Des critères de base, issus notamment des travaux de la Commission Européenne, sont à évaluer pour les traitements réalisés

1. Contrôles humains sur le traitement IA
2. La robustesse technique et la sécurité du traitement
3. Le respect de la vie privée et la gouvernance des données servant au traitement
4. La transparence autour du traitement
5. La diversité, non-discrimination et équité du traitement
6. L'impact environnemental et sociétal du traitement
7. Les responsabilités associées au traitement

Méthodologie d'évaluation des traitements IA éthiques

L'éthique doit être basée sur des principes partagés et au cœur de l'identité d'un groupe

Elle doit donner lieu à des critères guidant l'action de chacun

Un approfondissement des critères est possible par un débat contradictoire entre les membres du groupe

Objectif -> définir quelques principes et tests clefs auxquels se conformeront les traitements IA

Échelle de notation

0 Critère non pertinent
1 Critère non intégré à la réflexion
2 Critère intégré à la réflexion de manière informelle
3 Formalisation de réponses partielles sans documentation
4 Formalisation et documentation de réponses partielles
5 Formalisation et documentation de réponses complètes

Contrôles humains sur le traitement IA (1/7)

Critère	Aide à l'évaluation	Évaluation
Avez-vous réalisé une analyse d'impact du système IA sur les droits fondamentaux ? Avez-vous défini les modalités d'arbitrages entre droits et les voies de recours ?	Il s'agit ici d'évaluer si le traitement IA contrevient à des principes que toute charte éthique se doit de respecter, par exemple les droits de l'homme. Tout système bien pensé doit permettre un recours et définir des modalités d'arbitrages entre droits	/5
Quelle est la nature du résultat de l'IA (décision, conseil, etc.) ? Avez-vous évalué l'interférence de ce résultat dans le processus de décision humaine ?	Une aide à la décision ou une décision automatisée n'ont pas le même impact, qui doit être évalué pour ses conséquences directes et indirectes, exemple : l'humain peut-il relâcher son jugement ?	/5
Existe-t-il une répartition claire et constructive des tâches entre humains et IA ? Existe-t-il un risque de dépendance excessive au système IA ?	Un traitement est inséré dans un processus plus global, qu'il faut maîtriser et dont il faut évaluer les risques, notamment pour les humains impliqués dans le processus	/5
Avez-vous construit une description formelle des zones et niveaux de contrôles humains dans la chaîne d'opérations où intervient le système IA ? Existe-t-il une responsabilité globale des humains ou bien y a-t-il délégation totale ?	Une chaîne de responsabilité ne peut être établie sans répartition claire des tâches dans le processus entre humains et traitements automatisés. Un traitement ne peut être tenu pour responsable en soi	/5
Quelles mesures ont été mises en place pour faciliter l'audit de la gouvernance et de l'autonomie du système d'IA ?	Il faut distinguer l'auditabilité d'un modèle de celle de sa gouvernance. Un premier pas est la définition du RACI en MCO.	/5

Robustesse technique et sécurité (2/7)

