REFERENTIEL DE CERTIFICATION DE PROCESSUS POUR L'IA

Conception, développement, évaluation et maintien en conditions opérationnelles

Réf. rédacteur :
LNE/DEC/CITI/CH
LNE/DEC/IA/GA

Révision n°2.0

Approbation LNE : 12/07/2021

LABORATOIRE NATIONAL DE MÉTROLOGIE ET D'ESSAIS

REVISIONS DU DOCUMENT

Version	Date	Motif de la mise à jour
1.0	25/04/2021	Version initiale publiée lors de la consultation publique
2.0	12/07/2021	Version intégrant les commentaires reçus lors de la consultation publique

Table des matières

CHAPITRE I : GENERALITES

I.1) Contexte et objet

L'intelligence artificielle (IA) a connu ces dernières années d'importants développements dans de nombreux secteurs professionnels (robots industriels collaboratifs, robots d'inspection et de maintenance, système de mobilité autonome, etc.) et domestiques (robots d'assistance à la personne, dispositifs médicaux, assistants personnels, etc.).

Les niveaux de performance, de robustesse, d'éthique, d'explicabilité atteints par les différentes systèmes d'IA doivent encore être démontrés de manière fiable. Les utilisateurs finaux disposeront ainsi des garanties conditionnant l'acceptabilité de ces technologies. Ils pourront choisir parmi différentes solutions existantes grâce à des références communes objectives et non ambiguës. Les développeurs bénéficieront quant à eux de repères pour orienter leurs efforts de R&D et de contrôle.

Un système d'Intelligence Artificielle est un logiciel qui peut, pour un ensemble d'objectifs définis par des humains, générer des sorties telles que du contenu, des prédictions, des recommandations ou des décisions influençant l'environnement avec lequel elles interagissent ; et développé avec une ou plusieurs des approches et techniques suivantes :

1) les approches par apprentissage automatique[1] (ou Machine Learning) incluant l'apprentissage supervisé[2], non-supervisé ou par renforcement, utilisant une vaste variété de méthodes incluant l'apprentissage profond (ou Deep Learning)[3],

2) les approches logiques et basées sur la connaissance incluant la représentation de connaissances, la programmation logique inductive, les bases de connaissance, les moteurs d'inférence et déductifs, le raisonnement symbolique et les systèmes experts[4],

3) les approches statistiques, les estimations Bayésiennes, les méthodes de recherche et d'optimisation. [5]

[1] Apprentissage automatique : Processus par lequel un algorithme évalue et améliore ses performances sans l''intervention d''un programmeur, en répétant son exécution sur des jeux de données jusqu''à obtenir, de manière régulière, des résultats pertinents.

[2] Apprentissage supervisé : Apprentissage automatique dans lequel l'algorithme s'entraîne à une tâche déterminée en utilisant un jeu de données assorties chacune d'une annotation indiquant le résultat attendu.

[3] Cas particulier de l'apprentissage automatique (Machine Learning) reposant sur l'utilisation d'un algorithme de réseau de neurones à plusieurs couches. Plus le nombre de couches est important, plus l'apprentissage est dit profond.

[4] Système expert : Outil déterministe capable de répondre à des questions, en effectuant un raisonnement à partir de faits et de règles connues (s'appuyant sur les connaissances tirées de l'expertise humaine). Le moteur d'inférence du système d'inférence viendra utiliser les faits et règles pour produire de nouveaux faits, jusqu'à parvenir à la réponse à la question experte posée. Cette approche de l'IA ne repose pas sur de l'apprentissage automatique.

[5] Définition issue de la proposition de règlement du parlement européen et du conseil établissant des règles harmonisées concernant l'intelligence artificielle publiée le 21/04/2021.

La certification est une procédure par laquelle une tierce partie, l'organisme certificateur, donne une assurance écrite qu'un système d'organisation, un processus, une personne, un produit ou un service est conforme à des exigences spécifiées dans une norme ou un référentiel.

Ce référentiel de certification a vocation à accompagner cette transformation profonde de la société en apportant de la confiance dans les systèmes d'IA, afin d'en sécuriser les usages et de favoriser leur déploiement.

Il décrit les modalités de certification des **processus de conception, développement, évaluation et maintien en conditions opérationnelles (MCO) des systèmes d'IA reposant sur des algorithmes d'apprentissage automatique** conformes aux besoins de leurs utilisateurs. Ces besoins peuvent être définis en lien direct avec l'utilisateur, dans le cas de systèmes développés spécifiquement pour un client, ou bien, dans le cas où ces systèmes sont développés pour un client générique, décrits de façon non ambiguë et complète de manière à ce que les futurs utilisateurs aient pleinement conscience des avantages et limites de la fonctionnalité d'IA[6] développée.

Le référentiel est applicable aussi bien aux processus visant à développer et maintenir des fonctionnalités d'IA reposant sur de l'apprentissage incrémental[7], continu, qu'à celles dont le modèle appris est figé une fois la fonctionnalité d'IA déployée.

Les processus associés aux systèmes d'IA symboliques sont exclus du domaine d'application du référentiel, mais ceux associés à l'apprentissage automatique sont inclus y compris lorsqu'ils sont combinés à des processus symboliques (cas des systèmes hybrides[8]).

[6] Fonctionnalité d'Intelligence Artificielle : Fonctionnalité dotant un système d'IA de sa capacité d'analyse, de raisonnement ou de décision lui permettant de générer des sorties telles que du contenu, des prédictions, des recommandations ou des décisions influençant l'environnement avec lequel elle interagit.

[7] Apprentissage automatique réalisé sur des données regroupées en lots (batchs), les lots étant renouvelés périodiquement, au fur et à mesure de l'accumulation de nouvelles données tout au long du cycle de vie de la fonctionnalité d'IA.

[8] Système hybride : Système d'Intelligence Artificielle intégrant à la fois des techniques d'apprentissage automatique à partir de données et des modèles permettant d''exprimer des contraintes et d'effectuer des raisonnements logiques.

I.2) Domaine d'application : certification de processus

Un système intégrant une fonctionnalité d'IA est un produit au sens de l'ISO 9000 : 2015[9].

Ce référentiel vise à définir des exigences communes liées aux processus de conception, de développement, d'évaluation et de maintien en condition opérationnelle de tous types de fonctionnalités d'IA dotées d'apprentissage automatique. Il couvre de ce fait tous les secteurs d'activités dans lesquels les systèmes d'IA sont utilisés, afin de garantir l'application des meilleures pratiques favorisant la confiance accordée dans ces systèmes.

Il ne vise pas à définir des exigences applicables aux fonctionnalités d'IA en elles-mêmes et donc spécifiques à leurs utilisations.

Les processus de conception, de développement, d'évaluation et de maintien en conditions opérationnelles concernés par le référentiel sont définis de la façon suivante :

1. Le processus de conception consiste à transformer une expression de besoin en spécifications fonctionnelles.

2. Le processus de développement consiste à traduire ces spécifications en une version de la fonctionnalité d'IA prête à être évaluée.

3. Le processus d'évaluation consiste à vérifier et valider la conformité du système aux spécifications définies avant son déploiement.

4. Le processus de maintien en conditions opérationnelles vise à assurer la conformité de la fonctionnalité d'IA aux spécifications définies après son déploiement et tout au long de sa phase d'exploitation.

[9] Produit : élément de sortie d'un organisme qui peut être produit sans transaction entre l'organisme et le client. Un logiciel est constitué d'informations quel que soit le support de livraison (par exemple programme informatique). – ISO 9000 : 2015 §3.7.6

CHAPITRE II : PROCESSUS D'ATTRIBUTION ET DE SUIVI DES CERTIFICATS

II.1) Processus d'attribution de la certification

Le processus de certification est constitué de plusieurs étapes successives, dont les principales sont :

1. l'instruction du dossier de demande (audit d'étape 1),
2. la réalisation de l'audit de certification initial (audit d'étape 2),
3. le retour sur les fiches de non-conformité, le cas échéant,
4. l'étude du rapport d'audit par un comité de lecture qui formule un avis quant à la certification,
5. le cas échéant, l'émission du certificat une fois la décision de certification entérinée.

L'entreprise candidate à la certification doit, pour que l'organisme de certification puisse établir l'offre de certification initiale, lui transmettre la documentation décrivant précisément le périmètre couvert par la certification (organisation, équipes, effectifs) ainsi que les éventuelles certifications couvrant ce périmètre (notamment ISO 9001 ou ISO 27001).
La durée de l'audit de certification pourra être diminuée en cas de certification ISO 9001 sur un système de management couvrant les processus visés par la certification.

II.1.1) Audit d'étape 1

A l'acceptation de l'offre de certification initiale, l'entreprise doit transmettre à l'organisme de certification la documentation relative au(x) processus à certifier. Cette documentation doit décrire les processus de conception, de développement, d'évaluation et de maintien en conditions opérationnelles visés par la certification. Elle doit inclure les éléments suivants :

- Documents généraux liés aux processus à certifier :
  - la déclaration d'applicabilité (cf. III.1.1),
  - la définition des entrées et sorties des processus à certifier ainsi qu'une description de leurs interfaces (cf. III.1.2),
  - une liste des procédures mises en place dans le cadre des processus à certifier,
  - une liste des contrôles et objectifs permettant de démontrer la capacité des processus certifiés à atteindre les résultats attendus (cf. III.2.1),
  - une liste des prestataires externes intervenant dans les processus certifiés (cf. III.2.4),
  - une analyse de criticité des prestataires externes intervenant dans les processus à certifier (cf. III.2.4),
  - une analyse des contrôles effectués (cf. III.2.6),

- Documents liés à chaque fonctionnalité d'IA mise au point dans le cadre des processus à certifier :
  - spécifications relatives aux fonctionnalités d'IA (cf. III.3.1),

- analyses préliminaires de risques (cf. III.3.5),
- caractéristiques des domaines d'utilisation avec justification de pertinence des facteurs d'influence (cf. III.4.1.5),
- listes des contre-indications et non-indications (cf. III.4.1.6 & III.4.1.7),
- méthodes d'estimation des distributions réelles des données dans les bases d'apprentissage (cf. III.4.2.2.1),
- listes des évènements rares intégrés aux bases d'apprentissage ainsi que leur fréquence d'apparition et la méthode utilisée pour les déterminer (cf. III.4.2.2.2),
- distributions des cas couverts par les bases de test (cf. III.4.2.3.1),
- listes des évènements rares intégrés aux bases de test ainsi que leur fréquence d'apparition et la méthode utilisée pour les déterminer (cf. III.4.2.3.2),
- méthodes permettant de garantir la qualité du processus d'apprentissage initial (cf. III.4.3.1.),
- protocoles d'évaluation mis en œuvre (cf. III.5.1),
- analyses de risques détaillées (cf. III.5.11)
- descriptions des mécanismes de contrôle de l'évolution des performances dans le cadre de la MCO (cf. III.6.1.7),
- méthodes de mise à jour après déploiement (cf. III.6.2.1),

➢ Ensemble de la documentation tenue à disposition ou communiquée aux clients (Fiche produit dans le cas des fonctionnalités d'IA développées pour un client générique).

L'audit d'étape 1 consiste à déterminer si l'audit de conformité des processus (audit d'étape 2) est possible compte tenu du degré d'aboutissement de la documentation transmise par le demandeur. Pour ce faire, il est constaté :

- si le périmètre de certification est suffisamment précis et non ambigu,
- si les exclusions d'exigences sont dûment justifiées,
- si les principaux documents et procédures exigés par le présent référentiel sont bien présents.

Il s'agit de déterminer si les éléments nécessaires au fonctionnement des processus conformes au référentiel sont présents, et non d'en vérifier l'application, objet de l'audit d'étape 2.

A l'issue de l'audit d'étape 1, l'organisme de certification informe le demandeur du résultat.

Dans le cas où cette étape 1 conclut à l'irrecevabilité du dossier, il appartient au demandeur de la certification de répondre à l'organisme de certification en fournissant les documents manquants. Une offre complémentaire pourra dans ce cas être adressée par l'organisme de certification si un second audit d'étape 1 est nécessaire.

II.1.2) Planification de l'audit d'étape 2

Dans le cas où l'audit d'étape 1 est satisfaisant, le dossier est recevable et l'organisme prend contact avec l'entreprise candidate à la certification, afin de définir les lieux et dates de l'audit d'étape 2.

La durée de l'audit d'étape 2 peut être augmentée s'il est nécessaire de se déplacer sur plusieurs sites, si des sous-traitants interviennent dans les processus de conception, de développement, d'évaluation et de maintien en conditions opérationnelles des fonctionnalités d'IA, visés par la certification, dont la maitrise n'est pas assurée par le candidat à la certification et ne sont pas certifiés, ou encore s'il est nécessaire de faire appel à un interprète.

II.1.3) Réalisation de l'audit d'étape 2

L'entreprise doit appliquer toutes les exigences du présent référentiel, si elles sont applicables à ses processus. Tous les points du référentiel et des textes de référence relatifs aux processus visés par la certification sont examinés. Si l'entreprise candidate à la certification ne réalise qu'une partie des opérations visées (conception, développement, évaluation, MCO), seuls les processus liés aux activités concernées seront audités et certifiés et ceci doit être documenté dans la déclaration d'applicabilité. Seuls les processus certifiés seront mentionnés sur le certificat. Les informations détaillées concernant le périmètre de certification, issues de la déclaration d'applicabilité seront reprises en annexe du certificat.

L'ensemble des paragraphes du référentiel audités est mentionné dans le plan d'audit.

L'audit d'étape 2 a préférablement lieu chez le demandeur de la certification, sur le ou les sites où sont réalisées les activités de conception, de développement, d'évaluation et de maintien en conditions opérationnelles des fonctionnalités d'IA.

Le demandeur de la certification doit s'assurer de la disponibilité :

- d'interlocuteur(s) maîtrisant les processus mis en œuvre,
- toute autre personne jugée pertinente,
- des informations documentées exigées par le présent référentiel et permettant d'avoir la preuve que les processus sont conformes aux exigences du présent référentiel,
- des informations documentée permettant de démontrer la conformité des fonctionnalités d'IA développées aux exigences spécifiées (cf. III.3.1).

