The Positive Way

WAVESTONE

INTELLIGENCE ARTIFICIELLE ET CYBERSÉCURITÉ

PROTÉGER DÈS MAINTENANT LE MONDE DE DEMAIN

AUTEURS

CAROLE MEZIAT
carole.meziat@wavestone.com

LAURENT GUILLE
laurent.guille@wavestone.com

Cette publication a été réalisée avec la contribution d'Erwan NICOLAS, consultant en cybersécurité et confiance numérique.

Le constat est alarmant : 58% des entreprises du CAC 40 mentionnent lancer des projets d'IA dans leur dernier rapport annuel et 2 seulement font le lien avec la cybersécurité [1]. Comment empêcher de manière pragmatique et concrète une future catastrophe ?

L'histoire le montre, des attaques sur des systèmes d'intelligence artificielle ont déjà eu lieu. En 2017, à la sortie de l'iPhone X, Apple se vantait d'avoir créé un système de reconnaissance faciale extrêmement robuste. Une semaine plus tard, et pour un coût ne dépassant pas les 150$, l'entreprise de cybersécurité vietnamienne *Bkav* avait réussi à créer un masque capable de duper l'application.

L'Intelligence Artificielle (IA) est en train de révolutionner notre quotidien : voiture autonome, biométrie comportementale, médecine prédictive, *chatbot* intelligent, suggestion de contenu… De nouveaux usages apparaissent chaque jour mais le sujet de la gestion des risques de cybersécurité apportés par cette nouvelle technologie reste rarement abordé.

Quels sont les risques spécifiques associés à l'IA ? Quelles sont les bonnes questions avant de se lancer ? Quelles solutions de sécurisation pour accompagner ces innovations ? Comment les choisir et les mettre en place ?

[1] Wavestone - Financial communication cybermaturity index – 2019 edition CAC 40 findings.

52%

des entreprises françaises **utilisent des solutions IA ou en ont le projet**

1/3

d'entre elles y allouent déjà un **budget annuel supérieur à un million d'euros**

71%

des acteurs **anticipent un budget IA en forte croissance**

Étude réalisée en 2018 par Tata Consultancy Services auprès de 900 entreprises françaises.

INTELLIGENCE ARTIFICIELLE : AU-DELÀ DU *BUZZWORD*, QUEL CONCEPT ?

L'Intelligence Artificielle permet de **reproduire l'intelligence humaine.** Cela va du programme d'analyse médicale détectant une tumeur jusqu'à celui capable de conduire votre voiture en toute autonomie.

Parmi ces applications, celles dont les règles ont été fixées à l'avance par des experts se distinguent de celles dotées d'une **faculté à adapter leur comportement** en fonction de la situation. Pour ce second cas, les termes **d'apprentissage automatique, ou de *Machine Learning (ML)*** sont utilisés. Ces systèmes se basent sur un grand nombre de données, manipulées à l'aide d'ordinateurs de plus en plus puissants, et les analysent pour reconnaître automatiquement des *patterns* (motifs) servant de base à des décisions. Différentes méthodes d'apprentissage existent, non exhaustives : apprentissage supervisé, non-supervisé ou apprentissage par renforcement.

Ces systèmes, avec leur mécanisme d'apprentissage et leur faculté à adapter leurs seuils de décision de manière évolutive, **apportent une différence fondamentale avec les systèmes historiques.** Aussi, l'amalgame est souvent fait entre Intelligence Artificielle et *Machine Learning*, y compris pour évoquer le sujet de la cybersécurité. Ce focus ne fait pas exception à la règle et, derrière les mots « Intelligence Artificielle », **il adresse plus particulièrement la gestion des risques de cybersécurité liés à l'utilisation du *Machine Learning.***

DE NOUVEAUX DÉFIS POUR LES ÉQUIPES CYBERSÉCURITÉ

Pourquoi attaquer l'Intelligence Artificielle ?

La technologie est en plein essor et de plus en plus d'applications utilisant du *Machine Learning* s'inscrivent dans nos usages et manipulent nos données, voire agissent physiquement dans notre quotidien. L'étude des vulnérabilités de ces systèmes peut représenter un **investissement rentable pour les attaquants capables de revendre des données ou de monnayer certaines prises de décisions.** De plus, compromettre le module de reconnaissance faciale de l'iPhone ou faire dévier une voiture autonome de sa route est susceptible d'attirer un attaquant **avide de défis technologiques ou de renommée personnelle.**

Les attaquants vont principalement chercher à :

- **Détourner le fonctionnement de l'application d'IA,** en provoquant volontairement une décision erronée de l'application avec un jeu de données choisi. Le contournement d'un système de reconnaissance faciale permettant d'obtenir un accès logique ou physique non légitime et réaliser un vol en est un exemple.
- **Saboter le fonctionnement de l'IA,** en empêchant ou en perturbant le fonctionnement de l'application. L'attaquant vise une dégradation de l'image de marque ou un ralentissement des activités de l'entreprise ciblée. L'attaque du *chatbot* Microsoft Tay en 2016, présentée dans la suite du focus, est un exemple emblématique de sabotage.
- **Comprendre et « rétroconcevoir » le modèle** en étudiant son comportement. Le travail de traitement des données et de modélisation est souvent long et coûteux pour les entreprises et le résultat à forte valeur ajoutée. « Voler » puis revendre un modèle peut être très lucratif et des acheteurs seront présents pour gagner du temps dans la perpétuelle course à l'innovation numérique.
- **Dérober les données utilisées par l'application,** en les requêtant directement ou en recoupant les résultats fournis par l'application et en tentant d'en tirer des conclusions, voire voler les bases de données auxquelles l'IA a accès.

Comment attaquer l'Intelligence Artificielle ?

Les attaques touchant spécifiquement les applications basées sur du *Machine Learning* peuvent être rassemblées en trois catégories.

Empoisonnement, ou comment déplacer le centre de gravité de l'IA

Cette technique est celle qui trouve le moins d'équivalent avec les méthodes traditionnelles car elle **cible spécifiquement la phase d'apprentissage automatique.** Avec l'empoisonnement, un attaquant cherche à modifier le comportement de l'IA dans un sens choisi en influençant les données utilisées pour l'apprentissage.

L'attaquant cherche principalement à détourner le fonctionnement de l'application à son avantage ou à saboter l'application.

C'est par exemple ce qui est arrivé en 2016 à Microsoft Tay, *chatbot* construit par Microsoft pour étudier les interactions qu'avaient les jeunes américains sur les réseaux sociaux et en particulier Twitter. Microsoft Tay a été inondé pendant toute une nuit de *tweets* abusifs par un groupe d'utilisateurs malveillants du forum 4chan, ce qui a, en moins de 10 heures, fait basculer le *chatbot* du comportement d'un adolescent « normal » à celui d'un extrémiste notoire.

