GOUVERNEMENT
Liberté
Égalité
Fraternité

FRANCE NATION VERTE
Agir · Mobiliser · Accélérer

Charte éthique de l'IA

Pôle Ministériel Ecologie Energie Territoires

Cette charte éthique accompagne la feuille de route « IA et transition écologique » du pôle ministériel, publiée le 28 novembre 2023.

La charte éthique proposée pour le MTECT-MTE s'appuie sur les recommandations émises par le « Groupe d'experts de haut niveau » réuni par la Commission européenne[1]. La France, comme la plupart des Etats membres, les a retenues comme source de référence. Les ministères les adaptent ensuite à leur contexte particulier. Sur le fond, ces recommandations visent à assurer l'explicabilité et l'acceptabilité des outils par les utilisateurs et les usagers. Comme le veulent les recommandations du Groupe d'experts, la distinction n'est pas ici faite entre éthique et rigueur scientifique, mais englobe tous les aspects utiles à la mise en œuvre et concourant à l'acceptabilité.

La charte ministérielle se compose de deux parties : une destinée au chef de projet et l'autre au comité de pilotage. *In fine*, c'est le comité de pilotage qui est responsable du projet et qui doit contrôler sa bonne application. La charte est destinée à aider à la mise en œuvre des projets utilisant de l'IA. Elle a évidemment vocation à être améliorée au fil des retours.

Les questions de la version chef de projet ont pour vocation de l'inciter à se questionner dès la création du projet, y compris sur des aspects délicats étant donné l'état de l'art en *machine learning* et en *deep learning*, notamment la relation entre le résultat d'un algorithme et la prise de décision qui s'en suit. Elles visent à permettre au chef de projet de s'y confronter. Ainsi, ces questions sont à visée interne. La plupart du temps, une réponse par « oui/non » est suffisante. Ces réponses n'ont pas vocation à être diffusées.

Les questions de la version comité de pilotage sont des questions que le comité de pilotage devrait poser au chef de projet. Plus génériques et en principe davantage centrées sur des enjeux de pilotage et d'impact du projet, elles sont moins nombreuses. En toute logique, les réponses du chef de projet devraient être davantage rédigées.

Qui est concerné ?

La charte éthique est un élément de la feuille de route IA et transition écologique ministérielle. Toutes les structures du pôle ministériel sont donc invitées à la mettre en œuvre.

Quand doit-elle s'appliquer ?

Au plus tôt dans le projet, au risque de ne plus pouvoir le redresser par la suite.

---

[1] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

Version Comité de Pilotage

A. Contrôle humain et maîtrise des risques

A.1. Avez-vous réfléchi au niveau de contrôle humain approprié pour votre système d'IA ?

- ⇨ L'IA peut proposer des décisions aux humains. L'humain doit disposer des informations qui lui permettent de juger de la pertinence de la proposition de l'IA. Il s'agit de ne jamais laisser la machine décider seule, sans contrôle possible.

A.2. Avez-vous évalué le fonctionnement de votre système dans des situations ou des environnements imprévus ?

- ⇨ Par exemple, face à des situations auxquelles le système n'a pas été confronté lors de sa phase d'apprentissage, il faut s'assurer que celui-ci ne renvoie pas des résultats au mieux faussés, au pire discriminatoires. Idéalement, il faudrait que le système renvoie un message d'alerte indiquant que les résultats peuvent être inexacts de par cette situation imprévue.

B. Précision

B.1. Comment la précision des résultats est-elle mesurée ?

- ⇨ Les algorithmes d'IA les plus récents basés sur des techniques d'apprentissage profond étant des « boîtes noires », une inconnue subsiste quant à la qualité de leurs livrables. Cette question pose celle de l'explicabilité et donc de l'acceptabilité du produit. Tout élément qui conduit à mieux apprécier la justesse des résultats est donc crucial. Une méthode pragmatique peut être de tester le projet auprès de cercles d'utilisateurs de plus en plus étendus.
- ⇨ Toutefois, au même titre que la qualité des données fait l'objet de contrôles via des mesures de terrain ou par comparaison avec d'autres éléments indépendants, selon qu'elles soient de référence ou non, les traitements par IA devraient faire l'objet des procédures de qualification des résultats comparables et adaptées aux enjeux.

B.2. En cas de prédictions inexactes, avez-vous mis en place une série d'étapes pour résoudre le problème ?