Il est rappelé que l'audit repose sur un échantillonnage des informations disponibles. L'absence de non-conformité constitue une présomption de conformité et non une preuve de conformité aux exigences du référentiel.

II.1.4) Réponse aux fiches de non-conformité

Dans le cas où une ou plusieurs non-conformités (NC) sont constatées durant l'audit, une fiche décrivant la NC est rédigée par l'équipe d'audit puis présentée et transmise à l'entreprise en réunion de clôture d'audit. Une non-conformité peut être majeure ou mineure.

Une NC majeure est bloquante pour la certification : elle devra être corrigée en vue de l'obtention de la certification.

Une NC mineure n'est pas bloquante pour la certification mais devra être corrigée avant le prochain audit de suivi, sous peine de suspension de certificat.

La non-conformité est classée majeure lorsque, sur la base d'évidences objectives, l'une des situations suivantes est rencontrée :

- présence d'un risque significatif[10] pour la capacité du processus à maîtriser la conformité des fonctionnalités d'IA aux exigences spécifiées[11],
- absence d'un contrôle efficace de processus,
- non-respect systématique ou répété d'une exigence spécifiée,
- plusieurs non-conformités mineures associées à la même exigence.

Dans les autres cas, la non-conformité est classée mineure. La justification du classement en non-conformité mineure ne peut pas être basée sur des faits futurs.

L'entreprise candidaté a alors 20 jours ouvrés après la fin de l'audit pour lui retourner chaque fiche complétée par l'analyse des causes de la NC et les actions engagées pour y remédier.

Après analyse des actions proposées par le demandeur de la certification, le responsable d'audit (RA) se prononce sur leur pertinence, préconise le type de suivi nécessaire à la NC et transmet son rapport complet à l'organisme de certification dans les 10 jours ouvrés suivant la réception des réponses de l'entreprise.

II.1.5) Revue du rapport d'audit

Après réception des éventuelles fiches de NC complétées par le candidat à la certification et le RA, l'organisme de certification analyse le rapport d'évaluation et l'avis du responsable d'audit. A la lecture du rapport il peut demander des informations complémentaires au demandeur de la certification, avant le passage en comité de lecture.

II.1.6) Décision du comité de lecture

Le comité de lecture est chargé de rendre un avis sur la décision de certification dans le processus d'attribution, de surveillance, de retrait ou de suspension des certificats. Il est composé au minimum :

- d'un représentant de la direction de l'organisme de certification, n'ayant pas de responsabilités en tant que chef de projet certification et ne peut avoir participé à l'évaluation,
- d'un chef de projet certification indépendant du dossier,
- d'un chef de projet certification en charge de présenter le dossier.

Le comité est présidé par le représentant de la direction de l'organisme de certification et a pour mission :

- d'examiner les rapports d'évaluation et de formuler un avis et une recommandation sur les décisions à prendre, notamment sur le type et la durée du suivi d'une NC,
- le cas échéant, d'examiner dans un premier temps les appels contre les décisions de l'organisme de certification et de formuler un avis sur les suites à donner,
- d'évaluer la qualité des rapports d'audit.

[10] le niveau de risque doit être justifié dans la fiche de NC

[11] les exigences spécifiées peuvent être réglementaires, formulées par les clients ou par l'entreprise candidate à la certification (cf. III.3.1).

La décision de certification de l'organisme de certification s'appuie sur l'examen des éléments du dossier et du rapport d'audit. Chaque décision de certification est matérialisée par une notification au demandeur et le cas échéant par l'émission d'un certificat.

II.2) Surveillance du certificat

La certification est prononcée pour une durée de trois ans. Elle suit un cycle d'audit composé de l'audit de certification initial, d'un audit de suivi 1 la première année suivant la certification, puis d'un audit de suivi 2 la deuxième année.

Un nouveau cycle de trois ans est ensuite initié par un audit de renouvellement suivi de deux audits de suivi.

Pour éviter toute rupture de certification entre deux cycles, il est préconisé de réaliser l'audit de renouvellement avant l'échéance du certificat.

La durée d'un audit de suivi est le tiers de la durée d'un audit initial et la durée d'un audit de renouvellement est de deux tiers de la durée d'un audit initial.

Afin de planifier cette surveillance, l'organisme de certification envoie un questionnaire à l'entreprise dont les processus sont certifiés, afin de connaître les éventuelles évolutions apportées depuis l'audit précédent.

Les étapes ci-dessous sont identiques à celles de l'audit initial :

1. planification de l'audit de suivi / renouvellement,
2. réalisation de l'audit,
3. réponses aux éventuelles fiches de NC,
4. revue du rapport,
5. avis du comité de lecture et décision de renouvellement, suspension, retrait.

II.3) Résiliation, suspension et retrait du certificat

Les motifs de résiliation, suspension ou de retrait d'un certificat par l'organisme de certification sont les suivants :

- le non-respect des exigences contractuelles,
- le refus par l'entreprise de réaliser l'audit de suivi dans le délai imparti notifié par l'organisme de certification,
- le refus de mise en œuvre d'actions correctives requises dans le délai imparti notifié par l'organisme de certification,
- la demande d'annulation de tout ou partie de la certification par l'entreprise.

L'organisme de certification notifie alors formellement la suspension ou le retrait au titulaire, en indiquant dans le premier cas les conditions de levée de la suspension, notamment les mesures correctives à prendre le cas échéant.

Afin de lever une suspension, l'organisme de certification procède aux vérifications nécessaires pour rétablir la certification. Si tel est le cas, la suspension est levée et la certification remise en vigueur avec notification au titulaire.

Lorsque la certification est retirée ou suspendue, l'ex-titulaire de la certification se doit de cesser toute utilisation de la marque, sous peine de poursuites.

CHAPITRE III : EXIGENCES APPLICABLES

III.1) Définition des processus inclus dans le périmètre de certification : déclaration d'applicabilité

III.1.1. L'entreprise doit documenter, dans la déclaration d'applicabilité, ses processus de conception, développement, évaluation et maintien en conditions opérationnelles (MCO), couverts par la certification et leurs interfaces, les sites sur lesquels ils sont appliqués, les effectifs alloués à ces processus, et fournir une justification pour toute exclusion d'applicabilité des exigences du présent référentiel.
La conformité au présent référentiel ne peut être déclarée que si les exigences déclarées comme non applicables n'ont pas d'incidence sur la capacité de l'entreprise à assurer la conformité des fonctionnalités d'IA déployées et maintenues aux besoins spécifiés (cf. III.3.1).
Par exemple il peut être déterminé pour une entreprise ne fournissant pas de services après déploiement que le processus de MCO est de fait non inclus dans le périmètre de certification et que les exigences liées à la MCO sont non applicables. Si certaines exigences ne s'appliquent pas à certaines technologies ou types d'algorithmes ceci doit être clairement justifié dans la déclaration d'applicabilité.
Les processus et leurs interfaces peuvent par exemple être documentés via une cartographie processus.

III.1.2. L'entreprise doit établir, mettre en œuvre de façon maîtrisée et tenir à jour des processus de conception, de développement, d'évaluation et de maintien en conditions opérationnelles en accord avec les exigences du présent chapitre III dont l'objectif final est de s'assurer de la conformité des fonctionnalités d'IA déployées et maintenues aux exigences les concernant définies lors de la phase de conception (cf. III.3.1).
Pour cela l'entreprise doit :

- documenter les éléments d'entrée requis et les éléments de sortie attendus des processus. Les éléments de sortie doivent être :
  - conformes aux besoins identifiés,
  - cohérents avec les entrées des processus ultérieurs ;
- déterminer les interfaces des processus ;
- déterminer les ressources nécessaires pour assurer le bon fonctionnement de ces processus ;
- évaluer les processus et mettre en œuvre toutes les actions requises pour s'assurer de la capacité de ses processus à atteindre les résultats attendus (cf. III.2.5 et III.2.6).

III.2) Planification, fonctionnement, évaluation des processus

III.2.1. L'entreprise doit :

- concernant les risques identifiés liés à l'usage de la fonctionnalité d'IA (cf. III.3.5 et III.5.11) :
  - planifier des actions efficaces de mitigation[12] des risques ;
  - planifier l'évaluation des actions de mitigation des risques menées ;
- établir et documenter des contrôles et objectifs permettant de démontrer la capacité des processus certifiés à garantir que les fonctionnalités d'IA déployées et maintenues respectent les exigences les concernant qui ont été définies (cf. III.3.1). Les objectifs doivent être :
  - cohérents, notamment avec les exigences applicables,
  - mesurables,
  - pertinents pour la conformité des fonctionnalités d'IA développées,
  - surveillés,
  - communiqués,
  - mis à jour si nécessaire ;
- déterminer, pour atteindre chaque objectif ainsi défini :
  - les actions à mener,
  - les ressources nécessaires,
  - les personnes en charge de réaliser l'action,
  - la date d'échéance des actions à mener et d'atteinte de l'objectif,
  - la façon d'évaluer l'atteinte de l'objectif.

III.2.2. L'entreprise doit :

- déterminer les compétences nécessaires aux personnes qui participent à la capacité des processus à atteindre les résultats attendus,
- pouvoir démontrer la compétence de ces personnes (via des formations ou expériences par exemple).

III.2.4. Si des prestataires externes interviennent dans les processus à certifier, l'entreprise doit :

- documenter la criticité du prestataire au regard de la conformité des processus à certifier et de leur capacité à atteindre les résultats attendus ; s'assurer que les services fournis par les prestataires externes sont conformes aux exigences du présent référentiel, et permettent de répondre aux spécifications relatives à la fonctionnalité d'IA (cf. III.3.1) le cas échéant et ne compromettent pas la capacité des processus certifiés à atteindre les résultats attendus ;
- communiquer au prestataire concerné les modalités de contrôle/évaluation auxquelles il sera soumis.

[12] Les mesures de mitigation des risques peuvent être : éviter le risque, éliminer la source du risque, réduire la probabilité ou l'impact du risque (par exemple par l'ajout d'une redondance ou d'un système alternatif pouvant assurer le fonctionnement en cas de besoin), partager le risque (communiquer et / ou transférer le risque à l'utilisateur par exemple) ou accepter le risque sur la base d'une décision éclairée.

Par exemple il est possible de suivre ces conditions de sous-traitance via un audit du sous-traitant ou de tenir compte d'une certification suivant la norme ISO 9001 émise par un organisme accrédité et couvrant un périmètre adéquat

III.2.5 L'entreprise doit assurer la maîtrise des processus certifiés en contrôlant :

- la bonne réalisation des contrôles des processus planifiés (cf. III.2.1) aux étapes opportunes,
- que des activités de vérification et de validation ont été correctement effectuées aux étapes opportunes afin de s'assurer que les éléments de sortie des processus certifiés et la fonctionnalité d'IA en résultant satisfont aux exigences définies,
- que toutes les actions jugées pertinentes pour corriger une anomalie ont bien été mises en œuvres (cf. III.2.8),
- que les informations documentées sont bien disponibles et traçables (cf. III.2.4).

III.2.6 L'entreprise doit analyser les résultats des contrôles effectués à une fréquence définie, afin d'évaluer :

- la conformité de la fonctionnalité d'IA aux exigences définies (cf. III.3.1),
- la capacité des processus à atteindre les résultats attendus,
- l'efficacité des actions de traitement des risques liés à l'usage de la fonctionnalité d'IA identifiés,
- l'efficacité des actions correctives des anomalies détectées concernant la fonctionnalité d'IA ou les processus,
- la maîtrise des prestataires externes.

Cette évaluation doit être documentée et peut se faire via un audit interne.

III.2.7 L'entreprise doit faire évoluer ses processus et mettre en œuvre toutes les actions requises pour :

- analyser et corriger les anomalies détectées lors des contrôles, qu'elles concernant les fonctionnalités d'IA ou les processus,
- prévenir ou réduire les effets indésirables des fonctionnalités d'IA développées.

Il est par exemple possible de prévoir des revues de processus à intervalles réguliers ou de mettre à jour les processus suite à la détection d'anomalies lors des contrôles effectués.

III.3) Processus de conception

Les éléments d'entrée et sortie à documenter sont *à minima* :

Pour les éléments d'entrée :

- les exigences et attentes du client relatives à la fonctionnalité d'IA,
- les exigences réglementaires applicables,
- les exigences issues d'activités et/ou usages similaires (normes, règles de l'art, retours d'expérience).
  Le client peut être spécifique dans le cas où la fonctionnalité d'IA est développée pour un acteur en particulier ou bien générique dans le cas où il est prévu que la fonctionnalité d'IA soit distribuée largement.

Pour les éléments de sortie :

- les spécifications de la fonctionnalité d'IA,
- les exigences concernant la documentation associée,
  Par exemple doivent être définis les besoins en documentation utilisateur ou concernant la fiche produit. Il peut également être défini que les plans de test permettant la vérification des exigences relatives à la fonctionnalité d'IA doivent être associés à la documentation utilisateur.
- les besoins de communication avec le client et les utilisateurs,
- une analyse préliminaire de risques (cf. III.3.5).

III.3.1. L'entreprise doit définir les spécifications relatives à la fonctionnalité d'IA, les documenter et justifier les critères d'acceptabilité pour chacune des exigences ainsi définies. Ces éléments doivent être communiqués au client selon les modalités prévues au § III.7.4. La définition des spécifications doit notamment considérer les exigences éventuellement non formulées par les clients et couvrir les spécifications:

- d'usage du système dans son ensemble,
  par exemple, pour un système de reconnaissance faciale par caméra le descriptif pourrait être : "le système de reconnaissance biométrique vise à reconnaître, grâce à une caméra, les personnes qui se présentent au sas de sécurité de l'aéroport, à la porte d'entrée d'un particulier et dans la rue"
- d'usage de la fonctionnalité d'IA (la tâche automatisée et sa finalité doivent être précisées),
  par exemple, pour une tache de segmentation de mélanome sur des photos "l'algorithme segmente, grâce à un polygone à 10 sommets, la tache sombre détectée, et lui associe un label ("mélanome" ou "RAS")"
  Autre exemple d'application : reconnaissance d'un pattern d'attaque dans les logs d'un système d'information.
  Autres exemples de tâche macros : classification, classement, régression, etc.
- de documentation,
- de communication auprès du client (sur l'origine et la composition des bases de données d'apprentissage, les résultats des évaluations, l'explicabilité[13] et l'interprétabilité[14] de la fonctionnalité, les post-traitements suggérés, les performances attendues et contraintes éventuelles sur certaines ressources hardware, les contraintes éventuelles liées à la qualité et la maintenance des

[13] Explicabilité : Capacité d'une fonctionnalité d'IA de rendre compte explicitement des éléments conduisant ou ayant conduit à une évaluation ou à une décision, à partir de données et caractéristiques connues de la situation.
[14] Interprétabilité : Capacité de rendre compréhensible, pour une catégorie d'utilisateurs donnée, le fonctionnement d'un système d'intelligence artificielle.

capteurs, son caractère open-source ou non, les modifications apportées à la fonctionnalité d'IA après déploiement, les modalités de sous-traitance, sur la gestion des incidences négatives potentielles, etc.),

Par exemple, il peut être défini que les temps d'exécution dans le pire des cas doivent être communiqués pour plusieurs hardware cibles.

