraient soit appliquées *a posteriori* et ne seraient alors pas représentatives de la décision du modèle, soit nécessiteraient une refonte du modèle d'apprentissage.

Pour comprendre une erreur sur un cas individuel, une explication contre-factuelle [21] pourrait également s'avérer utile. Pour expliquer pourquoi une image appartient à une certaine classe (par exemple, les classes « ceinture » ou « sans ceinture », respectivement pour le téléphone), une nouvelle image composite est générée à partir de l'image de départ et de changements locaux [19]. Ces changements sont les changements minimaux permettant de faire basculer l'image d'une classe à une autre. Cette catégorie d'ex-

19. Une carte de chaleur, ou *saliency maps* en anglais, se superpose aux pixels d'une image et la teinte selon leur importance dans la décision.

20. Un test de randomisation consiste à réinitialiser aléatoirement les paramètres d'un modèle ou ses données d'entrée et d'observer l'effet sur la sortie de l'algorithme.

21. Une explication contre-factuelle pose la question : « pourquoi pas ? » Elle permet de créer un cas imaginaire où la décision aurait été l'inverse et de voir quels éléments ont fait basculer cette décision. Par exemple dans le cas de la ceinture, on imagine qu'un contraste plus élevé entre les habits et la ceinture aurait permis de mieux distinguer : si le conducteur avait porté un t-shirt blanc, alors le système aurait détecté la ceinture.

plication pourrait être utilisée par les agents de police pour comprendre une décision sur un cas particulier.
Une explication à base de concepts [22] [34, 24, 23] permettrait peut-être de capturer des éléments comme la position du corps ou de l'objet ceinture/ téléphone. Ce type d'explication sera plus évidente à interpréter pour un expert tentant de comprendre le fonctionnement, et les erreurs, de l'algorithme. Néanmoins, il est possible que ces méthodes ne servent qu'à souligner la présence de l'objet, limitant ainsi l'explication. Il en va de même pour les explications à base de texte générées par des méthodes de *captioning* [20], qui servent peu dans le cas d'une détection binaire.
Enfin, des méthodes utilisant des prototypes [23] [25] pourraient également servir à des experts pour comprendre le comportement général d'un système mais sont peu convaincantes pour une décision individuelle.

6 Discussion et recommandations

Les nouveaux dispositifs ceinture/ téléphone utilisent des algorithmes à base d'apprentissage profond, ce qui pose la question de leur explicabilité. Ces dispositifs envoient des images d'infractions à des agents de police, qui examinent les images pour constater, ou non, une infraction. Ces dispositifs de reconnaissance d'image n'ont pas les mêmes performances techniques que les radars classiques, et peuvent commettre des erreurs. Cependant, seules les images ayant un taux de confiance élevé seront envoyées aux agents. Ce type de dispositif est similaire dans son fonctionnement à d'autres dispositifs de vidéo dits « intelligents », tels que ceux qui seront expérimentés lors des Jeux Olympiques de Paris en 2024. Ces dispositifs ont de nombreux usages, comme la détection automatique d'infractions, d'évènements « suspects » ou de colis abandonnés [13]. Ils peuvent également être utilisés dans le secteur privé pour la sécurité des bâtiments, des banques, des magasins, etc. Le flux vidéo est alors pré-analysé par le système d'IA et une alerte est levée lorsqu'un élément que le système a été programmé à détecter est repéré. Ce fonctionnement correspond, selon le projet de loi relatif aux jeux Olympiques et Paralympiques de 2024, à un « signalement d'attention » : « *[les traitements algorithmiques] ne produisent aucun autre résultat et ne peuvent fonder, par eux-mêmes, aucune décision individuelle ni aucun acte de poursuite. Ils demeurent en permanence sous le contrôle des personnes chargées de leur mise en œuvre* » (art. 7§6-7) [31]. En cela, la situation relative au contrôle humain est similaire pour ces systèmes de vidéo « intelligentes » que pour les dispositifs de détection ceinture/ téléphone. Nos conclusions seraient donc valables plus généralement pour tout système s'appuyant sur la vérification humaine pour la classification d'images.
Notre analyse des textes légaux, de la certification pour l'IA de confiance et du processus envisagé pour les dispositifs de détection ceinture/ téléphone nous amène à la conclusion que les exigences détaillées d'explicabilité contenues dans le CRPA ont peu d'intérêt pour la personne ciblée par une décision dans la mesure où l'image parle pour elle-même, et permet de détecter une erreur de classification.