- ⇨ Le *machine learning* est un processus d'apprentissage itératif basé sur une évaluation continue des résultats, en particulier par les utilisateurs. Il est donc normal d'avoir un certain nombre de résultats inexacts au départ qui seront améliorés au fur et à mesure que le modèle rencontre des situations nouvelles. Sans cette évolution continue, le modèle pourra rapidement devenir obsolète en raison des évolutions législatives (évolution des textes) ou sociétales (acceptabilité de nouvelles contraintes ou de nouveaux usages).

C. Pérennité et reproductibilité

C.1. Avez-vous étudié la reproductibilité de votre projet pour d'autres cas d'usage ? Si oui, comment et avec quels résultats ?

- ⇨ Nous manquons de recul sur la reproductibilité des projets d'IA. Conduire le chef de projet à se poser cette question est une condition de la protection des investissements réalisés.

C.2. Que prévoyez-vous pour la maintenance, dans le long terme, de votre modèle d'IA (portabilité, réversibilité) ?

- ⇨ Comme dans tout projet d'exploitation de la donnée, la question de la portabilité (ici, par exemple, changer de cloud) et de la réversibilité (ici, récupérer des données, algorithmes et modèles) d'un projet IA doit être envisagée dès le lancement du projet. L'IA pose des questions encore mal cernées liées aux modèles qui méritent de s'y appesantir.

D. Respect de la vie privée et protection des données

Tout projet de données doit intégrer la protection des données personnelles. Le sujet a pris une telle ampleur, notamment dans les travaux sur le projet de règlement européen sur l'IA, qu'il paraît utile de prévoir un volet spécifique dans cette charte.

D.1. Y a-t-il des données personnelles dans les données d'entrée ? Dans les données de sortie ? Avez-vous consulté le responsable de la protection des données ?

- ⇨ Le RGPD impose la prise en compte de la protection des données personnelles. Il est parfois délicat de différencier des données personnelles d'autres données, d'où l'intérêt de consulter un référent.

D.2. Avez-vous prévu un mécanisme de contrôle pour consigner quand, où, comment, par qui et dans quel but les données ont été consultées ?

- ⇨ En cas de présence de données personnelles et selon leur sensibilité, ce dispositif peut être rendu obligatoire par le RGPD.

D.3. Avez-vous mis sur pied un mécanisme permettant à autrui de signaler des problèmes en rapport avec le respect de la vie privée et la protection des données ? Avez-vous mis en place le droit d'accès et de rectification des données personnelles ?

- ⇨ En cas de présence de données personnelles et selon leur sensibilité, ce dispositif peut être rendu obligatoire par le RGPD.

E. Eviter les biais

E.1. Avez-vous évalué les biais potentiels de votre modèle ? Si oui, quels sont-ils, y en a-t-il de critiques ? Si oui, que prévoyez-vous ?

- ⇨ Les biais constituent des menaces pour la mise en place d'IA sur le plan de l'acceptabilité et de la réputation des organismes. Les biais viennent principalement de sources de données dont la qualité n'a pas été évaluée pour cet usage, notamment sous l'angle social. Ils peuvent venir également de modèles entraînés par ailleurs.
- ⇨ Les biais peuvent apparaître à l'occasion d'un détournement du système par les utilisateurs pour satisfaire un besoin non traité.

F. IA durable, incidence sociale, société et démocratie

F.1. Avez-vous mis en place des mécanismes pour mesurer l'impact environnemental de la mise au point, du déploiement et de l'utilisation du système d'IA ? Avez-vous prévu des mesures pour réduire l'impact environnemental du cycle de vie de votre système d'IA ?

- ⇨ L'évaluation de l'impact environnemental du numérique fait partie des objectifs du ministère.

F.2. Quels sont les impacts possibles sur l'emploi et les compétences des agents destinés à utiliser les systèmes d'IA ?

- ⇨ L'acceptabilité de l'IA par les agents et l'encadrement dépend fortement de la gestion de ces impacts.

F.3. Avez-vous évalué l'incidence plus large de l'utilisation du système d'IA sur la société au-delà de l'utilisateur final, par exemple les parties prenantes susceptibles d'être indirectement concernées ?

- ⇨ Dans le domaine particulier de l'environnement, le citoyen doit être informé (charte de l'environnement adossée à la Constitution). Il est probable que les dispositifs d'IA doivent ainsi être rendus publics lorsqu'ils concernent l'environnement. Par conséquent, l'acceptabilité du déploiement du projet IA pourra dépendre également de la réaction globale de la société.