- de données d'entrée :
  - de types, de formats *(images au format png ou jpg, audio mpeg, mp3, structures spécifiques au format texte en xml, json, csv ou binaire spécifique etc.)* et de la compatibilité de ces formats avec d'autres solutions ou environnement (dans le cas où l'interopérabilité est un enjeu pour le client),
  - de source/mode d'acquisition *(capteurs internes à la solution matérielle, données importées d'une base de données, etc.)* utilisées par la fonctionnalité d'IA (si la source dépend des cas d'usage, celle-ci doit être spécifiée pour chaque cas d'usage),

    Par exemple, pour un système de reconnaissance faciale par caméra les données d'entrée sont acquises par des caméras. Dans le cas du sas sécurisé, la caméra doit être de tel modèle et dans le cas de la caméra de vidéosurveillance dans la rue, elle doit être de tel autre modèle.
  - de fréquence et flux d'alimentation pour chaque type de données d'entrée (le flux limitant doit être précisé),

    Par exemple, pour un système de reconnaissance faciale par caméra la fréquence d'échantillonnage des images, pour les algorithmes reposant sur la vidéo, ajouter le nombre d'images nécessaires à l'algorithme pour traiter l'information

    Autre exemple : fréquence d'échantillonnage audio, fréquence de sortie de capteurs de température, etc.
  - de présence et de nature d'attributs critiques[15],
- de données de sortie :
  - de types, de formats et de la compatibilité de ces formats avec d'autres solutions ou environnement (dans le cas où l'interopérabilité est un enjeu pour le client),
  - de fréquence et de flux de sortie,
- de domaine d'utilisation[16],
- d'apprentissage (notamment dans le cas où des réapprentissages sont possibles après déploiement),
- de niveaux d'autonomie (actions et contrôles humains sur les tâches automatisées),
- de performances[17], le cas échéant en termes de :
  - précision,
  - fiabilité,
  - temps d'exécution de l'apprentissage / temps d'exécution de la tache automatisée sur du matériel cible,

[15] Attribut critique : Attribut dont l'absence ou l'exactitude peuvent fausser grandement le résultat.

[16] Domaine d'utilisation : Description de l'environnement (conditions météorologiques, population visée, etc.) pour lequel la fonctionnalité d'IA est conçue.

[17] Performance : Degré selon lequel un système ou composant accomplit ses fonctions désignées avec un ensemble des contraintes données, telles que la vitesse, la précision ou l'utilisation de la mémoire, etc.

  - résilience[18] aux attaques et valeurs aberrantes,
  - reproductibilité[19],
- de confidentialité :
  - respect de la vie privée,
  - accès aux données,
- de transparence , le cas échéant en termes de :
  - explicabilité et interprétabilité (notamment les éléments requis et la durée de conservation de ceux-ci),
  - traçabilité, auditabilité des apprentissages et/ou résultats,
- de diversité, non-discrimination et équité,
- d'impact sociétal et environnemental,
- de maintenance après livraison,
- réglementaires,
- normatives,
- liées aux retours d'expérience,
- jugées nécessaires par le développeur de la fonctionnalité,
- liées aux parties intéressées jugées pertinentes,
- liées aux conséquences potentielles d'une défaillance de la fonctionnalité d'IA.

Il est possible qu'aucune exigence particulière ne soit définie par rapport à un des points ci-dessus mais ceci doit être clairement formalisé dans les spécifications tenues à disposition du client.

Par exemple, le client peut ne pas souhaiter définir d'exigence particulière en termes de traçabilité et/ou d'impact social et environnemental mais cela doit être formellement tracé.
Concernant la définition des besoins liés à l'éthique et notamment en termes d'action/contrôle humain, de respect de la vie privée, de transparence, de diversité, non-discrimination, équité, impact social et environnemental et responsabilité, il est, par exemple, possible de se baser sur les « lignes directrices en matière d'éthique pour une IA digne de confiance », publiées par le groupe d'experts indépendants de haut niveau sur l'intelligence artificielle constitué par la commission européenne en juin 2018.
Des exigences en termes d'impact environnemental peuvent par exemples être liées à la consommation d'énergie.
Les indicateurs permettant la validation de chacune des exigences de la solution finale doivent être clairs et compris du développeur et du client dans le cas d'une relation client/fournisseur.
Pour les seuils d'acceptabilité des performances, il peut être défini pour une fonctionnalité d'IA de recommandation de film un seuil de précision de 90% alors que ce même seuil ne serait pas acceptable dans le domaine médical avec des questions d'intégrité des personnes. De la même manière, par rapport au temps de prise de décision, le seuil de durée de prise de décision acceptable de 60 minutes est cohérent pour une analyse de sang mais non acceptable dans le domaine des voitures autonomes.
Les critères d'acceptabilité concernant la prise en compte d'une règlementation peuvent être binaires.
Les critères d'acceptabilité des conséquences potentielles d'une défaillance de la fonctionnalité peuvent être détaillés par rapport à des scénarios de défaillance spécifiquement définis.

[18] Résilience : Capacité d'une fonctionnalité d'IA à maintenir sa conformité aux exigences attendues en présence de données d'entrée extérieures à son domaine d'utilisation (par exemple en cas de panne, d'incident intentionnel ou non et/ou de sollicitation extrême).
[19] Reproductibilité d'une expérimentation : Fidélité d'une expérimentation selon un ensemble de conditions de reproductibilité (réalisée pour un même objet ou des objets similaires dans un ensemble de conditions qui comprennent des lieux, des opérateurs et des systèmes de mesure différents).

III.3.2. Les spécifications de la fonctionnalité d'IA (cf. III.3.1) doivent être tenues à disposition de toute personne participant à la conception, au développement, à l'évaluation ou au maintien en conditions opérationnelles de la fonctionnalité d'IA.
Par exemple les spécifications de la fonctionnalité d'IA peuvent être mises à disposition des clients par le biais de la fiche produit, disponibles sur le site internet du fournisseur, dans la documentation à destination du client etc.

III.3.3. Dans le cadre d'une relation client/fournisseur (BtoB), ces exigences et les critères d'acceptabilité doivent être établis en lien avec le client et l'entreprise doit, avant de lancer le développement de la fonctionnalité, vérifier et valider les exigences précédemment citées afin de s'assurer qu'elle est apte à y répondre.
Par exemple, la disponibilité des données nécessaires à l'apprentissage pour atteindre les performances visées doit être validée ou mentionnée comme une condition nécessaire d'utilisation de la fonctionnalité d'IA.

III.3.4. Les hypothèses de conception émises sur la fonctionnalité d'IA (notamment sur les hypothèses statistiques qui peuvent varier dans le temps) et l'approche retenue pour le choix des types modèles et l'évaluation de la fonctionnalité d'IA doivent être documentées. En cas de modification des hypothèses ou des exigences applicables à la fonctionnalité d'IA, l'entreprise doit s'assurer :

- qu'une analyse d'impact est réalisée afin de s'assurer que les modifications n'ont pas d'impact négatif sur la conformité des exigences relatives à la fonctionnalité d'IA,
- de la correcte information des personnes impliquées dans les processus certifiés.

III.3.5. Une analyse préliminaire de risques à la phase de développement pertinente et adaptée à l'usage de la fonctionnalité d'IA doit permettre d'identifier, évaluer et de documenter les risques associés à son utilisation et leurs potentiels impacts. Cette analyse doit prévoir le cas d'utilisation de données erronées pouvant être dues à des défauts de capteurs, des erreurs de formatage, des bugs dans le système de management des données ou des cyberattaques et porter sur les composants et sous-composants ainsi que sur les interfaces entre composants de la fonctionnalité d'IA. Les différents modes de défaillance de la fonctionnalité d'IA et leurs conséquences doivent être établis afin de permettre à l'utilisateur d'avoir conscience des risques résiduels auxquels il s'expose et qu'il accepte.
Par exemple, une mauvaise recommandation d'un film sur une plateforme de vidéo à la demande par une fonctionnalité d'IA n'a pas les mêmes conséquences qu'une fonctionnalité d'IA copilote proposant un mauvais virage en plein vol. Les impacts peuvent être quantifiés en termes de coût, de sûreté, sécurité, discrimination etc.

III.4) Processus de développement

III.4.1) Généralités

Les éléments d'entrée et sorties à documenter sont *à minima* :

Pour les éléments d'entrée :

- les spécifications définies (cf. III.3.1),
- l'analyse préliminaire de risques (cf. III.3.5),
- les besoins en documentation liée à la fonctionnalité d'IA (cf. III.3.1).

Pour les éléments de sortie :

- la fonctionnalité d'IA dont le domaine d'utilisation, les usages et les performances seront à évaluer,
- la documentation associée (manuel utilisateur, descriptif des modèles, etc.).

III.4.1.1. Le ou les types d'algorithmes ainsi que le type d'apprentissage utilisés par la fonctionnalité d'IA doivent être documentés au regard des contraintes de performance, de maintenance et d'explicabilité.

Les types d'algorithmes peuvent par exemple être neural networks, SVMs, random forest, etc. Les types d'apprentissage peuvent par exemple être apprentissage supervisé, non supervisé, par renforcement, etc.

III.4.1.2. Les contraintes éventuelles des ressources matérielles sur lesquelles l'apprentissage de la fonctionnalité d'IA pourra être réalisé doivent être documentées.

Par exemple le fait qu'un entraînement doive être réalisé sur le cloud ou en local sur le hardware du développeur ou sur le site de déploiement doit être documenté.

III.4.1.3. Si la fonctionnalité est déployée sur l'infrastructure du client, l'infrastructure (matérielle, système d'exploitation et logicielle), les types de déploiement (cloud public ou privé, on-premise etc.) supportés par la fonctionnalité d'IA et la dépendance à des technologies d'IA sous-jacentes doivent être documentés et communiqués au client selon les modalités prévues au § III.7.4.

Par exemple, la solution repose sur une application J2EE ou node.js et est compatible avec ou sans docker, sur AWS EC2 et GCP. Les modules X, Y et Z de NLP sont supportés et peuvent être intégrés de façon alternative à la fonctionnalité.

III.4.1.4. Les interfaces nécessaires à l'usage de la fonctionnalité d'IA doivent être documentées et communiquées au client selon les modalités prévues au § III.7.4.

Il peut par exemple s'agir des interfaces API (Application Programming Interface) ou des ICD (Interface control document).

III.4.1.5 Les caractéristiques du domaine d'utilisation visé, notamment ceux qui ont une influence sur les performances de la fonctionnalité d'IA, doivent être documentées. Pour chaque facteur d'influence analysé, la justification de sa pertinence ou de son exclusion vis-à-vis du domaine d'utilisation doit être documentée et communiquée au client selon les modalités prévues au § III.7.4.

Par exemple, pour de la reconnaissance automatique par caméra de panneaux routiers, une justification du choix de la luminosité ambiante comme facteur d'influence pourrait être qu'il a été constaté lors d'essais que la performance du système variait en fonction de la luminosité et qu'elle pouvait être approximée par une loi de

type y=f(x). A contrario, si la température n'a pas été retenue, ceci peut être justifié par des résultats d'essais montrant qu'elle n'avait pas d'influence.
Exemples d'autres facteurs d'influence pour la reconnaissance d'images : résolution image, contraste, distance de prise de vue etc.

III.4.1.6. Les contre-indications[20] doivent être documentées et communiquées au client selon les modalités prévues au § III.7.4.
Par exemple, si les tests ont démontré que l'algorithme de reconnaissance faciale ne fonctionnait pas pour les personnes portant des chapeaux, le port d'un chapeau constitue une "contre-indication".

III.4.1.7. Les non-indications[21] connues doivent être documentées et communiquées au client selon les modalités prévues au § III.7.4.
Par exemple, pour de la reconnaissance automatique par caméra de panneaux routiers une non-indication pourrait être l'utilisation de nuit sans éclairage public.
Autre exemple : si le domaine d'utilisation précise que l'algorithme de reconnaissance faciale a été conçu pour reconnaître des personnes sans chapeau, le port d'un chapeau constitue une "non-indication" (donnée d'entrée en dehors du domaine d'utilisation).

III.4.1.8 L'architecture globale du code source du projet ainsi que l'architecture réseau sous-jacente à la fonctionnalité d'IA et notamment les flux en entrée et sortie doivent être documentées. L'architecture réseau et les flux d'entrée/sortie doivent être communiqués au client selon les modalités prévues au § III.7.4.
Par exemple, un fichier de type "Readme" peut décrire l'objectif du projet, ainsi que l'arborescence des dossiers et fichiers du projet (par exemple un dossier pour chacune des thématiques suivantes : acquisition des données, analyse des données, modélisation des données, création des modèles avec un dossier contenant les fichiers notebooks et un dossier contenant les fichiers de code), quels sont les flux réseau (réseau d'entreprise, internet, cloud, etc.) à ouvrir et sécuriser (confidentialité, secret défense, etc.) en entrée et sortie de la fonctionnalité d'IA.

III.4.2) Maîtrise de la qualité des données

III.4.2.1 Exigences communes aux bases d'apprentissage et de test

III.4.2.1.1. La collecte et l'utilisation de données d'apprentissage, de validation et de test doivent être conformes à la réglementation en vigueur.
Par exemple en Europe, la conformité au RGPD peut être assurée par un processus de pseudonymisation/anonymisation des données tel que recommandé par la CNIL ou l'ENISA dans ses recommandations sur l'usage des technologies conformément aux dispositions en matière de protection des données et de respect de la vie privée sur les techniques et meilleures pratiques de pseudonymisation.