Toutes les applications utilisant de l'apprentissage automatique sont concernées par ce type d'attaques. Néanmoins, ces techniques sont **particulièrement redoutables lorsque la donnée utilisée pour l'apprentissage est peu maîtrisée :** donnée publique ou externe, fréquence d'apprentissage élevée…

Inférence, ou comment faire parler l'IA

Avec l'inférence, un attaquant expérimente, teste successivement différentes requêtes sur l'application et étudie l'évolution de son comportement.

L'attaquant cherche ici soit la récupération des données utilisées par l'IA (en apprentissage ou en production), soit le vol du modèle (ou de certains de ses paramètres).

Cela fait par exemple partie des **techniques les plus utilisées pour détourner le fonctionnement des solutions de sécurité**. Les attaquants souscrivent à la solution, envoient de multiples requêtes, récupèrent les sorties associées et utilisent toutes ces données pour entraîner un modèle identique à la solution de sécurité étudiée. Ils peuvent ensuite initier des attaques adverses plus facilement, étant propriétaires de l'algorithme et donc ayant accès à toutes les informations conduisant à la décision de la solution. Ils utiliseront ensuite les exemples adverses générés pour propager des attaques qui ne seront pas détectées.

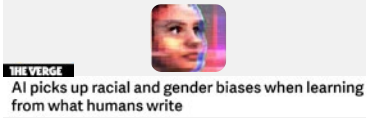
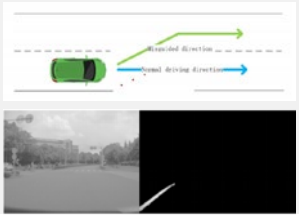
L'attaquant utilise toutes les informations fournies en sortie de l'application pour mener à bien son attaque. Par conséquent, ces techniques sont d'autant plus efficaces que **l'application est exposée à un grand nombre d'utilisations et que le résultat de la prise de décision est détaillé.** Par exemple, une application de reconnaissance d'image renvoyant un résultat avec son niveau de fiabilité « cette image est un chien avec un niveau de confiance de 90% et une autruche avec un niveau de confiance de 10% » sera beaucoup plus vulnérable à ce type d'attaques qu'une application qui renvoie uniquement la réponse la plus probable « cette image est un chien ».

L'évasion, ou comment duper l'IA

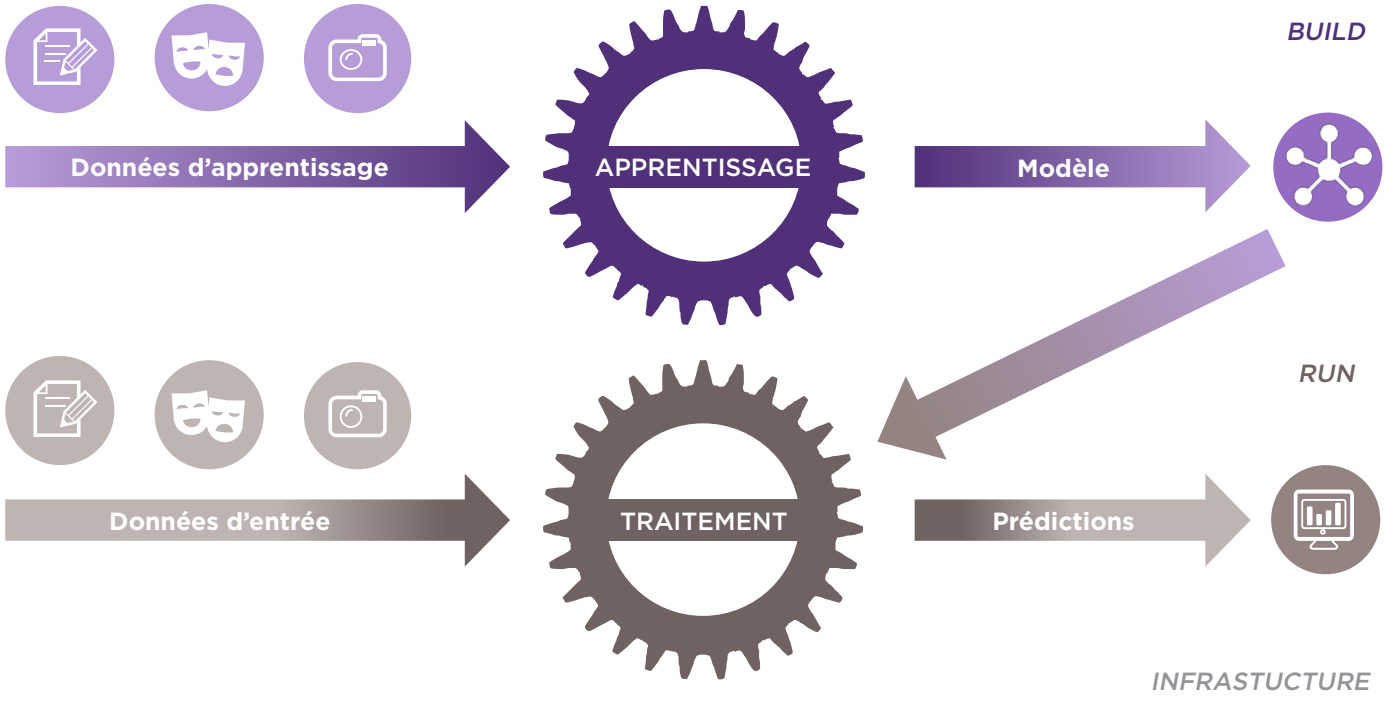
Avec l'évasion, un attaquant joue sur les données d'entrée de l'application afin d'obtenir une décision différente de celle normalement attendue par l'application. Il cherche à créer **l'équivalent d'une illusion d'optique pour l'algorithme, appelé exemple contradictoire** (*adversarial example*), en introduisant un « bruit » judicieusement calculé pour rester discret et ne pas être détecté.

L'attaquant cherche ainsi à détourner le comportement de l'application à son avantage et cible l'application en production, une fois l'apprentissage terminé.

SYNTHÈSE DES PRINCIPAUX MOYENS D'ATTAQUE SPÉCIFIQUES AU *MACHINE LEARNING*

Catégorie d'attaque	EMPOISONNEMENT	INFERENCE	ÉVASION
Exemple d'attaque	*Changer le centre de gravité de l'IA*  THE VERGE AI picks up racial and gender biases when learning from what humans write	*Extraction d'information depuis l'IA*  iPhone X's Face ID supposedly got hacked. We have questions	*« Illusions d'optique » pour l'IA* 
Motivations principales	**Détournement du fonctionnement** **Sabotage de l'application**	**Compréhension et rétroconception du modèle** **Vol de données**	**Détournement du fonctionnement**
Phase ciblée	**Apprentissage** *(build)*	**Apprentissage** *(build)*, **Traitement** *(run)*	**Traitement** *(run)*
Facteurs aggravants	**Fréquence d'apprentissage** (continue…) **Données non maîtrisées** (publiques, non authentifiées…)	**Verbosité des sorties** (fourniture du niveau de fiabilité…) **Exposition de l'application et des données d'apprentissage**	**Complexité des entrants** (images, son…) **Exposition de l'application** (lieux publics, Internet sans authentification…)

BRIQUES DE FONCTIONNEMENT D'UNE APPLICATION BASÉE SUR DU *MACHINE LEARNING*



La compromission du logiciel de pilotage automatique de la voiture autonome Tesla par un groupe de chercheur chinois travaillant pour le *Keen Security Labs* est une illustration de ce type d'attaques. Les chercheurs ont réussi à faire dévier de sa ligne une Tesla S, grâce à de simples autocollants collés sur les marquages des voies utilisées par la voiture. La reconnaissance du marquage routier et des éléments d'environnement, appris grâce à du *Machine Learning*, servent de base aux décisions d'orientation des voitures autonomes. Les attaquants ont ainsi réussi à fausser le marquage routier pour créer une situation ni prévue, ni apprise par la Tesla S.