Pour ces nouveaux dispositifs, il est nécessaire de réfléchir en amont aux besoins en explicabilité pour adapter le niveau de l'explication à la personne concernée et à l'objectif de l'explication. Dans le cadre des nouveaux dispositifs ceinture/téléphone, nous avons identifié plusieurs publics ayant des besoins différents d'explication. Pour les personnes ciblées, comme pour les agents en charge de constater les infractions, une communication de l'image suffira en général pour comprendre la classification. Contrairement à un score de risque, par exemple, une image parle pour elle-même.

Pour les concepteurs, les exploitants, les régulateurs, et les organismes de certification, les besoins en transparence et en explicabilité seront plus élaborés. Ces acteurs souhaiteront comprendre les classifications pour améliorer l'efficacité, l'équité, et la sécurité du système, et notamment contrôler et corriger les biais dans le modèle, et les erreurs individuelles. Pour ce faire, la documentation technique prévue par l'Annexe IV du projet de règlement européen *AI Act* sera utile et, dans la plupart des cas, nécessaire.

Une approche qui se contenterait de communiquer les paramètres et leur pondération, et/ou les caractéristiques de données d'entraînement, serait peu utile. Différentes techniques d'explicabilité individuelle - carte de chaleur, explications contre-factuelles, explication à base de concepts - aideront alors à comprendre les sources d'erreurs de classification. Une image suffira à constater l'erreur ; une explication individuelle sera nécessaire pour comprendre sa source.

Le concepteur et l'exploitant du système devront faire des choix de performance - prioriser la réduction de faux positifs par rapport aux faux négatifs ; définir des seuils de confiance. Ces choix devront faire l'objet de justifications particulières. Par exemple, si le système envoient des images avec un taux de confiance d'au moins 98%, pourquoi ce niveau de pourcentage et pas un autre ? Si le concepteur a fait des choix pour concilier biais et performance, comment justifier ces choix ? Si le système minimise les faux positifs mais en échange laisse passer beaucoup de faux négatifs, comment justifier cet équilibre ? Si le système n'envoie que des images avec un taux de confiance très élevé, comment lutter contre les biais humains liés à l'automatisation ?

La CNCDH recommande *« aux administrations de communiquer sous une forme intelligible les informations sur le fonctionnement de l'algorithme, ainsi que sur la part éventuellement prise par une intervention humaine dans le processus de décision. »* (Recommandation 19) [16]. L'explication du processus n'est alors pas seulement l'explication de la recommandation du système d'apprentissage mais de l'entièreté de la chaîne de décision.

22. Une explication à base de concepts cherche les attributs de l'image qui pourraient faire basculer la décision. Par exemple, un zèbre est classifié ainsi par ces rayures. On pourrait imaginer ici rechercher un concept de ceinture ou de bras levé à l'oreille dans le cas du téléphone.
23. Un prototype est une image "type" générée pour une classe donnée.

Remerciements

Cette recherche a été financée dans le cadre du projet LIMPID [24] (Projet ANR 20-CE23-0028).

Références

[1] Arrêté du 18 janvier 2012 relatif à l'homologation des systèmes de contrôle automatisé de franchissement d'une signalisation lumineuse fixe ou clignotante. NOR : DEVS1107065A.

[2] Arrêté du 31 décembre 2001 fixant les modalités d'application de certaines dispositions du décret n° 2001-387 du 3 mai 2001 relatif au contrôle des instruments de mesure. NOR : ECOI0200007A.