G. Minimisation des incidences négatives et documentation des arbitrages

G.1. Le projet comporte-t-il des valeurs antagonistes (par exemple sécurité routière et libertés individuelles) ? Si oui, comment sont-elles arbitrées ?

- ⇨ D'habitude, l'être humain gère ces valeurs antagonistes de façon implicite. Pour éviter des biais dommageables, il importe d'évaluer le mieux possible le type d'arbitrage réalisé par le modèle.

Version Chef de projet

A. Contrôle humain

A.1. Les personnes destinées à interagir avec les systèmes d'IA sont-elles informées du fait qu'elles ont une interaction avec un algorithme ou un système d'IA (aide à la prise de décision, conversation avec un robot) ?

⇨ Oui / Non

A.2. Avez-vous mis en place des mécanismes et des mesures pour garantir une supervision humaine, ou pour veiller à ce que les décisions soient prises sous la responsabilité globale d'êtres humains ?

⇨ L'IA peut proposer des décisions aux humains. L'humain doit disposer des informations qui lui permettent de juger de la pertinence de la proposition de l'IA. Il s'agit de ne jamais laisser la machine décider seule, sans contrôle possible.

A.3. Qui sont les personnes qui supervisent ou contrôlent, et à quel moment y a-t-il intervention humaine ou avec quels outils ?

⇨ Je ne sais pas / Question en cours d'analyse / Réponse à la question

B. Maîtrise des risques

B.1. Avez-vous envisagé le niveau de risque posé par le système d'IA dans votre cas d'utilisation spécifique ?

⇨ Oui / Non

B.2. Avez-vous envisagé si, et dans quelle mesure, votre système pourrait avoir un autre usage que celui pour lequel il a été développé ? Si celui-ci pourrait représenter un « mésusage » d'un système d'IA, avez-vous pris des mesures préventives appropriées contre un tel cas de figure ?

⇨ Oui / Non. Il s'agit de limiter a priori les finalités pour lesquelles le système d'IA pourra être utilisé, quitte à prévoir une possible extension de ses finalités a posteriori.

B.3. Avez-vous prévu des mesures ou systèmes pour veiller à l'intégrité et à la résilience du système d'IA face à de potentielles attaques ?

⇨ Oui / Non

B.4. Le système d'IA est-il conforme à des standards de sécurité spécifiques ?

⇨ La normalisation va avancer dans les définitions, à suivre.

B.5. Avez-vous évalué s'il est probable que le système d'IA cause des dommages ou préjudices aux utilisateurs ou à des tiers ?

⇨ Possible / Impossible / Probable / Improbable

B.6. Avez-vous évalué l'incidence probable d'une défaillance de votre système d'IA entraînant la production de résultats erronés ?

⇨ Oui / Non (on demande ici de se poser la question à la fois de la probabilité de l'occurrence des dommages et de l'incidence de l'irruption de ces dommages le cas échéant)

C. Précision

C.1. Avez-vous évalué le niveau de précision et la définition de la précision nécessaires du modèle d'IA dans le contexte du cas d'utilisation concerné ?

⇨ Oui / Non (La définition de la précision correspond à la cible de précision, alors que le niveau de précision attendu correspond à la dispersion des points autour d'un point exact)

C.2. Avez-vous mis en place des mesures pour veiller à ce que les données utilisées soient de bonne qualité et représentatives de l'environnement dans lequel va être déployé le système ?

⇨ Oui / Non

C.3. Avez-vous évalué le préjudice que causeraient des prédictions inexactes du système d'IA ?

⇨ Oui / Non (Des prédictions inexactes correspondent par exemple au fait que le cas de figure n'ait pas été introduit dans le modèle, qui ressortira donc des réponses sans fondement)

C.4. L'utilisateur est-il informé du niveau de précision auquel il peut s'attendre de la part du système ?

⇨ Oui / Non

D. Fiabilité et reproductibilité

D.1. Avez-vous vérifié si des contextes spécifiques ou conditions particulières doivent être pris en compte pour garantir la reproductibilité ?

⇨ Oui / Non (La reproductibilité correspond à la fidélité des résultats d'une même opération ou expérimentation répétée à des moments, en des lieux ou avec des opérateurs différents)

D.2. Que prévoyez-vous pour la maintenance, dans le long terme, de votre modèle d'IA (portabilité, réversibilité) ?