L'exposition à ces techniques d'évasion est d'autant plus importante **que l'application est largement exposée (utilisation grand public, contrôle d'accès limité) et les données d'entrée complexes (image, son...).**

Cette troisième catégorie vient compléter le panel des grands types d'attaques propres à l'utilisation de l'Intelligence Artificielle. Le scénario ultime étant une prise de contrôle totale d'un système d'IA par un attaquant, et son utilisation comme rebond pour attaquer une autre cible.

SIX POINTS CLÉ POUR RÉUSSIR SON PROJET D'INTELLIGENCE ARTIFICIELLE EN TOUTE SÉCURITÉ

Nous l'avons vu, les solutions d'IA, en pleine expansion et très convoitées, sont loin d'être à l'abri des cyberattaques. Les moyens utilisés sont parfois d'un genre nouveau, obligeant à challenger, et dans certains cas repenser, les mécanismes de sécurité applicables aux systèmes existants. Voici les six points à ne pas rater pour encadrer les risques liés aux projets métier d'Intelligence Artificielle.

1\ Protéger les données à chaque étape du projet

La donnée est le fondement des projets de *Machine Learning* dans les entreprises. Ces projets nécessitent la manipulation d'un grand volume et d'une grande variété de données. Avant même de parler de protection contre les fuites d'informations, il est essentiel de garantir que l'usage souhaité est **conforme aux réglementations** en vigueur, et notamment à celles liées à la protection de la donnée (RGPD, Hébergeur de Données de Santé, PCI-DSS...). Cela passe par la définition la plus claire possible de la finalité du projet et des traitements de données associés.

Ceci doit être pensé non seulement pour l'usage cible de l'application, mais également **pour toute la phase de développement et de test de la solution de *Machine Learning*,** particulièrement gourmande en données. En effet, une des opportunités clé du *Machine Learning* est de pouvoir établir des corrélations parmi un grand nombre de données qu'un humain serait incapable d'analyser. Les modèles sont donc testés avec le plus de données possibles pour pouvoir déterminer le plus efficace. Une fois le modèle sélectionné et validé, il est possible de **réduire le nombre et la nature des données utilisées au strict nécessaire, pour le passage en production.** Il convient donc, pour tout projet de *Machine Learning,* de :

/ **Désensibiliser les données d'apprentissage** dès que cela est possible, et en premier lieu pour les données à caractère personnel. Des solutions de génération de jeux de données fictives (ou synthétiques), produisant des données désensibilisées mais conservant leurs valeurs statistiques, émergent sur le marché.

Ces solutions, telles que *Mostly AI, HAZY* ou *KIProtect*, offrent un potentiel important pour traiter ces enjeux de *data privacy.*

/ **Valider la possibilité d'utilisation des données sensibles résiduelles** en les protégeant de manière adéquate (droit d'accès, chiffrement…), notamment pour la phase de développement.

/ **Sensibiliser et responsabiliser les équipes au cœur du développement des modèles** à la manipulation des données sensibles, incluant non seulement les *data engineers* et *data scientists,* mais également les métiers impliqués.

Les décisions prises par les solutions de *Machine Learning* n'étant plus basées sur des règles explicitement définies par des humains, mais apprises automatiquement, elles peuvent parfois devenir difficiles, voire impossibles à expliquer. À ce stade, il est donc également primordial d'évaluer le **niveau d'interprétabilité** nécessaire des algorithmes. Ceci est notamment le cas pour les algorithmes prenant des décisions à partir de données personnelles dans le cadre du RGPD, qui exige que toute décision importante ou de nature juridique puisse être expliquée.

2\ Protéger la plateforme *Big Data*

Dans le cadre des projets de *Machine Learning,* cette étape prend une dimension particulière. Les données y sont nombreuses et très concentrées, donc particulièrement exposées au risque de vol ou de modification d'information. Et les modèles eux-mêmes et leurs jeux d'entraînement constituent un précieux secret industriel qui, dans un contexte concurrentiel important, sont particulièrement convoités : vol du modèle, de ses seuils de décision…

Les systèmes *Machine Learning* ayant naturellement un comportement évolutif, il est plus difficile de détecter des modifications subtiles de comportement et donc d'éventuelles modifications malveillantes. Il est par conséquent essentiel d'**appliquer les bonnes pratiques de protection du *Big Data***, quel que soit l'environnement utilisé : sécurité périmétrique et cloisonnement, gestion des habilitations et des accès (sur chaque brique), gestion des comptes à privilèges, durcissement, maintien en conditions de sécurité, chiffrement des supports, traçabilité, sécurité des fournisseurs et contractualisation…

En parallèle, le développement des projets de *Machine Learning* nécessite souvent l'utilisation de technologies spécifiques, non incluses dans les *frameworks* de développement classiques de l'entreprise, dont il est **nécessaire d'évaluer et de valider le niveau de sécurité** avant passage en production ou généralisation.

Ces briques supportant les premières initiatives de *Machine Learning* serviront souvent de base pour la création de services à plus grande échelle, et la mise en place de ces bonnes pratiques dès les premières initiatives **bénéficiera à l'ensemble des futurs projets.**

3\ Sécuriser le processus d'apprentissage

L'apprentissage automatique est à la fois l'étape clé sur laquelle repose toute l'efficacité et la pertinence de la solution, et la vraie nouveauté par rapport aux systèmes existants. On comprend vite qu'il puisse constituer une cible de choix pour les attaquants. Il convient donc de dédier une réflexion toute particulière à la protection de cette étape. Cette protection doit se faire à deux niveaux : au niveau des données d'entraînement et au niveau de la méthode d'apprentissage.