[3] Arrêté du 4 juin 2009 relatif aux cinémomètres de contrôle routier. NOR : ECEI0912713A.

[4] Directive (UE) 2016/680 du Parlement européen et du Conseil du 27 avril 2016 relative à la protection des personnes physiques à l'égard du traitement des données à caractère personnel par les autorités compétentes à des fins de prévention et de détection des infractions pénales, d'enquêtes et de poursuites en la matière ou d'exécution de sanctions pénales, et à la libre circulation de ces données, et abrogeant la décision-cadre 2008/977/JAI du Conseil. OJL 119.

[5] Décret n°2001-387 du 3 mai 2001 relatif au contrôle des instruments de mesure. NOR : ECOI0100116D.

[6] Loi n° 2003-495 du 12 juin 2003 renforçant la lutte contre la violence routière. NOR : EQUX0200012L.

[7] Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés.

[8] Amina Adadi and Mohammed Berrada. Peeking Inside the Black-Box : A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6 :52138–52160, 2018.

[9] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity Checks for Saliency Maps. *arXiv :1810.03292*, 2020.

[10] Valérie Beaudouin, Isabelle Bloch, David Bounie, Stéphan Clémençon, Florence d'Alché Buc, James Eagan, Winston Maxwell, Pavlo Mozharovskyi, and Jayneel Parekh. Flexible and Context-Specific AI Explainability : A Multidisciplinary Approach. *SSRN Electronic Journal*, 2020.

[11] Astrid Bertrand, Adam Pearce, and Nithum Thain. Searching for unintended biases with saliency. *PAIR Explorables*, 2022.

[12] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, Lama Nachman, Rumi Chunara, Madhulika Srikumar, Adrian Weller, and Alice Xiang. Uncertainty as a Form of Transparency : Measuring, Communicating, and Using Uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 401–413, New York, NY, USA, 2021. Association for Computing Machinery.

[13] CNIL. Caméras dites « intelligentes » ou « augmentées » dans les espaces publics. Position sur les conditions de déploiement. Technical report, 2022.

[14] Cours de Justice de l'Union Européenne. Arrêt de la Cour (grande chambre) du 21 juin 2022, Ligue des droits humains contre Conseil des ministres, Affaire C-817/19.

[15] Laboratoire National de Métrologie et d'Essais (LNE). Référentiel de certification de processus pour l'IA - Conception, développement, évaluation et maintien en conditions opérationnelles, 2021.

[16] Commission Nationale Consultative des Droits de l'Homme(CNCDH). Avis relatif à l'impact de l'intelligence artificielle sur les droits fondamentaux. Technical report, 2022.

[17] Commission Européenne. Proposition de règlement du Parlement Européenn et du Conseil établissant des règles harmonisées concernant l'Intelligence Artificielle (législation sur l'Intelligence Artificielle) et modifiant certains actes législatifs de l'Union. COM/2021/206 final, April 2021.

[18] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust Physical-World Attacks on Deep Learning Visual Classification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1625–1634, 2018.

[19] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual Visual Explanations. 2019.

[20] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. Generating Visual Explanations. *arXiv :1603.08507*, 2016.

[21] TF1 INFO. Excès de vitesse, pas de ceinture, téléphone au volant : gare aux nouveaux radars. JT 20h du 30 juillet 2021. URL : `https://www.tf1info.fr/societe/autoroute-exces-de-vitesse-pas-de-ceinture-telephone-au-volant-rien-n-echappe-aux-nouveaux-radars-urbains-2192686.html`.

[22] Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9) :389–399, 2019.

[23] Dmitry Kazhdan, Botty Dimanov, Mateja Jamnik, Pietro Liò, and Adrian Weller. Now You See Me (CME) : Concept-based Model Extraction. *arXiv :2010.13233*, 2020.