III.4.2.1.2. Les données utilisées pour les tests doivent être distinctes des données utilisées pour l'apprentissage (entraînement et validation/développement). Les modalités de segmentation[22] (aléatoire, par date, par configuration, etc.) des jeux de données d'entraînement, de validation et de test doivent être documentées.

[20] Contre-indication : Scénario ayant fait l'objet d'un test et ayant conduit à une non performance.
[21] Non-indication : Scénario n'étant pas compris dans le domaine d'utilisation prévu de la fonctionnalité d'IA.
[22] Segmentation de données : Division d'un corpus de données en plusieurs bases (par exemple d'apprentissage et de test), soit à partir de critères objectifs (date, âge, etc.) soit de manière aléatoire.

Par exemple, pour de la prédiction du risque de cancer chez des personnes, une modalité de segmentation pourrait être que tous les dossiers des patients dont la date de naissance est entre le 1er janvier est le 30 mars sont utilisés pour la phase de test.
Par exemple, la validation croisée, par blocs, leave-one-out sont des stratégies de séparation.

III.4.2.1.3. Afin d'assurer la traçabilité des actions réalisées, les identités (ou à défaut des identifiants uniques), les rôles et responsabilités des personnes impliquées dans la conception des bases de données d'apprentissage et de test doivent être documentées. Les moyens d'accès au système d'information (SI) et notamment aux bases de données d'apprentissage et de test doivent être sécurisés. Les droits d'accès doivent être nominatifs, restreints aux seules personnes habilitées et revus à intervalles réguliers.
Par exemple, le processus d'habilitation d'accès peut prévoir la signature d'accords de non-divulgation et un contrôle des antécédents judiciaires. La revue des droits peut s'effectuer mensuellement ou à chaque arrivée/départ d'un collaborateur. La sécurisation du SI peut reposer sur la mise en place d'un SMSI (système de management de la sécurité de l'information) basé sur l'ISO 27001.

III.4.2.1.4. Si l'entreprise développant la fonctionnalité d'IA est en charge de la collecte des données, une sécurisation applicative de la collecte de données doit être mise en œuvre et documentée pour limiter le risque d'attaque par empoisonnement de données[23].
La sécurisation des applicatifs de collecte des données peut par exemple être assurée par des campagnes de tests d'intrusion et l'application de mesures de sécurité correctives.

III.4.2.1.5. Les données d'apprentissage et de test doivent être saines (non empoisonnées) et uniques[24]. Les méthodes utilisées pour s'assurer que les bases sont saines et ne contiennent pas de doublon doivent être documentées.
Par exemple, la garantie sur la fiabilité des données peut être apportée par le fait que les données ont été produites par l'entreprise développant la fonctionnalité d'IA et stockées sur un serveur sécurisé dont les accès sont restreints.
Par exemple, pour garantir l'absence de doublon dans la base d'apprentissage, chaque donnée peut être identifiée par un identifiant unique.

III.4.2.1.6. La qualité et la traçabilité des annotations[25] doivent être garanties. La méthode d'annotation à suivre par les annotateurs afin de garantir l'homogénéité des annotations et limiter la subjectivité de ces dernières doit être documentée. La méthode mise en œuvre pour assurer et mesurer la qualité des annotations des données doit être documentée.
Par exemple, cela peut être fait via une multi-annotation ou par une revue régulière des annotations.
Pour de la reconnaissance de plantes sur des images, la méthode d'annotation peut stipuler que chaque image est annotée par trois annotateurs et qu'une comparaison inter-annotateurs (afin d'analyser les écarts entre les annotations de chaque expert) ainsi qu'une comparaison intra-annotateur (pour mesurer l'homogénéité des annotations à travers un jeu de données de chaque annotateur) sont réalisées.

[23] Empoisonnement de données (Poisoning) : Attaque visant à manipuler les données d'apprentissage afin de contrôler le comportement de prédiction d'un modèle formé, de sorte que le modèle étiquette les exemples malveillants dans les classes souhaitées (par exemple, en étiquetant les courriels de spam comme sûrs).
[24] C'est-à-dire l'absence de copie de données dans la base
[25] Tâche qui consiste à attribuer à chaque donnée le label qui lui correspond. Par exemple, à attribuer le label « chien » ou « chat » à une base de photographies d'animaux. Ou encore à attribuer le label correct entre « nom », « prénom », « adresse », « date » ou « aucun » à chacun des mots d'un document. On distinguera l'annotation manuelle, lorsque cette tâche est effectuée par un humain, de l'annotation automatique, lorsque cette tâche est effectuée par un programme informatique.

Les compétences des annotateurs ne se limitent pas aux compétences techniques liées aux outils d'annotation mais également aux compétences/connaissances métier liées à l'usage de la fonctionnalité d'IA.

III.4.2.1.7. La collecte des données et leur annotation doivent être assurées par des outils adaptés et des personnels compétents au regard de la fonctionnalité d'IA développée (notamment la tache automatisée) mais également de son domaine d'utilisation et notamment de ses règles métier.

Par exemple, s'il est jugé que certaines tâches ne nécessitent pas de compétence particulière, il doit être possible de démontrer la facilité d'implémentation de ces tâches au sein du processus d'apprentissage par des non-experts.
Il peut également être défini que les personnes en charge du processus d'apprentissage doivent a minima comprendre le domaine d'application et l'utilisation qui sera faite de la fonctionnalité d'IA. Il est possible de se baser sur des entretiens, des QCM ou des tests pour évaluer cette compréhension.
Par exemple, des annotations de cancers sur des radios ne peuvent pas être réalisées par des personnes qui ne sont pas compétentes en cancérologie.

III.4.2.1.8. Dans le cas d'apprentissage sur des données structurées, une méthode efficace de gestion (détection et traitement) des valeurs manquantes parmi les attributs critiques des données d'apprentissage et de test doit être mise en œuvre et documentée.

Par exemple, pour du diagnostic automatique à partir de dossiers de patients, une analyse de criticité a permis d'identifier l'attribut « dossier patient » comme attribut critique pour la fonctionnalité d'IA, c'est-à-dire que son absence peut fausser de manière significative le résultat.
Ce caractère critique peut être défini dans le cahier des charges (ou guide d'annotation) ou par une analyse de la criticité.
L'absence de champ vide parmi les champs des dossiers patients utilisés est garantie car un algorithme de détection automatique de champ vide sur la plateforme numérique de saisie du dossier patient est présent, avec un message d'erreur, le cas échéant.
Par exemple pour une application servant à identifier des plantes, on peut penser que l'absence de l'attribut nom de la plante pour certaines données est critique alors que l'absence de l'horaire de prise de la photo ne l'est pas.

III.4.2.1.9. Une méthode efficace de gestion (détection et traitement) des données d'apprentissage, de test ou d'utilisation avec des valeurs erronées ou situées en dehors du champ d'application du modèle doit être mise en place et documentée.
Les données aberrantes[26] doivent être distinguées des données atypiques[27] ou en dehors du domaine de fonctionnement. À défaut, la justification de la suffisance du test et/ou la limite du test dans le contexte d'utilisation de la fonctionnalité d'IA doit être documentée.

Par exemple, des intervalles dans lesquels les valeurs de champs doivent être comprises pour être considérées comme viables peuvent être définis. Les anomalies peuvent par exemple être gérées par "outliers".
Pour une prédiction du cours de bourse, la garantie qu'aucune donnée de la base d'apprentissage n'est erronée peut être apportée par des analyses statistiques sur les données qui sont réalisées et permettent d'éliminer les données aberrantes, c'est-à-dire à l'extérieur de l'intervalle de confiance.
Les valeurs erronées ou en dehors du champ d'application du modèle comprennent les problèmes liés au format des données d'entrée. Il est possible pour garantir que les données de la base d'apprentissage soient bien dans un format référencé dans la liste des données d'entrée, de mettre en place un algorithme de vérification automatique de format, avec un message d'erreur, le cas échéant.

[26] Donnée aberrante : Donnée impossible en mode de fonctionnement nominal.
[27] Donnée atypique : Donnée représentant un évènement rare.

III.4.2.1.10. Si des exigences en confidentialité sont énoncées (cf. III.3.1), la présence de données personnelles et/ou sensibles, notamment à potentiel discriminatoire, doit être vérifiée et ces données doivent être marquées pour traitement ultérieur à l'apprentissage.
Par exemple dans le cas où le RGPD est applicable, ces données devront faire l'objet d'une suppression ou d'une anonymisation/pseudonymisation suivant par exemples les recommandations de la CNIL ou de l'ENISA dans ses recommandations sur l'usage des technologies conformément aux dispositions en matière de protection des données et de respect de la vie privée sur les techniques et meilleures pratiques de pseudonymisation.

III.4.2.2 Exigences spécifiques à la base d'apprentissage

III.4.2.2.1. La base de données d'apprentissage doit assurer un taux de couverture des cas d'usage du système suffisant au regard des exigences de performance et des risques associés à l'usage de la fonctionnalité d'IA, limiter les biais de sélection et doit être équilibrée vis-à-vis des cas d'usage et des objets sous étude. Les cas couverts par la base et leur distribution doivent être documentés. La méthode d'estimation de la distribution réelle des données ou des clusters de données dont la base d'apprentissage constitue un échantillon doit être documentée. La méthode pour assurer le rééquilibrage de la base d'apprentissage (initiale ou suite à un rééchantillonnage) en fonction des cas d'usage et des objets sous étude doit être documentée.
Par exemple pour le cas de la navigation d'un véhicule autonome, le cas "insertion sur l'autoroute" apparaît 107 fois dans la base ce qui représente 1% des cas représentés. Cette distribution est donc en équivalence (statistique) avec la distribution opérationnelle qui a été estimée grâce à la collecte de données terrain qu'on peut remettre en perspective avec la proportion réelle de ce cas d'usage en conduite sur autoroute.
Si un cas se produit 1% du temps mais entraine systématiquement la mort de la personne si le système agit mal, le système doit être plus performant dans ce cas-là et ce cas doit se refléter dans la distribution des données.
Par exemple les techniques d'équilibrage statistique (data upsampling, weighting of the loss, etc.) des catégories doivent être décrites.
Par exemple via stratification sampling comme les SMOTE (synthetic minority oversampling techniques) ou sous-échantillonnage de la classe majoritaire.
La documentation de l'équilibrage et de la couverture des jeux de données peut être fournie par des outils de visualisation de données.

III.4.2.2.2. La base de données d'apprentissage doit intégrer les événements rares ou à défaut la fonctionnalité d'IA doit être capable de détecter l'occurrence de tels événements et, le cas échéant, de signaler l'invalidité potentielle des résultats à l'utilisateur. L'entreprise doit documenter la liste des évènements rares[28] intégrés à la base de données ainsi que leur fréquence d'apparition et la méthode utilisée pour les déterminer.
Pour une fonctionnalité de reconnaissance des usagers de la route par caméra, la reconnaissance de personnes marchant à côté de leur vélo est identifiée comme cas rare représentant 0,001% des cas soit 107 occurrences dans la base d'apprentissage.
Par exemple, la base de données d'apprentissage pourrait ne pas intégrer de cas où la température extérieure à Grenoble est supérieure à 50°C, mais la fonctionnalité doit dans ce cas prévoir un avertissement explicite pour l'utilisateur si le cas se présente.

[28] Evènement rare : Événement de risque non nul sur un temps infini, mais qui n'est observable que quelques fois sur un grand nombre d'observations.

Les méthodes mises en place pour augmenter les bases de données (ajouts de bruits, de transformations métamorphiques, etc.) et générer des corner cases (GAN[29], VAE[30], etc.), des données antagonistes (adversariales), peuvent être décrites.

III.4.2.2.3. Les données d'apprentissage doivent être précises, c'est-à-dire avec une incertitude limitée, et fidèles à ce qu'elles représentent dans la réalité. Les méthodes mises en œuvre pour calculer cette incertitude et garantir cette fidélité doivent être documentées.

III.4.2.2.4. La quantité de données d'apprentissage doit être suffisante au regard des exigences de performance et des risques associés à l'usage de la fonctionnalité d'IA et ceci doit être vérifié par une méthode documentée. Tout processus automatique générant de la donnée d'apprentissage complémentaire doit être documenté.

La quantité de données peut dépendre de la complexité du cas d'usage, ainsi que le type d'algorithme utilisé mais l'objectif est d'avoir une quantité de données suffisante pour converger vers le résultat souhaité. La collecte devra donc se faire sur le plus de données possible par l'intermédiaire des sources internes et externes en s'assurant d'éviter le sur-apprentissage.

Par exemple, dans le cas d'une application non-critique, qui pourrait être boot-strappée avec peu de données, la méthode permettant d'évaluer la qualité des modèles ou de leur application doit être documentée, ainsi que le processus permettant d'améliorer la base de données d'apprentissage et les modèles appris au cours du temps.

Par exemple, les processus comme les rotations d'images visant à produire des données d'apprentissage doivent être documentés.

III.4.2.3 Exigences spécifiques à la base de test

III.4.2.3.1. La base de données de test doit assurer un taux de couverture des cas d'usage du système suffisant au regard des exigences de performance et des risques associés à l'usage de la fonctionnalité d'IA. Les cas couverts par la base de test et leur distribution doivent être documentés.

Par exemple, pour la navigation d'un véhicule autonome, si l'on considère parmi le descriptif des scénarios couverts la condition météorologique "neige", le nombre d'apparition dans la base est 107 ce qui fait une fréquence de 1%, en équivalence (statistique) avec la distribution opérationnelle qui a été estimé grâce à la collecte de données terrain.

III.4.2.3.2. La base de données de test doit intégrer des événements rares. L'entreprise doit documenter la liste des évènements rares intégrés à la base de données ainsi que leur fréquence d'apparition et la méthode utilisée pour les déterminer.

Par exemple, pour une fonctionnalité de reconnaissance des usagers de la route par caméra, la reconnaissance de personnes marchant à côté de leur vélo est identifiée comme cas rare représentant 1% des cas soit 107 occurrences dans la base d'apprentissage.