> Des projets d'un genre nouveau qui nécessitent d'être créatif sur les scénarios de risques à envisager et innovant sur les mesures à déployer

Les algorithmes sont entrainés sur un échantillon de données appelé jeu d'apprentissage. Ces données sont utilisées comme des exemples de base que les algorithmes développés doivent permettre de généraliser. La constitution du jeu d'apprentissage est une étape essentielle de tout projet de *Machine Learning*. Celui-ci doit être suffisamment important pour pouvoir être généralisé, suffisamment représentatif pour ne pas introduire de biais, et suffisamment lisible pour que l'extraction des données fournisse le plus de valeur possible au modèle. **Influer sur le jeu d'apprentissage permet d'influer sur le comportement de l'application de *Machine Learning.***

Plusieurs mesures peuvent être mises en place pour **assurer la fiabilité du jeu d'apprentissage utilisé** et lutter contre les attaques de type empoisonnement décrites plus haut.

/ Dans un premier temps, allonger la durée d'apprentissage pour élargir au maximum le jeu de données d'entraînement en utilisant de l'***Advanced Learning*** permet de diminuer l'impact de chaque donnée d'entrée sur le fonctionnement du modèle et donc de rendre la solution moins sensible aux altérations partielles du jeu d'apprentissage.

/ Ensuite, **différents contrôles peuvent être mis en place sur toute la phase d'apprentissage :** validation systématique par un expert métier du domaine concerné, mise en place de contrôles d'intégrité sur les jeux de données pour détecter les altérations à plusieurs moments de la phase d'apprentissage, définition de seuils d'utilisation de données issues d'une même source (localisation, individu, IP…) pour réduire les risques d'*oversampling*, définition de *blacklists* (mots clés, *patterns*) à retirer systématiquement du jeu d'apprentissage (ex. : vocabulaire injurieux pour un *chatbot*), détection

des évolutions de comportement du modèle au fil des réentraînements (détection d'évolutions brusques d'un entraînement à l'autre, comparaison des évolutions à fréquence ou à nombre de réentraînements réguliers…).

Il est clair que moins le jeu d'apprentissage est maîtrisé (données publiques ou issues d'un fournisseur externe) et stable (apprentissage ponctuel *vs* continu), plus son risque de compromission sera élevé, et donc plus les moyens mis en œuvre pour protéger le jeu d'apprentissage devront être importants. La clé sera de **trouver le meilleur compromis entre le besoin métier de quantité, variété et d'actualisation des données d'apprentissage et les besoins de sécurité et de maîtrise des données.**

En parallèle, il est également nécessaire de définir des bonnes pratiques d'apprentissage sécurisé, historiquement non existantes. La **sécurisation des méthodes d'apprentissage représente tout un nouveau champ de recherche en cybersécurité.** Il s'agit ici de définir quels sont les algorithmes à privilégier d'un point de vue sécurité, comment les utiliser, quelles options activer lors du développement des algorithmes… L'application de bonnes pratiques comme le RONI (*Reject On Negative Impact*), permettant de supprimer du jeu d'apprentissage les données ayant un impact négatif sur la précision du modèle, ou le *Bootstrap Aggregating*, permettant de stabiliser le modèle, sont des exemples de techniques permettant de construire des solutions plus robustes.

4\ Sécuriser l'application

Une bonne partie des tentatives d'attaque peut être contenues en **appliquant les bonnes pratiques de développement sécurisé déjà largement déployées dans les entreprises** (par exemple, les règles de *Open Web Application Security Project*, ou OWASP). Ceci est d'autant plus important que les profils *data scientists* sont issus de parcours plutôt teintés statistiques qu'informatique, et donc souvent très peu sensibilisés aux sujets de sécurité par rapport aux autres developpeurs plus classiques, et que les projets sont souvent menés directement par les métiers, indépendamment des équipes IT plus habituées à gérer la sécurité dans les projets.

Néanmoins, celles-ci ne suffisent pas à protéger contre tous les cas de fraudes liés à l'utilisation du *Machine Learning*. L'essentiel des mesures de sécurité spécifiques au *Machine Learning* se concentre sur trois aspects : maîtriser ses entrants, fiabiliser le traitement et contrôler ses sorties.

Maîtriser ses entrants

- **Protection de la chaîne d'acquisition des données :** il est primordial que personne ne puisse modifier les données entre le moment où elles sont acquises et le moment où elles sont fournies à l'algorithme pour effectuer les prédictions. Pour cela, la chaîne de traitement de la donnée doit être conçue pour garantir une protection de bout en bout, en limitant les canaux d'entrée offerts aux utilisateurs, en appliquant une gestion d'accès stricte, en chiffrant les flux associés…
- **Filtrage des données d'entrées :** cela est similaire au filtrage des données d'entrée requis pour les applications web classiques. Cela passe par une vérification de format (type de données, exhaustivité des informations entrées ou extraites…), ou de cohérence des données d'entrée (écart par rapport aux données anticipées, aux données historisées…), par la détection de bruit (signaux faibles) dans les données d'entrée, puis par le rejet ou le nettoyage des données avant qu'elles ne soient traitées par l'application. La détection de bruit (*Noise Prevention*) est par exemple particulièrement utilisée dans le cadre d'applications de reconnaissance d'image pour lutter contre les attaques adverses.
- **Détection et blocage des comportements utilisateurs suspects :** cela consiste à mettre en place des mécanismes permettant de détecter les tentatives d'attaques par inférence, par exemple en détectant une affluence de requêtes similaires ou de même origine, en détectant des anomalies successives sur les formats d'entrées… La détection de comportements suspects utilisateurs (UEBA, *User and Entity Behavior Analytics*) est d'ailleurs un domaine d'application majeur d'utilisation du *Machine Learning* pour améliorer la cybersécurité.

Fiabiliser le traitement

- ***Adversarial Training* :** cette méthode s'utilise en amont de la mise en production, pendant la phase d'apprentissage, et consiste à apprendre au modèle des exemples d'attaques possibles et à y associer une prise de décision. Cette technique est très utilisée dans le cadre de la reconnais-

BIAIS D'APPRENTISSAGE, ÉTHIQUE ET RISQUES RÉPUTATIONNELS

L'influence des jeux de données et les risques de biais d'apprentissage associés dépassent largement la problématique cybersécurité et soulèvent des questions d'éthique. Sans vigilance, des biais d'apprentissage naturels peuvent s'introduire et avoir des conséquences réputationnelles extrêmement néfastes. En témoigne l'exemple des algorithmes de reconnaissance d'image de Google Photos classifiant des photos de personnes noires en « gorilles » en 2015.

La Commission Européenne s'intéresse de près à ces sujets et a récemment publié des recommandations sur l'éthique de l'IA (*Ethics Guidelines for Trustworthy Artificial Intelligence*, que l'on peut retrouver sur le site de la Commission Européenne).

sance d'image, en utilisant des *Generative Adversarial Networks* (GAN) pour générer automatiquement des images contradictoires incluant du bruit, et rendre les modèles plus robustes aux attaques par évasion.