[24] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory

24. `https://limpid.telecom-paris.fr/`

Sayres. Interpretability Beyond Feature Attribution : Quantitative Testing with Concept Activation Vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning*, pages 2668–2677. PMLR, 2018.

[25] Sunnie S. Y. Kim, Nicole Meister, Vikram V. Ramaswamy, Ruth Fong, and Olga Russakovsky. HIVE : Evaluating the Human Interpretability of Visual Explanations. *arXiv :2112.03184*, 2021.

[26] Alexandra Kirsch. Explain to whom ? Putting the User in the Center of Explainable AI. In *Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML 2017 co-located with 16th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2017)*, Bari, Italy, 2017.

[27] Scott Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. *arXiv :1705.07874*, 2017.

[28] Aymeric Poulain Maubant. Pour y voir plus clair sur les notions de transparence et d'explicabilité en IA, 2020. URL : `https://medium.com/@AymericPM/pour-y-voir-plus-clair-sur-les-notions-de-transparence-et-dexplicabilit%C3%A9-en-ia-c0db2e96ae62`.

[29] Winston Maxwell. *Le contrôle humain des systèmes algorithmiques - un regard critique sur l'exigence d'un "Humain dans la boucle"*. HDR, Université Paris 1 Panthéon- Sorbonne, 2022.

[30] Winston Maxwell, Valérie Beaudouin, Isabelle Bloch, David Bounie, Stéphan Clémençon, Florence d'Alché Buc, James Eagan, Pavlo Mozharovskyi, and Jayneel Parekh. Identifying the 'Right' Level of Explanation in a Given Situation. *SSRN Electronic Journal*, 2020.

[31] Assemblée Nationale. Projet de loi relatif aux jeux Olympiques et Paralympiques de 2024 et portant diverses autres dispositions, April 2023.

[32] Angélique Négroni. De redoutables radars urbains expérimentés, LE FIGARO, July 2021. URL : `https://www.lefigaro.fr/actualite-france/de-redoutables-radars-urbains-experimentes-20210728`.

[33] Independent Hight-Level Expert Group on Artificial Intelligence set up by the European Commission (HLEG). Ethics guidelines for trustworthy AI. Technical report, European Commission, 2019.

[34] Jayneel Parekh, Pavlo Mozharovskyi, and Florence d'Alché Buc. A Framework to Learn with Interpretation. *arXiv :2010.09345*, 2021.

[35] Xavier Pin. V° Circulation routière - Fasc. 102 : Circulation routière - Constatation des infractions routières. In *JurisClasseur Lois pénales spéciales*. 2019.

[36] Gabrielle Ras, Marcel van Gerven, and Pim Haselager. Explanation Methods in Deep Learning : Users, Values, Concerns and Challenges. *arXiv :1803.07517*, 2018.

[37] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once : Unified, Real-Time Object Detection, 2016. arXiv :1506.02640.

[38] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You ?" : Explaining the Predictions of Any Classifier. *arXiv :1602.04938*, 2016.

[39] Cynthia Rudin. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *arXiv :1811.10154*, 2019.

[40] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM : Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[41] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity : The all convolutional net. *arXiv :1412.6806*, 2015.

[42] Catherine Tessier. Ethique et IA : analyse et discussion. In *Conférence Nationale en Intelligence Artificielle (CNIA)*, 2021.

[43] Richard Tomsett, Dave Braines, Dan Harborne, Alun Preece, and Supriyo Chakraborty. Interpretable to Whom ? A Role-based Model for Analyzing Interpretable Machine Learning Systems. *arXiv :1806.07552*, 2018.

[44] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–15, New York, NY, USA, 2019. Association for Computing Machinery.

[45] Xinru Wang and Ming Yin. Are Explanations Helpful ? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces*, pages 318–328, New York, NY, USA, 2021. Association for Computing Machinery.

[46] Matthew D. Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, Lecture Notes in Computer Science, pages 818–833. Springer International Publishing, 2014.