Les méthodes mises en place pour augmenter les bases de données (ajouts de bruits, de transformations métamorphiques, etc.) et générer des corner cases réalistes (GAN, VAE, etc.), des données antagonistes (adversariales), peuvent être décrites.

[29] GAN (ou RAG, Réseau antagoniste génératif): Un GAN est un modèle génératif constitué de deux réseaux : un générateur en charge de produire une donnée, et un discriminateur responsable de déterminer si cette donnée est réelle ou issue du générateur.

[30] VAE : Le VAE (variational autoencoder) est un modèle génératif, comme le GAN, qui repose sur une technique d'apprentissage non supervisé impliquant deux modèles couplés (le codeur pour la reconnaissance et le décodeur pour la génération) capables de former des représentations de données tout en ignorant les bruits associés aux données.

III.4.2.3.3. Les données brutes de la base de test doivent avoir été générées par des capteurs similaires (qualité, etc.) à ceux du système déployé ou une étude doit avoir été réalisée pour évaluer le biais introduit par l'utilisation de capteurs différents et une méthode doit avoir été mise en place pour corriger (transformations artificielles appliquées aux données d'apprentissage, etc.) et/ou diminuer l'impact de ce biais.
Les données brutes de la base de test doivent faire l'objet de prétraitements similaires à ceux dont sont l'objet les données d'entrée de la fonctionnalité d'IA une fois le système déployé.
Par exemple, dans le cas de la reconnaissance de panneaux routiers sur des photos, les images de la base de données de test doivent être produites par le même appareil photo que celui utilisé en conditions réelles ou une analyse de biais doit être réalisée. Par exemple, si les données issues de la base de données ouverte "German Traffic Sign Recognition Benchmark" sont utilisées, un corpus de test composé d'images générées par l'appareil photo du système déployé peut être généré et les performances comparées à celles obtenues sur la base de données de test issues de la GTSRB, mais isolée de la base de données d'apprentissage.

III.4.2.3.4. Toutes les étapes de transformation des données doivent être documentées. A défaut, une étude doit avoir été réalisée pour évaluer le biais introduit par la réalisation des prétraitements sur la base de données de test et les données en entrée de la fonctionnalité d'IA du système déployé et/ou diminuer l'impact de ce biais.
Le processus de normalisation des données est un exemple de prétraitement à lister.

III.4.3) Maîtrise du processus d'apprentissage

Les éléments d'entrée et sorties du processus d'apprentissage qui doivent être identifiés et documentés sont à minima :

Pour les éléments d'entrée :

- les spécifications définies (cf. III.3.1),
- l'analyse préliminaire de risques (cf. III.3.5),
- le descriptif des données,
- le descriptif des méthodes mises en œuvre pour la collecte et l'annotation des données,
- le descriptif des étapes de modification/traitement sur les données dans le cadre du processus d'apprentissage.

Pour les éléments de sortie :

- modèles développés,
- la fonctionnalité d'IA dont le domaine d'utilisation, les usages et les performances seront à évaluer,
- la documentation associée (manuel utilisateur, descriptif des modèles, etc.).

III.4.3.1. Une méthode doit être mis en œuvre et documentée pour garantir la qualité du processus d'apprentissage initial. Elle doit permettre :

- d'éviter le sous[31] ou sur-apprentissage[32]
- de minimiser les erreurs au centre ou en queue de distribution.

[31] Sous-apprentissage : L'apprentissage n'a pas été suffisant pour permettre à la fonctionnalité d'IA de modéliser de manière satisfaisante les relations sous-jacentes des données.
[32] Sur-apprentissage : Le modèle développé correspond si étroitement aux données d'apprentissage qu'il ne parvient pas à faire des prédictions correctes sur de nouvelles données.

Cette méthode devra donc inclure :

- des méthodes de contrôle mises en place pour remédier aux risques de sous ou sur-apprentissage et le cas échéant un descriptif des méthodes mises en œuvre si un sous ou un sur-apprentissage est détecté,
- une justification de la pertinence (choix des seuils, des indicateurs, etc.) et de l'efficacité de ces méthodes de contrôle est à présenter à partir des outils de mesure associés et des enregistrements des résultats,
- la justification des fonctions de perte sélectionnées ou de la ventilation du jeu de test effectuée pour tester sur des sous-parties.

Par exemple, pour éviter le sur-apprentissage : les méthodes suivantes peuvent être mises en œuvre : hold-out (séparation bases de données d'apprentissage et de test), validation croisée, augmentation de données (transformation métamorphiques, perturbations, etc.), régularisation L1/L2, dropout, etc.
Exemples de type de fonctions de perte : Empirical Risk Minimisation, Structural Risk Minimisation, etc.
Exemples de méthodes de convergence : descente de gradient, monitoring du loss entre le batch d'entraînement et le batch de validation (lorsque la performance diverge entre les deux batchs, on peut dire qu'on atteint un optimum local), techniques d'early stopping, etc.

III.4.3.2. De manière à assurer la reproductibilité du développement des modèles, l'entreprise doit tracer, archiver et documenter tous les changements apportés aux outils de développement, ceci sur une durée adaptée au contexte de l'application et à la durée de vie des modèles.
Par exemple, l'entreprise doit archiver la liste des librairies logicielles utilisées et leur version et le suivi des modifications du code source des outils de développement. La présence d'un document ou d'un enregistrement dans un outil interne spécifiant les prérequis des noms de librairies nécessaires et leurs versions est acceptable.

III.4.3.3. Des moyens d'archivage des versions successives de la fonctionnalité d'IA, des modèles et données, adaptés au contexte (notamment réglementaire ou de capacité de stockage), permettant de rejouer des situations lorsque nécessaire doivent être mis en œuvre. La durée d'archivage et les méthodes d'archivage (échantillonnage, Time-to-live pour la conservation de données, etc.) doivent être clairement documentées (notamment au regard des capacités de stockage), ainsi que la procédure permettant de récupérer les éléments nécessaires pour rejouer une situation.
Par exemple, la version d'une fonctionnalité d'IA peut être déterminée par une version du code source avec les valeurs des paramètres, ou par les hyperparamètres couplés aux données d'apprentissage.
La traçabilité des différentes versions peut être apportée par un identifiant unique garantissant l'intégrité et l'unicité de la version (hashage du couple code source/paramètres par exemple).
Les versions de la fonctionnalité d'IA et des modèles peuvent être archivés via des outils comme Git ou SVN. Les différentes versions des données peuvent être stockées sur des bases distinctes ou leurs versions taguées dans la base.

III.4.3.4. Afin d'assurer la traçabilité des décisions, les prédictions doivent porter l'information d'identifiant unique du modèle utilisé pour réaliser la prédiction/décision ou des informations permettant de retrouver cet identifiant.
Par exemple, un numéro unique attaché à chaque modèle peut être intégré aux métadonnées de la prédiction.

III.4.3.5. Si des besoins d'explicabilité sont définis, la chaîne de causalité menant à la décision finale de l'algorithme doit pouvoir être tracée. Le chemin de décision (data, modèles, statistiques) utilisé pour chaque modèle doit pouvoir être retrouvé.

Par exemple, dans le cas d'un système d'IA reposant sur des algorithmes de type Random Forest, on doit être capable de retrouver quel(s) arbre(s) de décision est /sont à l'origine du résultat.

III.4.3.6. Si des besoins de confidentialité précédemment définis (cf. III.3.1) l'imposent, le processus d'apprentissage doit empêcher de remonter aux données d'origine, notamment en cas d'utilisation d'apprentissage distribué (federated learning)[33].
Cela peut par exemple être démontré par l'absence de données en clair ou l'impossibilité de retrouver les données d'origine (données sensibles, identité des personnes, etc.) par recoupement d'information.

III.4.3.7. Si le processus d'apprentissage peut engendrer des biais, des algorithmes de débiaisage en phase de prétraitement, traitement ou post-traitement doivent être utilisés et cette utilisation doit être documentée.

III.4.3.8. Le choix et la stratégie d'adaptation de la valeur des hyperparamètres en fonction des résultats atteints doivent être justifiés et documentés.
Par exemple : le choix de réitérer ou non le processus de recherche des hyperparamètres est justifié par l'atteinte de résultats correspondants aux besoins métiers spécifiés. L'adaptation des hyperparamètres peut être faite via un apprentissage sur des données complémentaires ou l'on peut décider d'utiliser des méthodes de remédiation des biais.

[33] Federated learning (apprentissage fédéré) : Technique d'apprentissage automatique réalisée sur des corpus de données distribués sur plusieurs périphériques ou serveurs décentralisés.

III.5) Processus d'évaluation

Les éléments d'entrée et sorties du processus d'évaluation qui doivent être identifiés et documentés sont à minima :

Pour les éléments d'entrée :

- les spécifications fonctionnelles,
- les données de test respectant les exigences définies (notamment III.4.2),
- le prototype de fonctionnalité d'IA dont le domaine d'utilisation et les performances seront à évaluer.

Pour les éléments de sortie :

- les protocoles, outils et métriques d'évaluation,
- les résultats du processus d'évaluation,
- la fonctionnalité d'IA dont le domaine d'utilisation et les performances auront été validés ;
- l'analyse de risque finale ;
- la documentation associée (manuel utilisateur, descriptif des modèles, etc.).

III.5.1. Un protocole d'évaluation se basant sur des mesures de performance doit être mis en œuvre durant et à la fin du processus d'apprentissage et documenté (notamment l'intervention humaine dans les vérifications de performance).
Le choix des métriques d'évaluation des performances, leur mode de calcul et leur implémentation doivent être pertinents, justifiés et documentés.
Par exemple, les métriques utilisées pour mesurer les performances de la fonctionnalité d'IA peuvent reposer sur des calculs de taux d'erreurs, taux de faux positifs, taux de faux négatifs, taux de vrais positifs, taux de vrais négatifs, ROC, rappel/sensibilité (recall), précision (precision), F-mesure (F1 score), justesse (accuracy), indice de Jaccard, ZoneMap, maximum d'espace stable, critères de pertinence, analyses de sensibilité, etc. Dans le cas d'ajout de nouvelles métriques, il doit pouvoir être présenté un compte-rendu des résultats de validation de la métrique.

III.5.2. En adéquation avec le contexte de l'application (usage et analyse de risques), le protocole d'évaluation doit inclure des moyens d'identification de facteurs d'influence sur les performances et de biais[34] potentiels. En particulier, dans le cas où différents sous-groupes ont été identifiés dans les spécifications fonctionnelles (cf. III.3.1) ou dans l'analyse de risques (cf. III.3.5), des évaluations par sous-groupes doivent être réalisées et les mesures de performance par sous-groupes doivent être disponibles.

[34] Inclination au préjugé envers ou contre une personne, un objet ou un point de vue. Par exemple, dans les systèmes d'IA fondés sur les données, tels que ceux produits par apprentissage automatique, des biais présents dans la collecte de données et l'entraînement peuvent être à l'origine de la présence de biais dans le système d'IA. Dans les systèmes d'IA fondée sur la logique, comme les systèmes fondés sur des règles, le biais peut résulter de la manière dont un ingénieur des connaissances envisage les règles s'appliquant à un contexte particulier. Le biais peut également résulter de l'apprentissage en ligne et de l'adaptation par interaction. Il peut également se manifester à travers la personnalisation, par laquelle les utilisateurs reçoivent des recommandations ou des informations correspondant à leurs préférences. Il n'est pas nécessairement le résultat d'un biais humain et de la collecte de données par des êtres humains. Le biais peut, par exemple, se manifester dans les circonstances des contextes limités dans lesquels le système est utilisé, auquel cas il n'est pas possible de le généraliser à d'autres contextes. Un biais peut être positif ou négatif, intentionnel ou involontaire. Dans certains cas, le biais peut entraîner des résultats discriminatoires et/ou injustes.

Par exemple, ces vérifications peuvent s'appliquer à la statistical parity difference, disparate impact ou consistency. Les facteurs d'influence peuvent être découverts par des techniques d'explicabilité/interprétabilité, par exemple par shap, gradcam, methodes formelles.

III.5.3. Le protocole d'évaluation doit inclure une évaluation de l'éventuel sur/sous apprentissage.
Ceci peut être évalué par validation croisée, comparaison des mesures de performances sur bases de test avec celles réalisées sur la base d'apprentissage, etc.

III.5.4. Le protocole d'évaluation doit inclure une évaluation de la résilience et de la robustesse[35] de la fonctionnalité d'IA permettant de caractériser son comportement dans le domaine de fonctionnement normal mais aussi en mode dégradé avec des valeurs fausses, extrêmes, issues de défauts de capteur, de test de résistance ou d'attaques adversariales. Les comportements observés dans les cas de sortie du domaine d'utilisation défini pendant la phase de conception (cf. III.3) doivent être documentés et un résumé doit être communiqué au client selon les modalités prévues au § III.7.4.
Par exemple, pour une fonctionnalité d'IA qui répond au téléphone en relation client, le comportement de la fonctionnalité doit être évalué en fonctionnement normal mais également lorsque le client tient un langage incompréhensible en raison d'une mauvaise réception (robustesse) ou d'un micro défectueux (résilience).
La vérification de la robustesse peut s'appuyer par exemple sur des techniques classiques de validation croisée, de vérification de l'écart par rapport à la marge d'erreur, d'interprétation abstraite, etc.

III.5.5. Le protocole doit garantir la reproductibilité des expérimentations et la répétabilité[36] des mesures de performance mises en œuvre dans le cadre des évaluations. Des analyses de reproductibilité et de répétabilité doivent être réalisées et documentées. Le protocole d'évaluation doit permettre la détection et la traçabilité des cas de non répétabilité des mesures et de non reproductibilité des expérimentations dont un résumé devra être communiqué auprès du client selon les modalités prévues au § III.7.4. La variabilité de la performance doit être en accord avec les marges de reproductibilité et répétabilité définies lors du processus de conception (cf. III.3.1).
Par exemple, pour démontrer la reproductibilité des expérimentations, les résultats des évaluations pourront être comparés avec ceux obtenus par un organisme tiers. Pour démontrer la répétabilité des mesures de performance, plusieurs évaluations pourront être réalisées par un même évaluateur sur les mêmes données de test, et les résultats comparés.