- ***Randomization*** **:** cette méthode répond au même objectif que l'adversarial training, et consiste à ajouter un bruit aléatoire à chaque donnée. L'ajout de ce bruit aléatoire avant chaque traitement rend plus difficile pour un attaquant de prédire la perturbation à ajouter à une entrée pour arriver à ses fins. L'algorithme est entraîné sur des données auxquelles ce bruit aléatoire a également été ajouté, l'intensité de ce bruit étant optimisée pour obtenir le meilleur compromis entre l'exactitude de l'algorithme et sa robustesse face aux attaques par évasion. L'efficacité de la *randomization* est mathématiquement prouvée, contrairement à d'autres techniques de fiabilisation des algorithmes qui offrent uniquement des garanties empiriques. Ses résultats face aux attaques par évasion sont comparables à l'*adversarial training*, la *randomization* étant beaucoup moins coûteuse en temps de calcul.

- ***Defensive Distillation*** **:** cette technique consiste à utiliser deux modèles successifs afin de minimiser les erreurs de décision. Le premier modèle est entraîné pour maximiser la précision de l'algorithme (100% de probabilité que l'image soit un chat plutôt qu'un chien). Le second modèle est entraîné avec les sorties du premier algorithme qui présentent une incertitude faible (par exemple celles qui répondent à 95% de probabilité que l'image soit un chat plutôt qu'un chien). Le second modèle, « distillé », rend la tâche de deviner le mode de fonctionnement de l'algorithme plus complexe pour les attaquants et le rend ainsi plus robuste. Cette technique met un obstacle de plus sur le chemin de l'attaquant, qui pourra néanmoins, moyennant plus de temps et de ressources, étudier, comprendre et compromettre le fonctionnement global de la solution.

- ***Ensemble Learning*** **:** pour fiabiliser les prédictions, plusieurs modèles peuvent être utilisés simultanément, chacun étant basé sur des algorithmes et des fonctionnements différents, afin de combiner et / ou comparer leurs résultats avant la prise de décision. Cette technique, coûteuse en temps et en ressources, part de l'hypothèse que tous les algorithmes ne sont pas sensibles aux mêmes variations des données d'entrée, et donc qu'une attaque fonctionnant sur l'un des algorithmes pourrait ne pas fonctionner sur un autre. Cette hypothèse est néanmoins aujourd'hui contestée, des recherches montrant que des modèles utilisant des méthodes différentes, entrainés sur les mêmes données, sont sensibles aux mêmes attaques par évasion. Pour les cas d'usage les plus sensibles, elle peut néanmoins rajouter une difficulté pour les attaquants, rendant les décisions du modèle moins lisibles.

Contrôler ses sorties

- ***Gradient Masking*** : cela vise à réduire le risque de *reverse engineering* du modèle en limitant la verbosité de l'application de *Machine Learning*, et notamment concernant les scores de décision utilisés par l'algorithme. Cela réduit le risque de *reverse engineering* de l'algorithme via un jeu itératif sur les paramètres d'entrée.

- ***Protection de la chaîne de sortie :*** de la même manière que pour la chaîne d'acquisition des données, l'ensemble des prédictions effectuées par l'algorithme doivent être protégées contre les tentatives d'accès. Seul le résultat de la décision doit être accessible. Les mêmes mécanismes vont être utilisés ici à savoir : chiffrement, contrôle d'accès…

- **Détection des sorties suspectes :** comme les systèmes basés sur le *Machine Learning* sont évolutifs, les décisions peuvent évoluer malgré un contexte a priori similaire (en fonction du temps, d'un individu…). Afin de limiter les détournements, des vérifications peuvent être mises en place, par exemple, en comparant un résultat par rapport à des indicateurs de référence et en levant une alerte en cas de doute. Par exemple, avant d'initier un virement bancaire à un destinataire spécifique, une vérification que le montant n'est pas plus de 10 fois supérieur aux montants moyens de l'année passée vers ce même destinataire peut être effectuée. Cela permet à la fois de détecter les erreurs et les potentielles malveillances. La manière de traiter les anomalies doit ensuite être définie au cas par cas : arrêt du traitement, demande de réauthentification, alertes auprès des équipes…

- **Modération et *blacklist*** des données de sortie : pour un contrôle des sorties du modèle, une liste de réponses interdites peut également être mise en place ainsi qu'une modération manuelle *a posteriori* des résultats. Peut également être envisagée l'application de contraintes automatiques sur les sorties comme par exemple en forçant les valeurs à rester dans des plages autorisées.

5\ Définir sa stratégie de gestion de risques et de résilience

Comme évoqué en introduction, le terme Intelligence Artificielle regroupe une très grande variété d'applications ; **toutes n'ont pas le niveau de sensibilité de la voiture autonome.** Prenons l'exemple d'un *chatbot* intelligent : le niveau de risque dans le cas d'un *chatbot* de conseil passif, destiné à fournir des conseils personnalisés en réponse à des questions comme « Quel est le montant de remboursement de mes lunettes ? », sera moindre que celui d'un *chatbot* transactionnel, capable d'effectuer des opérations telle la création en ligne d'une carte de tiers payant. Il est par conséquent ici aussi essentiel de réaliser une **analyse de risques afin de savoir où prioriser les efforts.** Ces analyses de risques doivent prendre en compte les risques spécifiques liés à l'utilisation du *Machine Learning* (présentés dans la première partie de ce focus).

Afin d'industrialiser la démarche, **des *guidelines* de sécurité par typologie de projets d'IA** peuvent être définies. Ces *guidelines* peuvent par exemple être regroupées par type de données d'entrée (image, parole, texte, données structurées…),

MESURES DE SÉCURITÉ À INTÉGRER À CHAQUE ÉTAPE DU PROJET D'INTELLIGENCE ARTIFICIELLE

par fréquence d'apprentissage (continue, ponctuelle, régulière) ou par niveau d'exposition de la solution (publique, interne...). Il sera parfois nécessaire de dédier du temps et de recourir à de l'expertise afin de traiter ces problématiques d'un genre nouveau.

Aucun système n'étant à l'abri d'une attaque, il est clé d'anticiper les attaques et de définir **une stratégie de résilience adaptée aux spécificités du *Machine Learning* :**

- Réfléchir aux mécanismes de haute disponibilité et de sauvegarde de l'application. En particulier, le sabotage de l'application peut se faire au niveau de l'apprentissage, en compromettant les données utilisées et en rendant le modèle caduc. Il est essentiel de **conserver une copie des états passés du modèle et d'anticiper les éventuels besoins de retour arrière.**
- S'interroger sur les **traces à conserver en cas de besoin d'investigation.** Le *Machine Learning* a parfois un effet *black box*, qui limite la capacité à comprendre de quelle manière le modèle arrive à une conclusion, ce qui représente un obstacle à l'investigation lors d'un incident. Les bonnes pratiques de traçabilité traditionnelles ne prennent pas en compte cette complexité liée au *Machine Learning* ; il est donc clé de définir des exigences de traçabilité spécifiques au *Machine Learning*, permettant notamment de **garder une trace des paramètres qui conduisent aux décisions prises par les algorithmes.** De telles pratiques faciliteront par exemple la coopération avec la justice en cas de besoin.