⇨ Par exemple :
- https://anr.fr/fileadmin/documents/2019/ANR-modele-PGD.pdf
- https://opidor.fr/horizon-europe-guide-et-modele-de-dmp/

La deuxième référence concerne la dimension européenne mais aussi un outil de saisie développé par le CNRS/INIST.

D.3. Avez-vous une procédure dédiée dans le cas où le système d'IA renvoie des résultats avec un faible score de confiance ?

⇨ Oui / Non / En cours de création (Un faible score de confiance peut être dû à l'évolution des modèles dans le temps ; les conditions initiales du monde réel étudiées peuvent

avoir changé et il est possible que le modèle fournisse des niveaux avec un score de confiance qui s'effondre)

E. Respect de la vie privée et protection des données

Tout projet de données doit intégrer la protection des données personnelles. Le sujet a pris une telle ampleur, notamment dans les travaux sur le projet de règlement européen sur l'IA, qu'il paraît utile de prévoir un volet spécifique dans cette charte.

E.1. Avez-vous évalué si vos ensembles de données, d'entrée et de sortie, contiennent des données à caractère personnel ?

⇨ Oui / Non

E.2. Avez-vous réfléchi à des manières de mettre au point le système d'IA ou d'entraîner le modèle sans utiliser (ou en utilisant de manière limitée) des données à caractère personnel ?

⇨ Oui / Non

E.3. Avez-vous intégré des mécanismes de notification et de contrôle concernant les données à caractère personnel en fonction du cas d'utilisation (comme un consentement valable et la possibilité de révoquer le consentement, le cas échéant) ?

⇨ Oui / Non

E.4. Avez-vous pris des mesures pour renforcer le respect de la vie privée, par exemple des mesures de chiffrement, d'anonymisation et d'agrégation ?

⇨ Oui / Non

F. Traçabilité et explicabilité

F.1. Avez-vous évalué la mesure dans laquelle la décision du système influence les processus décisionnels de l'organisation ou du métier, dans lequel il a vocation à s'intégrer ?

⇨ Oui / Non

F.2. Pouvez-vous retrouver quelles données ont été utilisées par le système d'IA pour prendre certaines décisions ou faire certaines recommandations ?

⇨ Oui / Non

F.3. Pouvez-vous retrouver quel modèle ou quelles règles ont conduit aux décisions ou recommandations du système d'IA ?

⇨ Oui / Non

G. Eviter les biais

Les biais constituent des menaces pour la mise en place d'IA sur le plan de l'acceptabilité et de la réputation des organismes. Les biais viennent principalement de sources de données dont la qualité n'a pas été évaluée pour cet usage, notamment sous l'angle social.

Dans notre cadre, les biais de données peuvent être temporels (couverture temporelle inadéquate, le plus souvent faute d'alternative) ou géographiques (manque de diversité des zones traitées, pratiques locales spécifiques).

G.1. Avez-vous communiqué aux utilisateurs du système d'IA les risques potentiels que le système a de produire des biais ?

- ⇨ Oui / Non (Les biais d'un système d'IA sont par exemple des biais discriminatoires liés à la représentativité des données d'entraînement.)

G.2. Avez-vous prévu une stratégie ou un ensemble de procédures pour éviter de créer ou de renforcer des biais dans le système d'IA ?

- ⇨ Oui / Non

G.3. Avez-vous prévu un mécanisme permettant à autrui de signaler des problèmes liés aux biais, à la discrimination ou aux mauvaises performances du système d'IA ?

- ⇨ Oui / Non

H. Participation des parties prenantes

H.1. Avez-vous réfléchi à un mécanisme pour inclure la participation de différentes parties prenantes dans la mise au point et l'utilisation du système d'IA ?

- ⇨ Oui / Non

H.2. Avez-vous préparé la voie à l'introduction du système d'IA au sein de l'organisation destinée à l'adopter, en informant et en mobilisant au préalable les personnels concernés et leurs représentants ?

- ⇨ Oui / Non

I. Minimisation et documentation des incidences négatives

I.1. Avez-vous réalisé une analyse des risques ou de l'impact du système d'IA qui tienne compte des différentes parties prenantes qui sont directement et indirectement concernées ?

- ⇨ Oui / Non

I.2. Les tierces parties peuvent-elles signaler de potentielles vulnérabilités, risques ou biais dans le système d'IA ?

- ⇨ Oui / Non