III.5.6. En fonction des besoins spécifiés, de l'analyse préliminaire de risques (cf. III.3.5) établie et de la criticité de la fonctionnalité d'IA, une séparation des rôles entre les activités de développement/entraînement et d'évaluation et validation doit être mise en œuvre de façon appropriée.
Par exemple dans le domaine nucléaire ou de la défense, ou en cas de risque d'atteinte à l'intégrité d'une personne, il peut être jugé nécessaire de séparer totalement les équipes en charge de l'apprentissage et de l'évaluation/validation alors qu'il pourrait être jugé acceptable que des personnes différentes mais d'une même équipe réalisent ces taches pour une fonctionnalité d'IA qui proposerait du contenu sur le web. La représentation

[35] Robustesse : Capacité d'une fonctionnalité d'IA à maintenir sa conformité aux exigences attendues en présence de données d'entrée situées à l'intérieur du domaine d'utilisation prévu.
[36] Répétabilité d'une mesure : Fidélité de mesure selon un ensemble de conditions de répétabilité (réalisée sur un même objet ou des objets similaires dans une courte période de temps, avec une même procédure de mesure, de mêmes opérateurs, un même système de mesure, de mêmes conditions de fonctionnement et un même lieu).

de cette séparation peut être effectuée via des organigrammes, fiches de poste, compte-rendu de tests, logs de personnes sur la plateforme de développement, etc.

III.5.7. Au moins une partie des évaluations doivent être réalisées en conditions réelles de fonctionnement, ou, à défaut, lors des premières mises en œuvre effectives de la fonctionnalité d'IA, selon un protocole approprié. En fonction de l'analyse préliminaire de risques réalisée (cf. III.3.5), un niveau d'écart maximum de performance par rapport aux évaluations en environnement contrôlé[37] doit être assuré et tracé. Le choix des différentes modalités d'évaluation en conditions réelles mises en œuvre doit être documenté.
Par exemple, des tests différents permettant d'évaluer les performances peuvent être simulés numériquement ou réalisés en environnement contrôlés en laboratoire et comparés à ceux réalisés en environnement réel. Concernant les véhicules autonomes, des tests de roulage sur routes ouvertes complètent ainsi les tests sur piste ou simulateur.

III.5.8. Les environnements de test[38] pour l'évaluation (virtuels et réels) doivent avoir fait l'objet d'une qualification[39] qui doit être documentée.
Par exemple la méthode et les résultats d'une comparaison inter-laboratoires peuvent être enregistrés dans un outil de suivi qualité. Si une certification existe pour le type d'environnement de test mis en place, cette dernière pourra servir de qualification.

III.5.9. Les résultats des évaluations (et notamment tout écart aux exigences définies pendant le processus de conception (cf. III.3.1)) doivent être documentés, permettant ainsi de vérifier la conformité de la fonctionnalité d'IA à toutes les exigences définies et communiqués au client selon les modalités prévues au § III.7.4.
Par exemple, des écarts aux spécifications en termes de performance, de robustesse, de sûreté/sécurité, d'éthique, d'explicabilité, d'interprétabilité ou de transparence peuvent être identifiés malgré les procédures mises en place pour répondre aux spécifications mais tous ces écarts doivent être communiqués au client.

III.5.10. La vérification du respect des exigences réglementaires établies pendant le processus de conception (cf. III.3.1) doit permettre de s'assurer que la fonctionnalité d'IA n'est pas validée si une des exigences n'est pas remplie.
Par exemple, le fait que la fonctionnalité d'IA aille à l'encontre d'une réglementation applicable au client doit être identifié et empêcher le déploiement de cette fonctionnalité chez ce dernier.

III.5.11. L'analyse préliminaire de risques établie pendant le processus de conception (cf. III.3.5) doit être mise à jour en tenant en compte des résultats des processus de développement et d'évaluation de la fonctionnalité d'IA, notamment en prenant en compte les changements de comportement aux limites ou en dehors du domaine de fonctionnement défini. Cette analyse de risque finale doit être documentée et les risques résiduels doivent être communiqués au client selon les modalités prévues au § III.7.4.

[37] Environnements contrôlés : Bancs de test physiques situés dans un laboratoire de manière à contrôler les conditions d'expérimentation (température, humidité, etc.).
[38] Environnements de test virtuels : Simulateurs intégrant des modèles numériques de l'environnement et des fonctionnalités d'IA à évaluer.
Environnements de test réels : Lieux d'expérimentation correspondant aux conditions réelles de fonctionnement des fonctionnalités d'IA à évaluer.
[39] Qualification : Évaluation d'un élément en vue d'identifier ses qualités et ses défauts avant son utilisation.

La caractérisation des changements de comportement de la fonctionnalité peut reposer sur des méthodes statistiques reconnues.

III.5.12. Dans le cas d'une entreprise développant et utilisant sa propre fonctionnalité d'IA, un processus de décision préalable à l'autorisation de déploiement, décrivant la méthode et les critères considérés pour aboutir à la décision, au regard des résultats de l'évaluation doit être documenté et mis en œuvre.

III.6) Processus de maintien en conditions opérationnelles

III.6.1) Généralités

Les éléments d'entrée et sorties à documenter sont *à minima* :

Pour les éléments d'entrée :

- les aspects contractuels de maintien en conditions opérationnelles (délais, durée du support, etc.)
- une fonctionnalité d'IA ayant été déployée,
- un contexte d'utilisation ayant potentiellement évolué,
- de nouvelles données d'entraînement,
- une modification des exigences réglementaires applicables,
- les défauts remontés par le client ou le système,
- de nouvelles exigences issus d'activités et/ou usages similaires (normes, règles de l'art, retours d'expérience).

Pour les éléments de sortie :

- une fonctionnalité d'IA, potentiellement ré-entraînée, dont les usages et performances auront été validés et respectant les exigences énoncées,
- la documentation associée.

III.6.1.1. L'entreprise doit prendre en compte pour son processus de MCO :

- les exigences relatives à la fonctionnalité d'IA validées avec l'utilisateur (voir III.3.1),
- les conséquences indésirables potentiellement associées à une défaillance de la fonctionnalité d'IA déployée,
- l'utilisation et la durée de vie prévue de la fonctionnalité d'IA au sein de la solution,
- les retours d'information des utilisateurs et du système,
- la caractérisation du nouveau domaine d'utilisation après chaque mise à jour.

Par exemple dans le cas d'une solution identifiée pour être utilisée pendant plusieurs décennies, l'entreprise devra s'assurer de disposer des compétences internes sur la durée de vie prévue de la fonctionnalité ou devra formellement communiquer à son client la durée du support prévu de la fonctionnalité.

III.6.1.2. La communication avec les utilisateurs doit inclure :

- le traitement des retours d'information des utilisateurs concernant la fonctionnalité d'IA, y compris les réclamations ;
- les informations relatives à tout changement dans le domaine d'utilisation de la fonctionnalité d'IA.

III.6.1.3. Le processus de MCO de la fonctionnalité d'IA doit prévoir les aspects et les besoins en communication du client liés à l'assistance, la formation, la pérennité du modèle, les éventuelles actions préventives et correctives et les évolutions et mises à jour de la fonctionnalité d'IA.

Par exemple, il peut être exigé du client que l'utilisateur soit informé des modifications apportées par le fournisseur à la fonctionnalité d'IA ou au système incluant cette fonctionnalité (si la fonctionnalité d'IA n'est pas cloisonnée) et pouvant affecter sa performance.

L'assistance comprend le dépannage en cas de panne, le débugage, la remontée de fausse alarme ou faux positif pour demande de correction dans une prochaine mise à jour etc.

III.6.1.4. Si l'entreprise développant la fonctionnalité d'IA est en charge du maintien des modèles, les données d'apprentissage et de test doivent être actualisées par rapport à l'environnement d'utilisation qu'elles représentent. La fréquence de mise à jour des données doit être en adéquation avec la réalité et les données doivent être datées si nécessaire.
Par exemple, pour la reconnaissance de panneaux routiers, au début de chaque mois, une vérification est faite sur le site de la sécurité routière sur les nouveaux panneaux créés et ceux supprimés. S'il est constaté une modification dans la liste officielle de panneaux, une campagne d'acquisition d'images pour ce nouveau panneau est lancée pour procéder à un réapprentissage puis à des tests.

III.6.1.5. Les ressources nécessaires et les opérations à réaliser (recommandées ou obligatoires) après déploiement doivent être communiquées au client selon les modalités prévues au § III.7.4.
Par exemple, la nécessité d'utiliser une configuration X pour des modèles de petite taille et une configuration minimum Y pour des gros modèles doit être communiquée. En cas de nécessité de redémarrage hebdomadaire du système, ceci doit être communiqué également.
Par exemple, la consommation CPU/RAM/GPU des autres composants du système peut être surveillée via des moyens de monitoring du hardware (CPU load, GPU load, etc.). Le temps de réponse moyen de l'API ou de mise à jour du modèle peut également être surveillé ainsi que la latence de récupération des données.
Il reste cependant de la responsabilité du client, de mettre en œuvre un processus efficace de surveillance de l'état et des performances des autres composants du système dont dépend la fonctionnalité d'IA.
Par exemple, si c'est à l'utilisateur de régénérer des modèles, les modalités de mise à jour du modèle doivent être documentées et communiquées au client.

III.6.1.6. Si la fonctionnalité d'IA doit répondre à des exigences d'explicabilité et d'interprétabilité, les éléments d'explication et d'interprétation fournis par la fonctionnalité d'IA suite à une prise de décision automatique doivent être justifiés et conservés pour une analyse a posteriori, au regard des exigences, notamment réglementaires et liées à la criticité du système.
Ces éléments doivent être clairement décrits et cette description doit être communiquée au client selon les modalités prévues au § III.7.4.
Par exemple, pour un algorithme d'aide au diagnostic, il peut être défini que les explications renvoyées en langage naturel par la fonctionnalité d'IA sur les raisons qui ont conduit à son diagnostic doivent être conservées pendant 10 ans.

III.6.1.7. Le processus de MCO doit prévoir un mécanisme de contrôle efficace, régulier, et informatif pour l'utilisateur, de l'évolution de la performance. Ce mécanisme doit permettre la détection et la traçabilité de la dégradation[40] et de la dérive[41] des performances de la fonctionnalité d'IA déployée. Ce mécanisme de contrôle doit être documenté, notamment en termes de fréquence et de durée de contrôle et permettre d'identifier les limites du modèle.

[40] Dégradation : Régression des performances d'une fonctionnalité d'IA au cours du temps.
[41] Dérive : changement dans la distribution des données (data drift) ou dans les relations statistiques entre les variables cibles et les autres variables du système (concept drift), qui intervient au cours du temps, de manière instantanée ou progressive, prévisible ou imprévisible. Ces changements peuvent rendre le modèle construit sur d'anciennes données incompatible avec les nouvelles données et une mise à jour régulière du modèle est nécessaire. Souvent, la cause du changement est cachée (contexte caché), inconnue a priori, ce qui peut rendre la tâche d'apprentissage plus compliquée.

Par exemple, la supervision des performances peut être assurée par le suivi de logs, de métriques, via des alertes applicatives standard.
Ce mécanisme de contrôle peut également intégrer des vérifications concernant les propriétés statistiques des éléments d'entrée (distribution notamment) afin de détecter les changements dans le contexte d'utilisation.
Les métriques suivies peuvent être liées au modèle (justesse, rappel, déviation, etc.), techniques (nombre de requêtes, temps de réponse, etc.) ou aux retours utilisateurs (réactions des utilisateurs aux décisions de la fonctionnalité d'IA, validation des prédictions, etc.).
Par exemple en NLP, une extraction journalière d'une partie du corpus peut être réalisée pour évaluation et comparaison au modèle initial.
La dérive des performances suivies peut être due aux données ou à un contexte d'utilisation changeant.
Par exemple, pour des applications critiques, il peut être jugé nécessaire d'intégrer au système d'IA un mécanisme automatique de monitoring ou d'autodiagnostic (interface graphique, script, built-in self-test) permettant de surveiller l'état (fonctionnel) du système.

III.6.1.8. Les erreurs constatées en fonctionnement doivent pouvoir être archivées et récupérées à des fins de mise à jour de la fonctionnalité d'IA le cas échéant.

III.6.1.9. En cas de problème rencontré par un utilisateur, de dérive ou de dégradation des performances observées, une analyse d'impact doit être réalisée afin de déterminer les actions à mettre en œuvre pour y palier. Cette analyse d'impact doit également prendre en compte l'analyse des résultats que l'IA a produit antérieurement à la détection de cette dérive ou dégradation des performances.
Par exemple, lorsque la fonctionnalité d'IA prend une mauvaise décision, un agent peut être capable de le notifier au développeur (à l'aide d'une interface, d'une procédure, etc.) et cette information peut être intégrée dans le processus de maintien en conditions opérationnelles de la fonctionnalité d'IA.
Il peut par exemple être analysé suite à un problème rencontré par un utilisateur que les sorties de la fonctionnalité d'IA étaient imprécises ou fausses depuis un certain temps. Dans ce cas il peut par exemple être décidé que les clients doivent être informés de la remise en cause des résultats fournis antérieurement par la fonctionnalité d'IA.

III.6.1.10. Les évolutions et mises à jour de la fonctionnalité d'IA, ainsi que les évaluations de performance réalisées suite à un apprentissage incrémental ou continu doivent être tracées, archivées et documentées, ceci sur une durée adaptée au contexte de l'application et à la durée de vie de la fonctionnalité d'IA. Cet archivage des versions doit permettre une traçabilité précise de ces versions et l'accès à des versions antérieures dans une mesure adaptée au contexte de l'application.
Par exemple, la version d'une fonctionnalité d'IA peut être déterminée par une version du code source avec les valeurs des paramètres, ou par les hyperparamètres couplés aux données d'apprentissage.
Par exemple, pour l'archivage des versions de code source, des outils comme git ou SVN sont des solutions acceptables.
La traçabilité des différentes versions peut être apportée par un identifiant unique garantissant l'intégrité et l'unicité de la version (hachage SHA256 du couple code source/paramètres par exemple).