6\ Bien réfléchir avant d'externaliser

Dans certains contextes ou pour certains cas d'usage, les solutions d'IA sont parfois délivrées par des fournisseurs externes. Faire appel à des solutions externes comporte toujours son lot de risques, et au-delà de la nécessité de **challenger les éditeurs sur les points évoqués précédemment,** certains points de vigilance supplémentaires spécifiques à l'usage du *Machine Learning* doivent être encadrés pendant la phase de contractualisation.

Le premier point d'attention touche à la **propriété intellectuelle.** Et la question de la propriété se pose à la fois pour les données mais également sur le modèle entraîné. Qui est propriétaire du modèle entraîné ? Qui en est responsable ? À la fin du contrat, qui le récupère ? Sous quel format ? Ces questions sont clé avant la prise de décision d'externalisation. Le marché actuel de l'IA voit fleurir de multiples solutions jeunes, innovantes, très spécialisées et ainsi particulièrement exposées à une probable consolidation du marché dans les années à venir. Que se passerait-il si votre concurrent direct décidait d'investir dans une de ces technologies innovantes et rachetait le fournisseur avec lequel vous travaillez depuis plusieurs mois ? Ces risques sont à prendre en compte pendant les phases de contractualisation afin de ne pas se retrouver devant le fait accompli.

Un second point de vigilance a trait à l'exposition à l'inférence évoquée plus haut. Un des intérêts de faire appel à une solution externe est la possibilité d'obtenir un modèle entraîné sur un plus grand nombre et une plus grande variété de données, le fournisseur pouvant faire bénéficier la solution d'un entraînement sur les données de plusieurs de ses clients. Ceci rend ainsi la solution plus performante qu'en l'entraînant uniquement sur les données d'un seul client. Mais ce point invite à **s'interroger sur le cloisonnement de l'application entre les différents clients.** Si cette question se pose dans tout contexte de mutualisation, avec le *Machine Learning*, il ne s'agit pas simplement de vérifier que le fournisseur applique bien les bonnes pratiques de cloisonnement sur ses infrastructures et applications, **il s'agit également de déterminer si**

mutualiser l'entraînement des modèles peut conduire à la divulgation de données ou informations confidentielles de clients individuels. La question peut par exemple se poser à l'ajout d'un nouveau client dans le périmètre d'apprentissage, les nouvelles données entrées peuvent faire varier les décisions prises par la solution dans un sens qui permet de déduire certaines informations sur le client en question. Cette mutualisation peut également avoir d'autres conséquences, comme conduire à une divergence de comportement du modèle à la suite de son entraînement sur un nouveau jeu de données à l'acquisition d'un nouveau client. Il s'agit alors d'être particulièrement attentif à ce type de problématiques lors du choix de la solution et de choisir en conscience le meilleur **compromis entre les bénéfices métier d'un entraînement mutualisé et les risques de sécurité et sur la vie privée associés.** Il peut également être intéressant d'intégrer des clauses spécifiques permettant de limiter les risques, par exemple des clauses de maximum de pondération relative des clients les uns par rapport aux autres dans l'apprentissage effectué, et de mettre en place des indicateurs de suivi de la performance de la solution de *Machine Learning* pour vérifier l'amélioration du modèle dans le temps.

Enfin, il est également essentiel d'anticiper avant la souscription **les besoins de réversibilité** en fin de contrat avec le fournisseur. Au-delà des problématiques de propriété intellectuelle déjà évoquées plus tôt, les besoins de récupération ou de suppression des données d'entrainement ou des règles apprises doivent être matérialisés dans le contrat. Et les questions peuvent être plus subtiles que dans le cas de solutions classiques : dois-je imposer le réentrainement du modèle du fournisseur sans mes données après mon départ dans le cas d'un entrainement mutualisé ? Vais-je vouloir récupérer les règles issues du moteur d'analyse du fournisseur en fin de contrat ? Dans ce cas, quelles sont les contraintes à respecter par le fournisseur pour que leur exploitation soit compatible avec mes infrastructures ? Saurai-je réunir les compétences techniques en interne pour les exploiter ou dois-je prévoir un transfert de compétences *Machine Learning* dans le contrat ?

DES STANDARDS DE SÉCURISATION DU MARCHÉ EN COURS DE DÉFINITION

Les usages de l'IA en étant à leur début, leur maturité sécurité pâtit aujourd'hui de la jeunesse de la technologie. De nombreux mécanismes d'attaque et de défense sont d'ailleurs encore **au stade théorique de recherche.**

De multiples initiatives commencent à voir le jour pour **définir les standards qui régiront la sécurité des applications d'IA de demain.** Ces *Think tanks* et Groupes de Travail, parmi lesquels l'*AI Security Alliance* (US), l'*Information Commissioner's Office AI auditing framework* (UK) et le *Centre for European Policy Studies* (UE), sont à suivre de près dans les mois qui viennent afin d'alimenter au fur et à mesure ces *guidelines* de sécurisation de l'IA de mesures concrètes.

UN PRÉREQUIS : MOBILISER LES MÉTIERS SUR LA PROTECTION DE CES NOUVEAUX SYSTÈMES D'INTELLIGENCE ARTIFICIELLE

Une grande partie des initiatives actuelles d'Intelligence Artificielle dans les entreprises en est encore à l'état de recherche ou d'expérimentation. L'objectif est dans un premier temps de démontrer la valeur que l'IA peut apporter aux différents métiers de l'entreprise. Ces *proofs of value* (POV) se déroulent souvent sur des cas d'usages précis basés sur les données existantes de l'entreprise. Les **plannings sont courts, de manière à démontrer en quelques semaines seulement un retour sur investissement sur la base d'indicateurs concrets.** La sécurité reste alors rarement intégrée dans ces phases d'expérimentation. Pourtant, des données sensibles y sont souvent déjà utilisées et le planning de déploiement en production qui suivra le succès du POV sera généralement court et ne permettra que difficilement l'intégration des mesures de sécurité. Par ailleurs, la généralisation à d'autres périmètres ou l'extension des cas d'usages interviendra souvent rapidement.

Dans ce contexte, anticiper les besoins d'intégration des mesures de sécurité en amont est clé pour suivre le rythme des demandes au moment du passage à l'échelle et garantir leur prise en compte par les équipes cybersécurité.

Il peut être pertinent de prendre les devants en **publiant un *position paper* ou une note globale de sécurité sur le sujet.** Ceci permettra de sensibiliser les décideurs et les métiers aux enjeux de sécurité liés aux projets d'Intelligence Artificielle, avant même le démarrage des initiatives, et d'inculquer les bons réflexes à l'ensemble de l'écosystème dans l'entreprise.

En parallèle, **recenser et catégoriser les expérimentations et projets d'IA en cours** et à venir permet de commencer à définir les mesures de sécurité à intégrer en priorisant les besoins réels de l'entreprise : quelle est la nature des données manipulées (sensibles, personnelles, santé…) ? Les solutions sont-elles plutôt développées en interne ou plutôt souscrites auprès de fournisseurs spécialisés ? Quels types de données sont utilisés (texte, image, son…) ? Suivant les réponses, les mesures de sécurité à respecter pourront être très différentes.

Puis, si cet état des lieux révèle que de nombreux projets sensibles sont menés ou bien qu'une nette augmentation est anticipée dans les années à venir, il sera nécessaire de pouvoir accompagner ces projets plus finement. Avoir des **profils cybersécurité comprenant la *datascience* et capables de définir des mesures concrètes de sécurité** (*logs*, sauvegardes, validation des *frameworks* utilisés…) deviendra rapidement indispensable.

LA NOUVELLE MENACE À ANTICIPER : LE *DEEPFAKE*

En mars 2019, dupée par une voix artificielle imitant son PDG, une entreprise se fait dérober 220 000 euros. L'information, révélée à l'été 2019, fait le *buzz*. Le PDG de cette entreprise du secteur de l'énergie a reçu un appel du dirigeant de sa maison mère allemande lui demandant d'effectuer un virement de 220 000 euros sur le compte d'un fournisseur hongrois. Le virement a été fait. Son interlocuteur aurait en réalité été une **voix synthétique** imitant celle du dirigeant, créée par un **logiciel de génération de voix basé sur l'Intelligence Artificielle**.

Même si cette histoire reste à confirmer, des doutes persistant sur le fait que la voix ait vraiment été générée par un système d'Intelligence Artificielle, ce type d'attaque est aujourd'hui plus que possible vue l'avancée des systèmes d'usurpation de voix, voire de vidéos, les **célèbres « *Deepfake* »**. Et ces attaques vont être difficiles à gérer. En effet, comment se douter de quelque chose face à la voix familière de son patron ?

QU'EST-CE QUE LE *DEEPFAKE* ?

Le *Deepfake* désigne la modification d'images, audios ou vidéos via de l'Intelligence Artificielle (et notamment du « *Deep Learning* ») pour présenter une vision falsifiée (« *fake* ») de la réalité.

Un fichier *Deepfake* est créé en utilisant deux systèmes concurrents d'Intelligence Artificielle : l'un s'appelle le **générateur** et l'autre le **discriminateur**. Le générateur crée un faux fichier (image, audio, vidéo...) et demande ensuite au discriminateur de déterminer si le fichier est réel ou faux. Ensemble, le générateur et le discriminateur forment ce qu'on appelle un ***Generative Adversarial Network*** (GAN).

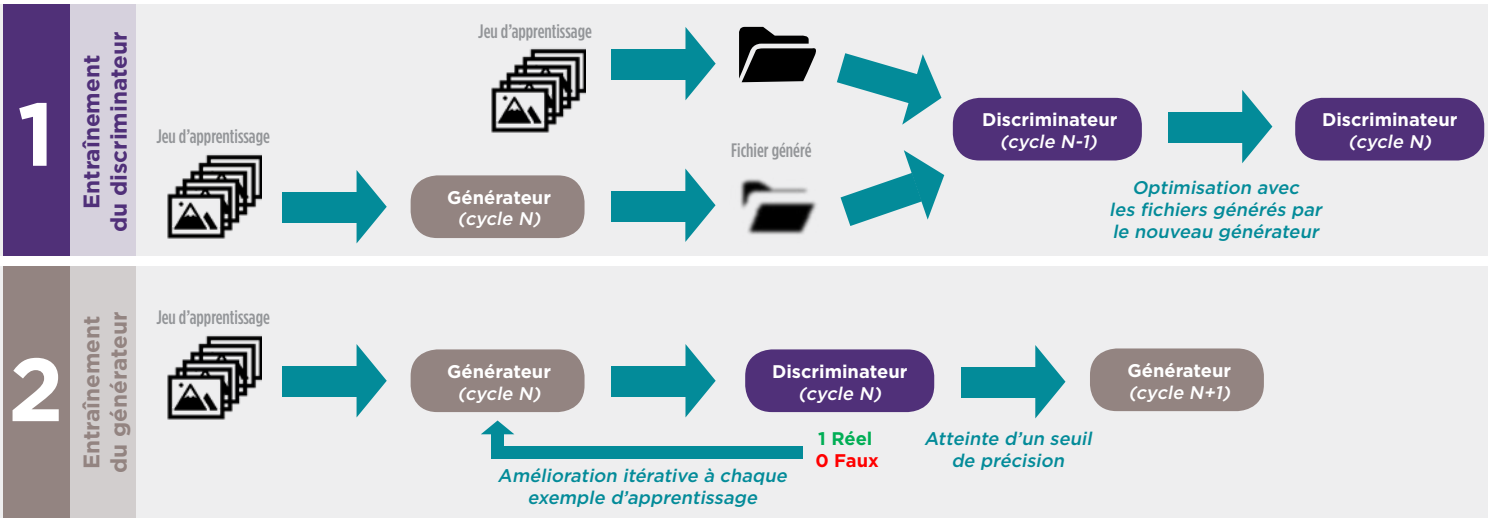
Un jeu de données d'apprentissage est fourni au générateur pour initialiser le processus. Ensuite, la méthode fonctionne par cycles de deux étapes :

1. **Entraînement du discriminateur** : Le discriminateur est entraîné à différencier les vrais fichiers des faux fichiers créés par le générateur.

2. **Entraînement du générateur** : Le générateur crée des fichiers, qui sont évalués par le discriminateur. Ceci permet au générateur de s'améliorer en produisant des fichiers de plus en plus crédibles (jugés comme réels) pour le discriminateur.

Une fois que le générateur a suffisamment progressé dans la crédibilité des fichiers créés, le discriminateur est à nouveau entraîné à différencier de vrais fichiers des faux fichiers créés par le générateur (nouvelle étape « 1 »). Le discriminateur ré entraîné est maintenant utilisé pour faire évoluer le générateur dans le cadre d'une nouvelle étape « 2 ». Ce cycle est **répété autant de fois que nécessaire** pour atteindre le niveau de précision souhaité.

FONCTIONNEMENT D'UN *GENERATIVE ADVERSARIAL NETWORK*



Il existe différentes formes de *Deepfake,* parmi lesquelles :

- Les ***Deepfake* audios**, qui imitent la voix d'une personne ciblée à partir d'échantillons de sa voix. Ils permettent de lui faire prononcer un texte donné en entrée.
- Le ***face-swapping***, qui remplace dans une vidéo le visage d'une personne par celui d'une personne ciblée à partir de sa photo.
- Le ***Deepfake lip-synching***, qui, dans une vidéo, adapte les mouvements du visage d'une personne ciblée à partir d'un fichier audio d'une autre personne. Il permet de lui faire prononcer le discours contenu dans le fichier audio en question sans qu'elle ne l'ait prononcé.
- Le ***Deepfake puppetry***, qui génère une vidéo d'une personne ciblée à partir d'une vidéo d'acteur fournie en entrée. Il est ainsi possible de créer une vidéo dans laquelle la cible reproduit un discours joué par l'acteur. Un exemple célèbre de cette technique est une vidéo de Barack Obama énonçant un discours simulé par Jordan Peele, dans laquelle la gestuelle de l'ancien président est reproduite de manière extrêmement réaliste.