III.6.2) Maîtrise de l'apprentissage après déploiement

III.6.2.1. La qualité du processus d'apprentissage après déploiement doit être maîtrisée. Cette maîtrise doit s'appuyer sur une méthode de mise à jour documentée (spécifiant notamment l'intervention humaine dans le processus de mise à jour) et des points de contrôle permettant d'assurer cette maîtrise.
Par exemple, la documentation devra inclure le descriptif de la procédure de mise à jour incrémentale avec les points de contrôle, la fréquence, etc. mais aussi le descriptif du rôle de chaque intervenant dans ce processus

incrémental. L'entreprise pourra enregistrer dans ses outils internes les preuves de réalisation des contrôles définis (par exemple la correcte réalisation des vérifications sur les annotations des données acquises après le déploiement de la fonctionnalité d'IA).

III.6.2.2. Les parties du modèle concernées par une mise à jour doivent être clairement documentées et une analyse d'impact de cette mise à jour doit être réalisée et documentée. Les modalités de communication au client en cas de mise à jour de la fonctionnalité doivent être communiquées au client selon les modalités prévues au § III.7.4.

III.6.2.3. La mise à jour de la fonctionnalité d'IA doit suivre le même processus de développement et d'évaluation qu'un déploiement initial à moins que les différences soient dûment justifiées, documentées et communiquées au client selon les modalités prévues au § III.7.4.
Les performances de l'ancienne et de la nouvelle version de la fonctionnalité IA doivent être comparées avant le déploiement d'une mise à jour, afin d'assurer une non dégradation des performances.
Par exemple, si le réapprentissage peut être lancé par l'utilisateur via une interface graphique, les données manipulées doivent notamment subir les mêmes traitements que lors d'un apprentissage initial (confidentialité, nettoyage des données, etc.).

III.6.2.4. Il doit être possible de revenir à la dernière version fonctionnelle de la fonctionnalité d'IA en cas de dysfonctionnement de la mise à jour.

III.7) Communication avec les clients & fiche produit

III.7.1. L'entreprise doit communiquer au client, selon les modalités prévues au § III.7.4, une description des post-traitements (suggérés ou intégrés) pour l'utilisation de la fonctionnalité d'IA.

Par exemple, le mode de décodage de la prédiction doit être précisé ainsi que la chaîne de post-traitement intégrée et celle suggérée.

III.7.2. L'entreprise doit communiquer au client, selon les modalités prévues au § III.7.4, les procédures d'étalonnage et plans de maintenance des capteurs nécessaires à l'exploitation de la fonctionnalité d'IA.

III.7.3. Dans le cas d'un développement revendiqué comme open-source, l'entreprise doit communiquer au client, selon les modalités prévues au § III.7.4 :

- le type de licence du code applicatif de la fonctionnalité IA et la présence ou non d'options payantes,

Par exemple, la licence peut être une licence GPL, Apache, etc. S'il est indiqué que la fonctionnalité d'IA repose sur une technologie ou un code Open Source, les outils et fonctionnalités en option nécessitant une licence supplémentaire (modèle "freemium") doivent être précisés.

- une description de la façon dont les briques open-source sont utilisées,

Il est par exemple possible de distinguer le code "générique" open-source (ex. brique NLP) du code "métier" spécifique au cas d'usage (implémentation des intentions/réponses métier par exemple).

- une liste des fournisseurs de service indépendants qui maîtrisent la fonctionnalité d'IA présentées comme open Source.

Par exemple, il doit être précisé si la solution n'est maîtrisée que par une entreprise unique, par des ESN, des prestataires de support ou de maintenance de logiciels open Source.

III.7.4. Certaines informations, liées aux processus ou aux fonctionnalité d'IA et visées par les différentes exigences du référentiel comme étant à communiquer au client, doivent être tenues à disposition ou spontanément communiquées aux clients. Les clients sont distingués en 2 catégories :

- client spécifique dans le cas où la fonctionnalité d'IA est développée pour répondre aux besoins d'un client spécifique (cas typique d'une relation client/fournisseur BtoB),
- client générique dans le cas où la fonctionnalité d'IA est développée pour répondre à un besoin identifié et défini par le développeur de la fonctionnalité d'IA pour un ensemble de clients (cas typique d'une solution sur étagère et/ou à destination de clients particuliers).

Si l'entreprise fournit des fonctionnalité d'IA pour des clients spécifiques, elle doit prévoir les moyens adéquats pour tenir à disposition ou communiquer spontanément aux clients ces informations.

Si l'entreprise fournit des fonctionnalité d'IA pour des clients génériques, l'entreprise doit communiquer aux clients une fiche produit récapitulant l'ensemble de ces informations.

La typologie du client (spécifique ou générique) sera vérifiée via la présence d'un contrat ou d'un document équivalent (commande, offre) signé par les deux parties permettant de vérifier le caractère non générique du client.

Si l'entreprise fournit des fonctionnalité d'IA pour des clients spécifiques, elle peut aussi se baser sur une fiche produit récapitulant l'ensemble des informations à communiquer et/ou à tenir à disposition.

Le type de communication (spontanée ou information tenue à disposition pour un client spécifique ou communiquée via la fiche produit pour les clients génériques) est précisé dans le tableau ci-dessous :

exigence	information visée	client **spécifique**		client **générique**
		doit être **tenu à disposition**	doit être **communiqué**	doit être **communiqué via la fiche produit**
III.3.1	les spécifications relatives à la fonctionnalité d'IA et les critères d'acceptation de chacune des exigences	X		X
III.3.1	éléments de communication avec le client définis lors de la phase de conception	selon ce qui a été défini	selon ce qui a été défini	
III.4.1.3	l'infrastructure (matérielle, système d'exploitation, logicielle), les types de déploiement (cloud public ou privé, on-premise etc.) supportés par la fonctionnalité d'IA et la dépendance à des technologies d'IA sous-jacentes		X	X
III.4.1.4	les interfaces nécessaires à l'usage de la fonctionnalité d'IA		X	X
III.4.1.5	les caractéristiques du domaine d'utilisation visé, notamment les principaux facteurs d'influence sur les performances de la fonctionnalité d'IA		X	X
III.4.1.6	les contre-indications		X	X
III.4.1.7	les non-indications connues		X	X
III.4.1.8	l'architecture réseau sous-jacente à la fonctionnalité d'IA et notamment les flux en entrée et sortie	X		X
III.5.4	un résumé des comportements observés dans les cas de sortie du domaine d'utilisation défini pendant la phase de conception		X	X
III.5.5	un résumé des cas de non répétabilité des mesures et de non reproductibilité des expérimentations		X	X
III.5.9	les résultats des évaluations permettant de vérifier la conformité de la fonctionnalité d'IA à toutes les exigences définies pendant le processus de conception	X		X

III.5.11	risques résiduels issus de l'analyse de risque finale mise à jour après évaluation de la fonctionnalité d'IA		X	X
III.6.1.5	les ressources à utiliser et les opérations à réaliser (recommandées ou obligatoires) après déploiement		X	X
III.6.1.6	la description des éléments d'explication et d'interprétation fournis par la fonctionnalité d'IA	X		X
III.6.2.2	les modalités de communication du client en cas de mise à jour de la fonctionnalité	X		X
III.6.2.3	éventuelles différences entre le processus appliqué pour la mise à jour et le processus normal de développement et évaluation		X	X
III.7.1	une description des post-traitements (suggérés ou intégrés) pour l'utilisation de la fonctionnalité d'IA.		X	X
III.7.2	les procédures d'étalonnage et plans de maintenance des capteurs nécessaires à l'exploitation de la fonctionnalité d'IA		X	X
III.7.3	si développement revendiqué comme open-source : - type de licence du code applicatif de la fonctionnalité IA et la présence ou non d'options payantes, - une description de la façon dont les briques open-source sont utilisées, - une liste des fournisseurs de service indépendants qui maîtrisent la fonctionnalité d'IA présentées comme open Source.	X		X

CHAPITRE IV : ENGAGEMENTS DU TITULAIRE DE LA CERTIFICATION

IV.1) Engagements

Les entreprises certifiées sont seules responsables de la conformité de leurs processus, les contrôles de l'organisme de certification ne pouvant se substituer à leurs responsabilités.

Les entités, pour les processus certifiés, s'engagent à :

- réaliser exclusivement des fonctionnalités d'IA en respectant les processus certifiés ;
- mettre en œuvre les changements appropriés en cas de nouvelles exigences ;
- prendre toutes les dispositions nécessaires pour la réalisation des évaluations initiales et de surveillance :
    - transmission du dossier technique et le cas échéant des échantillons nécessaires, accès aux sites, zones, personnels et sous-traitants le cas échéant concernés par l'évaluation,
    - instruction des non conformités formulées dans les rapports d'audit,
    - participation d'observateurs le cas échéant,
- ne communiquer que des informations loyales et sincères ;
- informer sans délai l'organisme de certification des changements pouvant avoir des conséquences sur la conformité du processus ou la validité de la certification émise (changement de statut juridique, modification/mise à jour du processus certifié, etc.).
- conserver un enregistrement de toutes les réclamations dont il a eu connaissance concernant la conformité aux exigences de certification et mettre ces enregistrements à la disposition de l'organisme de certification sur demande, et
    - prendre toute action appropriée en rapport avec ces réclamations et les imperfections constatées dans les produits qui ont des conséquences sur leur conformité aux exigences de la certification;
    - documenter les actions entreprises.

IV.2) Usage de la marque LNE – Intelligence artificielle

Seules les entreprises titulaires de la certification « LNE Intelligence artificielle » pour un ou plusieurs de leurs processus certifiés peuvent utiliser la marque « LNE Intelligence artificielle » sur leurs supports de communication. La marque de certification devra être utilisée conformément à la charte graphique, publiée par le LNE, en vigueur.

Lorsque le titulaire prévoit l'apposition du marquage LNE (marque LNE – Intelligence artificielle), il doit respecter les dispositions destinées à s'assurer du bon usage de la marque :

- ne pas utiliser la certification obtenue d'une manière qui puisse nuire au LNE, ni faire de déclaration ou de communication sur la certification de ses processus qui puisse être considérée comme trompeuse ou non autorisée ;
- toute référence à la certification avant la notification de celle-ci est interdite ;

- en cas de retrait de certification, la référence à cette certification retirée est interdite : tout moyen de communication qui y fait référence doit cesser d'être utilisé ;
- ne faire des déclarations sur la certification qu'en cohérence avec le certificat émis par le LNE ;
- reproduire les certificats dans leur intégralité, avec les annexes le cas échéant, en cas de fourniture à un tiers ;
- toute référence à la certification LNE Intelligence artificielle dans la publicité, la présentation de tout service, ainsi que sur les documents commerciaux de toute nature qui s'y rapportent doit reprendre au minimum les informations suivantes :
    - le numéro du certificat ;
    - l'adresse du site internet du LNE.

Tout usage ou référence abusif de la marque « LNE Intelligence artificielle », qu'il soit l'objet du titulaire du certificat ou d'un tiers, fera l'objet de poursuites en application de la réglementation en vigueur concernant la publicité mensongère et la propriété intellectuelle.

La liste des processus certifiés est disponible sur le site www.lne.fr.

CHAPITRE V : ELABORATION ET VALIDATION DU REFERENTIEL

V.1) Comité de marque

V.1.1) Modalités de fonctionnement

Il est constitué un comité de marque dont les attributions sont de :

- donner un avis sur les règles de certification et ses évolutions,
- donner un avis sur les projets d'actions de communication ou de promotion relatifs à la marque.

Le comité de marque se réunit au minimum une fois tous les 18 mois en réunion ordinaire. Des comités extraordinaires peuvent être organisés chaque fois que nécessaire (par exemple en vue de modifier les règles de certification).

Préalablement à la réunion du comité, le LNE transmet aux membres du comité, un ordre du jour de la séance, accompagné, le cas échéant, des documents associés. Le LNE rédige le compte-rendu des observations et propositions formulées en réunion de comité. Ce compte-rendu est adressé à tous les membres du comité. Le cas échéant, un bureau du comité ou des groupes de travail pourront compléter le dispositif pour gagner en efficacité.

La composition nominative du comité de marque est approuvée par le directeur général du LNE ou son délégataire, chaque membre en étant ensuite informé. Le mandat des membres est de 3 ans, il est renouvelable par tacite reconduction.

L'exercice des fonctions de membre du Comité de marque est strictement personnel. Toutefois, en cas d'absence, un suppléant est désigné et nommé dans les mêmes conditions que le titulaire.

V.1.2) Rôle, engagements et composition du comité

Les membres du comité s'engagent à :

- contribuer de par leur expertise au bon fonctionnement de la marque de certification ;
- conserver la confidentialité des échanges et informations communiqués au cours des réunions du comité de marque et ceci jusqu'à leur publication par le LNE ;
- participer régulièrement aux réunions ;
- contribuer pleinement par leur avis objectif à garantir l'impartialité des avis formulés ;
- contribuer au développement de la marque de certification et promouvoir les prestations certifiées.

Le comité est composé comme suit :

- 3 représentants des clients certifiés ou en cours de certification ;
- 2 représentants des associations ou organismes représentatifs des utilisateurs ou à défaut les utilisateurs eux-mêmes ;
- 1 représentant institutionnel ou académique.

A l'occasion du comité de marque, le LNE recueille l'avis des membres du comité qui participent à l'évolution du référentiel. Le LNE assure le secrétariat du comité.

V.1.3) Groupe de travail

Pour la conduite de certains travaux ponctuels, d'ordre technique et ne nécessitant pas la convocation de l'ensemble des membres du comité de marque, il peut être créé un groupe de travail dont les membres sont désignés nominativement et choisis parmi ceux du comité de marque. Dans le cas d'un groupe de travail, il peut être fait appel à des professionnels ou des personnalités extérieures au comité.

Les missions de ce groupe de travail sont précisées par le comité de marque ; ses attributions seront généralement limitées à l'élaboration de projets, de propositions ou à la fourniture de compléments d'information sur un sujet donné pour le compte du comité de marque.

V.2) Modalités d'élaboration et de validation du référentiel

Le présent référentiel a été élaboré par le LNE, à partir des documents de travail issus des réunions du groupe d'experts et du comité, comprenant les développeurs de fonctionnalités d'IA, les donneurs d'ordres, les utilisateurs.