DES TECHNIQUES À LA PORTÉE DE TOUS

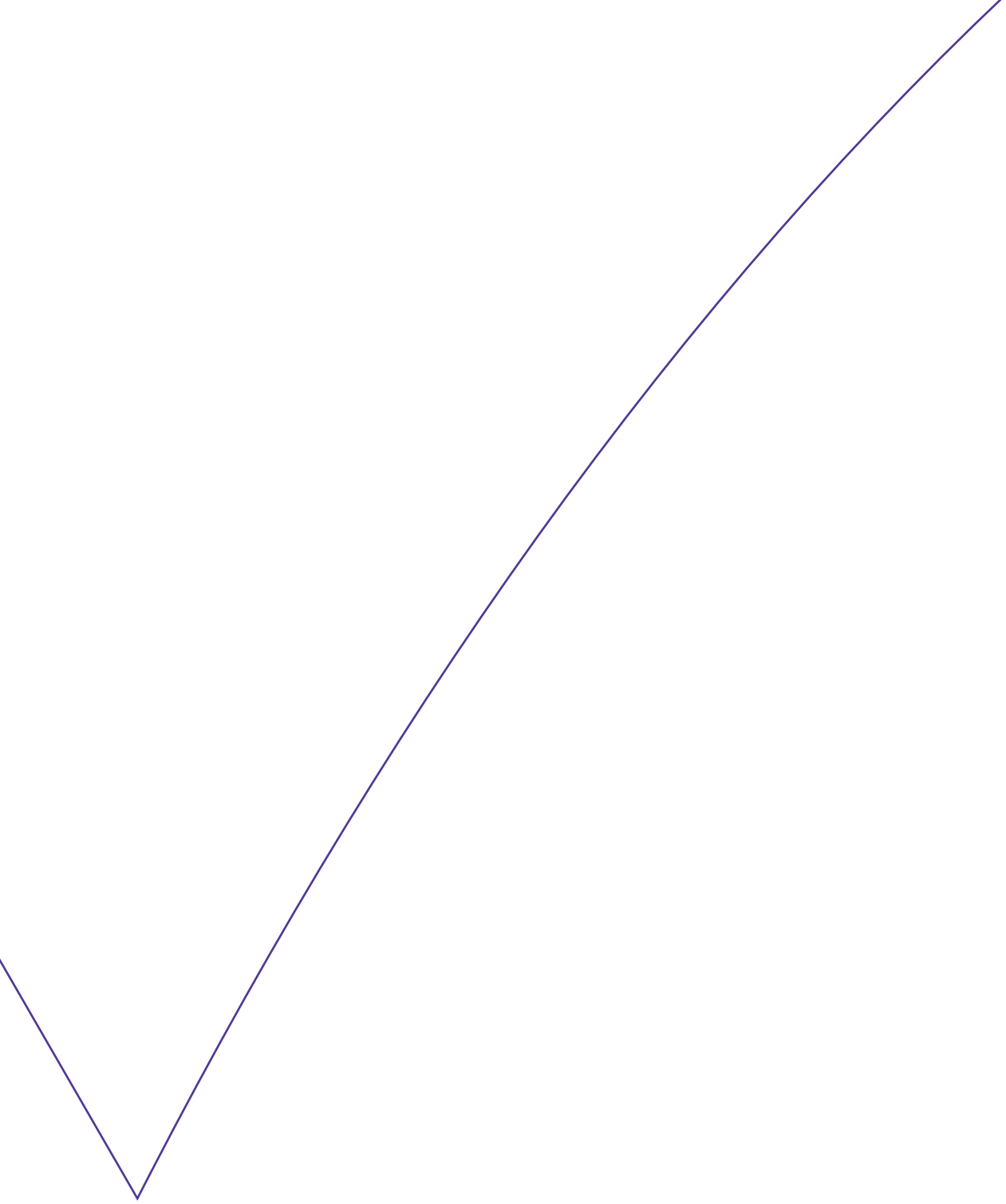
Le *boom* du *Machine Learning* conduit à la fois à des algorithmes de plus en plus performants, mais également à une généralisation de l'accès à ces algorithmes, à l'instar d'applications **grand public** comme Lyrebird (application gratuite de *Deepfake* audio) ou Zao (application chinoise de *face-swapping* ayant fait le *buzz* à sa sortie durant l'été 2019). Ces applications, créées dans un objectif de divertissement, représentent un nouvel outil performant, accessible et facile d'utilisation pour des personnes malveillantes. Il ne fait aucun doute que les attaques utilisant ce type d'applications augmenteront dans les prochaines années. Fraude au président, atteinte à l'image, création de fausses preuves juridiques, contournement d'authentification biométrique : les cas d'usages envisageables sont nombreux et effrayants.

LE *DEEPFAKE*, UNE FATALITÉ ?

Vues les perspectives de risque, des solutions sont activement recherchées pour se protéger contre l'usage des *Deepfake*. Le sujet est complexe et pris très au sérieux. À l'approche de la campagne électorale américaine de 2020, **Facebook** a lancé son « ***Deepfake Detection Challenge*** », concours public de développement d'outils *anti-Deepfake*, avec pas moins de 10 millions de dollars à la clé pour le vainqueur.

Les manipulations opérées pour créer les *Deepfake* laissent des traces. Des **éléments non naturels**, comme le nombre de clignements d'œil, l'orientation relative des éléments ou distorsions du visage, **peuvent être détectés**. Ces méthodes reposent sur les imperfections des techniques de *Deepfake* utilisées, et restent efficaces jusqu'à ce que les attaquants comprennent et adaptent leurs méthodes pour passer sous les seuils de détection.

Des **solutions de préventions** sont également à l'étude, comme la création de **filtres *anti-Deepfake*** appliqués aux médias en amont de leur publication. Ils introduisent dans l'image un bruit imperceptible à l'œil nu mais perturbant l'apprentissage des algorithmes *Deepfake* (de manière similaire aux exemples contradictoires présentés plus tôt). D'autres solutions proposent des mécanismes innovants destinés à garantir l'authenticité des fichiers, à l'image d'*Amber Authenticate*, solution basée sur la *Blockchain*.



The Positive Way

WAVESTONE

www.wavestone.com

Dans un monde où savoir se transformer est la clé du succès, Wavestone s'est donné pour mission d'éclairer et guider les grandes entreprises et organisations dans leurs transformations les plus critiques avec l'ambition de les rendre positives pour toutes les parties prenantes. C'est ce que nous appelons « The Positive Way ».

Wavestone rassemble plus de 3000 collaborateurs dans 8 pays. Il figure parmi les leaders indépendants du conseil en Europe, et constitue le 1er cabinet de conseil indépendant en France.

Wavestone est coté sur Euronext à Paris et labellisé Great Place To Work®.