Pour la validation de ce référentiel, le LNE a la responsabilité :

- d'identifier les parties intéressées concernées ;
- de s'assurer de la pertinence des parties intéressées sélectionnées ;
- de s'assurer de leur représentativité, sans prédominance de l'une d'entre elles ;
- de recueillir leur point de vue.

Sur la base du retour d'expérience, le référentiel est passé en revue au sein d'un comité de marque spécifiquement constitué, intégrant l'ensemble des parties intéressées. Son approbation est effectuée selon la même méthodologie que la première version.

V.3) Modalités de transition entre deux versions du référentiel

En cas de mise à jour du référentiel, des modalités de transition entre la nouvelle version du référentiel et la précédente doivent être définies. Elles doivent fixer la période transitoire, le caractère obligatoire ou non de cette transition, les modalités de vérification de l'application de la transition pour les titulaires de certificats et les modalités de communication associées à cette transition.

Ces modalités de transition doivent être validées par le comité de marque et prendre en compte :

- les entreprises titulaires de certificat(s) ;
- les entreprises en cours de certification parmi lesquelles
    - celles dont le processus de certification a démarré,
    - celles dont le processus n'est pas entamé.

CHAPITRE VI : RECOURS ET TRAITEMENT DES PLAINTES

VI.1) Recours contre décision

Le titulaire ou demandeur de la certification peut contester la décision prise par courrier avec accusé réception.

Dans un premier temps, l'organisme de certification procède au réexamen du dossier au vu des éléments factuels motivant le recours. Il notifie le maintien ou la nouvelle décision au demandeur dans un délai de 15 jours ouvrés à réception du recours.

Dans le cas où le demandeur désire maintenir son recours contre décision, il le notifie à l'organisme de certification par lettre recommandée avec accusé réception, dans un délai de 15 jours ouvrés. Ce recours, non suspensif de la décision de l'organisme de certification, doit être motivé. Il est instruit par l'organisme de certification dans les 21 jours ouvrés suivant sa réception et donne lieu, lorsqu'il concerne la décision de certification, à examen par le comité de lecture. L'organisme de certification informe l'auteur du recours, du maintien ou non de sa décision.

En cas de maintien du recours après instruction et soumission au comité de marque pour avis, le recours est présenté au Comité de Certification et de Préservation de l'Impartialité de l'organisme de certification, qui après examen, propose ses conclusions. La décision finale est notifiée par l'organisme de certification à l'entreprise.

VI.2) Traitement des plaintes

Toute plainte concernant des processus certifiés fait l'objet d'un examen par l'organisme de certification, afin de confirmer si la plainte concerne effectivement des processus certifiés par l'organisme de certification. L'entité formulant une plainte doit étayer celle-ci en fournissant des preuves factuelles.

A réception de celles-ci, l'organisme de certification les examine et le cas échéant contacte l'entreprise concernée.

L'Entreprise concernée doit alors informer l'organisme de certification des suites apportées et tenir à disposition de l'organisme de certification, les enregistrements relatifs à la plainte ainsi qu'aux actions entreprises pour la résoudre. La vérification de la mise en place des actions annoncées peut faire l'objet d'examens supplémentaires à la charge de l'Entreprise.

Dans le cadre du suivi de l'Entreprise, l'organisme de certification examine les enregistrements relatifs aux plaintes et réclamations et vérifie que les corrections et actions correctives appropriées ont été entreprises.

CHAPITRE VII : ANNEXES

VII.1) Lexique

Annotation : Tâche qui consiste à attribuer à chaque donnée le label qui lui correspond. Par exemple, à attribuer le label « chien » ou « chat » à une base de photographies d'animaux. Ou encore à attribuer le label correct entre « nom », « prénom », « adresse », « date » ou « aucun » à chacun des mots d'un document. On distinguera l'annotation manuelle, lorsque cette tâche est effectuée par un humain, de l'annotation automatique, lorsque cette tâche est effectuée par un programme informatique.

Apprentissage automatique : Processus par lequel un algorithme évalue et améliore ses performances sans l''intervention d''un programmeur, en répétant son exécution sur des jeux de données jusqu''à obtenir, de manière régulière, des résultats pertinents.

Apprentissage incrémental : Apprentissage automatique réalisé sur des données regroupées en lots (*batchs*), les lots étant renouvelés périodiquement, au fur et à mesure de l'accumulation de nouvelles données tout au long du cycle de vie de la fonctionnalité d'IA.

Apprentissage profond : Cas particulier de l'apprentissage automatique (Machine Learning) reposant sur l'utilisation d'un algorithme de réseau de neurones à plusieurs couches. Plus le nombre de couches est important, plus l'apprentissage est dit profond.

Apprentissage supervisé : Apprentissage automatique dans lequel l'algorithme s'entraîne à une tâche déterminée en utilisant un jeu de données assorties chacune d'une annotation indiquant le résultat attendu.

Attaque adversariale : Génération ou augmentation d'une donnée d'entrée d'une fonctionnalité d'IA afin de la mettre en difficulté et la tester dans un cas limite.

Attribut critique : Attribut dont l'absence ou l'exactitude peuvent fausser grandement le résultat.

Biais : Inclination au préjugé envers ou contre une personne, un objet ou un point de vue. Par exemple, dans les systèmes d'IA fondés sur les données, tels que ceux produits par apprentissage automatique, des biais présents dans la collecte de données et l'entraînement peuvent être à l'origine de la présence de biais dans le système d'IA. Dans les systèmes d'IA fondée sur la logique, comme les systèmes fondés sur des règles, le biais peut résulter de la manière dont un ingénieur des connaissances envisage les règles s'appliquant à un contexte particulier. Le biais peut également résulter de l'apprentissage en ligne et de l'adaptation par interaction. Il peut également se manifester à travers la personnalisation, par laquelle les utilisateurs reçoivent des recommandations ou des informations correspondant à leurs préférences. Il n'est pas nécessairement le résultat d'un biais humain et de la collecte de

données par des êtres humains. Le biais peut, par exemple, se manifester dans les circonstances des contextes limités dans lesquels le système est utilisé, auquel cas il n'est pas possible de le généraliser à d'autres contextes. Un biais peut être positif ou négatif, intentionnel ou involontaire. Dans certains cas, le biais peut entraîner des résultats discriminatoires et/ou injustes.

Contre-indication : Scénario ayant fait l'objet d'un test et ayant conduit à une non performance.

Corner case : Scénario situé aux limites du domaine d'utilisation effectif du système intelligent, constituant ainsi une situation difficile à traiter par la fonctionnalité d'IA.

Dégradation : Régression des performances d'une fonctionnalité d'IA au cours du temps.

Dérive : changement dans la distribution des données (data drift) ou dans les relations statistiques entre les variables cibles et les autres variables du système (concept drift), qui intervient au cours du temps, de manière instantanée ou progressive, prévisible ou imprévisible.
Ces changements peuvent rendre le modèle construit sur d'anciennes données incompatible avec les nouvelles données et une mise à jour régulière du modèle est nécessaire. Souvent, la cause du changement est cachée (contexte caché), inconnue a priori, ce qui peut rendre la tâche d'apprentissage plus compliquée.

Domaine d'utilisation : Description de l'environnement (conditions météorologiques, population visée, etc.) pour lequel la fonctionnalité d'IA est conçue.

Donnée aberrante : Donnée impossible en mode de fonctionnement nominal.

Donnée atypique : Donnée représentant un évènement rare.

Efficience : Degré selon lequel un système ou composant réalise ses fonctions désignées avec une consommation minimale de ressources.

Empoisonnement de données : Attaque visant à manipuler les données d'apprentissage afin de contrôler le comportement de prédiction d'un modèle formé, de sorte que le modèle étiquette les exemples malveillants dans les classes souhaitées (par exemple, en étiquetant les courriels de spam comme sûrs).

Environnements de test virtuels : Simulateurs intégrant des modèles numériques de l'environnement et des fonctionnalités d'IA à évaluer.

Environnements contrôlés : Bancs de test physiques situés dans un laboratoire de manière à contrôler les conditions d'expérimentation (température, humidité, etc.).

Environnements de test réels : Lieux d'expérimentation correspondant aux conditions réelles de fonctionnement des fonctionnalités d'IA à évaluer.

Evènement rare : Événement de risque non nul sur un temps infini, mais qui n'est observable que quelques fois sur un grand nombre d'observations.

Explicabilité : Capacité d'une fonctionnalité d'IA de rendre compte explicitement des éléments conduisant ou ayant conduit à une évaluation ou à une décision, à partir de données et caractéristiques connues de la situation.

Fonctionnalité d'Intelligence Artificielle : Fonctionnalité dotant un système d'IA de sa capacité d'analyse, de raisonnement ou de décision lui permettant de générer des sorties telles que du contenu, des prédictions, des recommandations ou des décisions influençant l'environnement avec lequel elle interagit.

Federated learning (apprentissage fédéré) : Technique d'apprentissage automatique réalisée sur des corpus de données distribués sur plusieurs périphériques ou serveurs décentralisés.

GAN (ou RAG, Réseau antagoniste génératif) : Un GAN est un modèle génératif constitué de deux réseaux : un générateur en charge de produire une donnée, et un discriminateur responsable de déterminer si cette donnée est réelle ou issue du générateur.

IA statique/figée : Une IA statique est entraînée une seule fois avant d'être déployée. Aucun réapprentissage n'est possible/autorisé tout au long du cycle de vie du produit.

Interprétabilité : Capacité de rendre compréhensible, pour une catégorie d'utilisateurs donnée, le fonctionnement d'un système d'intelligence artificielle.

Non-indication : Scénario n'étant pas compris dans le domaine d'utilisation prévu de la fonctionnalité d'IA.

Performance : Degré selon lequel un système ou composant accomplit ses fonctions désignées avec un ensemble des contraintes données, telles que la vitesse, la précision ou l'utilisation de la mémoire, etc.

Produit : élément de sortie d'un organisme qui peut être produit sans transaction entre l'organisme et le client. Un logiciel est constitué d'informations quel que soit le support de livraison (par exemple programme informatique).

Qualification : Évaluation d'un élément en vue d'identifier ses qualités et ses défauts avant son utilisation.

Répétabilité d'une mesure : Fidélité de mesure selon un ensemble de conditions de répétabilité (réalisée sur un même objet ou des objets similaires dans une courte période de

temps, avec une même procédure de mesure, de mêmes opérateurs, un même système de mesure, de mêmes conditions de fonctionnement et un même lieu).

Représentativité : Qualité d'un échantillon constitué de façon à correspondre à la population dont il est extrait.

Reproductibilité d'une expérimentation : Fidélité d'une expérimentation selon un ensemble de conditions de reproductibilité (réalisée pour un même objet ou des objets similaires dans un ensemble de conditions qui comprennent des lieux, des opérateurs et des systèmes de mesure différents).

Résilience : Capacité d'une fonctionnalité d'IA à maintenir sa conformité aux exigences attendues en présence de données d'entrée extérieures à son domaine d'utilisation (par exemple en cas de panne, d'incident intentionnel ou non et/ou de sollicitation extrême).

Robustesse : Capacité d'une fonctionnalité d'IA à maintenir sa conformité aux exigences attendues en présence de données d'entrée situées à l'intérieur du domaine d'utilisation prévu.

Segmentation de données : Division d'un corpus de données en plusieurs bases (par exemple d'apprentissage et de test), soit à partir de critères objectifs (date, âge, etc.) soit de manière aléatoire.

Sous-apprentissage : L'apprentissage n'a pas été suffisant pour permettre à la fonctionnalité d'IA de modéliser de manière satisfaisante les relations sous-jacentes des données.

Sur-apprentissage : Le modèle développé correspond si étroitement aux données d'apprentissage qu'il ne parvient pas à faire des prédictions correctes sur de nouvelles données.

Système critique : Un système critique est un système dont la panne peut avoir des conséquences graves, notamment pour les biens, les personnes, l'environnement, la survie d'une entreprise, etc. Une liste détaillée de systèmes critiques est proposée dans la proposition de règlement du parlement européen et du conseil établissant des règles harmonisées concernant l'intelligence artificielle publiée le 21/04/2021, Annexe III.

Systèmes d'IA : Un système d'Intelligence Artificielle est un logiciel qui peut, pour un ensemble d'objectifs définis par des humains, générer des sorties telles que du contenu, des prédictions, des recommandations ou des décisions influençant l'environnement avec lequel elles interagissent ; et développé avec une ou plusieurs des approches et techniques suivantes :

1) les approches par apprentissage automatique (ou Machine Learning) incluant l'apprentissage supervisé, non-supervisé ou par renforcement, utilisant une vaste variété de méthodes incluant l'apprentissage profond (ou Deep Learning),

2) les approches logiques et basées sur la connaissance incluant la représentation de connaissances, la programmation logique inductive, les bases de connaissance, les moteurs d'inférence et déductifs, le raisonnement symbolique et les systèmes experts,
3) les approches statistiques, les estimations Bayesiennes, les méthodes de recherche et d'optimisation.

Système expert : Outil déterministe capable de répondre à des questions, en effectuant un raisonnement à partir de faits et de règles connues (s'appuyant sur les connaissances tirées de l'expertise humaine). Le moteur d'inférence du système d'inférence viendra utiliser les faits et règles pour produire de nouveaux faits, jusqu'à parvenir à la réponse à la question experte posée. Cette approche de l'IA ne repose pas sur de l'apprentissage automatique.

Système hybride : Système d'intelligence artificielle intégrant à la fois des techniques d'apprentissage automatique à partir de données et des modèles permettant d''exprimer des contraintes et d'effectuer des raisonnements logiques.

Système mécatronique : Un système mécatronique est un système mécanique enrichi des fonctions offertes par les technologies de l'électronique, de l'automatique et de l'informatique, en particulier les fonctions de captage, de traitement et de communication de l'information, de manière à percevoir son milieu environnant, à communiquer et à agir sur ce milieu.

VAE : Le VAE (variational autoencoder) est un modèle génératif, comme le GAN, qui repose sur une technique d'apprentissage non supervisé impliquant deux modèles couplés (le codeur pour la reconnaissance et le décodeur pour la génération) capables de former des représentations de données tout en ignorant les bruits associés aux données.

lne.fr

CRÉER
LA
CONFIANCE