Critère	Aide à l'évaluation	Évaluation
Avez-vous établi les natures des attaques que pourrait subir le système IA (antagonistes, pollution data, infrastructure physique, cyber-attaques, etc.) ? Avez-vous formalisé des solutions selon le type d'attaque ?	L'acculturation aux attaques possibles sur un système IA est importante, tout comme formaliser une checklist. Voir ici : https://blog.f-secure.com/adversarial-attacks-against-ai/	/5
Avez-vous formalisé la liste des doubles usages possibles du système d'IA ? Avez-vous formalisé et/ou mis en place les mesures préventives appropriées ?	Le double usage s'entend ici comme la tendance d'un algorithme à pouvoir servir à la fois des usages bons ou mauvais, notamment via certaines applications militaires.	/5
Quelles mesures ont été mises en place pour veiller à l'intégrité et à la résilience du système IA face aux attaques identifiées ?	Il peut s'agir de mesure techniques ou de processus métiers assurant la résilience du système selon le type d'attaque considéré	/5
Existe-t-il un "Bouton arrêt" ou un mécanisme de suspension du système IA en cas de besoin ou urgence ?	Il faut bien veiller à prévoir une solution de remplacement du traitement IA au cas où il intervient dans un processus métier critique	/5
Avez-vous réalisé une analyse d'impact (technique, opérationnelle, utilisateurs, etc.) d'une défaillance, imprécision ou indisponibilité du système IA ?	Différents niveaux doivent être pris en compte dans l'analyse d'impact : une défaillance peut avoir comme conséquence technique la non production du résultat de sortie, opérationnelle l'interruption d'un processus métier, utilisateur l'indisponibilité d'un service plus ou moins critique, etc.	/5
Avez-vous défini différents niveaux de fonctionnement et les scénarios associés de gouvernance du système ? Avez-vous testé de tels scénarios ?	Il est recommandé de définir a minima un scénario optimal et dégradé de fonctionnement du traitement. La version dégradée peut correspondre à un arrêt du système et le remplacement du traitement dans le processus métier	/5
Avez-vous analysé les risques de dommages ou préjudices, raffinés selon le public concerné ? Avez-vous établi des règles de responsabilité ?	Une bonne cartographie des risques selon les populations se trouve ici : https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/confronting-the-risks-of-artificial-intelligence Les règles de responsabilité peuvent être difficile à établir selon le type de traitement. Il s'agira d'une approche au cas par cas.	/5
Avez-vous défini une politique d'assurance contre les risques recensés pour couvrir les dégâts potentiels provoqués par le système d'IA (personnes, dommages matériels ou financiers, etc.) ?	Une fois la cartographie des risques réalisée, une politique d'assurance peut être mise en place, même si l'absence de chaîne de responsabilité claire dans certains cas empêche l'émergence de solutions claires, exemple : la voiture autonome	/5
Quelle mesure de précision du modèle a été retenue ? Quelles garanties avez-vous autour de la précision du modèle ? Avez-vous bien défini un cadre d'amélioration de cette précision, via une mise à jour des données et réentraînements réguliers par exemple ?	La précision d'un modèle appelle la notion d'intervalle d'incertitude, dont l'acceptabilité dépend du cas d'usage. Or beaucoup de modèles sont développés sans mesure d'incertitude, ce qui est essentiel pour la mise en production et aboutit souvent à des contraintes fortes (phase de tests allongée vs prise de risque sur population plus importante en A/B testing, etc.)	/5
Avez-vous défini une méthodologie pour vérifier la reproductibilité des résultats du modèle ? Existe-t-il un niveau de documentation suffisant pour assurer la facilité de reprise par des tiers ?	Il faut distinguer performance, robustesse (stabilité de performance face à un événement) et reproductibilité. Le niveau le plus fondamental est la reproductibilité, sans laquelle aucune approche scientifique n'est possible. Il ne s'agit pas de relancer l'exécution de l'algorithme, mais de vérifier la stabilité des résultats en environnement de production avec sa dynamique réelle de processus	/5
Avez-vous bien délimité les causes possibles de défaillance du système ? Avez-vous testé ces scénarios de défaillance ? Quel niveau d'information des utilisateurs sur les scénarios de défaillance ?	-	/5

Respect de la vie privée et gouvernance des données (3/7)

Critère	Aide à l'évaluation	Évaluation
Avez-vous précisé les différents types de données utilisées, notamment les données personnelles et/ou sensibles ?	Toujours intégrer à la documentation du projet un dictionnaire des données utilisées, avec un tag pour les variables correspondant à des données personnelles et/ou sensibles	/5
Avez-vous adopté une stratégie de minimisation des données personnelles et/ou sensibles utilisées pour construire le système IA ?	La minimisation des données utilisées est importante tant du point de vue de la sécurité que de la maintenabilité du modèle et donc sa robustesse. N'intégrer que les variables à contribution établie via un incrément de performance	/5
Existe-t-il des mécanismes de notification/validation réversible d'utilisation des données personnelles et/ou sensibles ? Existe-t-il des mécanismes de contrôle et de signalement des problèmes autour des données personnelles et/ou sensibles ?	Il s'agit ici de mécanisme en amont d'un traitement IA en particulier, intégré par exemple dans un processus de gestion des consentements pour l'utilisation des données personnelles.	/5
Avez-vous pensé à des techniques de renforcement du respect de la vie privée (mesures de cryptage, d'anonymisation et d'agrégation) ?	Voir ici pour un ensemble de techniques validées par la CNIL : https://www.cnil.fr/sites/default/files/atoms/files/wp216_fr_0.pdf	/5
Avez-vous mobilisé un responsable de protection des données personnelles type DPO à un stade précoce du projet ?	La bonne pratique est de prévoir un atelier avec un responsable type DPO en fin de cadrage, avant d'entamer la réalisation	/5
Avez-vous défini des normes et des processus (comme les tests de qualité data) pour garantir la qualité et integrité des données en entrée du système IA ?	Les définitions métiers doivent être claires avant tous tests techniques. Voir par exemple le framework Test Driven Design pour un processus possible de garantie de qualité data	/5
Avez-vous défini des procédures pour limiter l'accès aux données utilisées par le système IA ou générées par les utilisateurs de ce système ? Quels contrôles avez-vous mise en place ? L'accès aux données est-il tracé ?	Il s'agit là essentiellement d'un sujet d'architecture data du système développé. Un système d'authentification des accès avec journalisation comme il en existe sur le cloud est très souhaitable	/5

Transparence (4/7)

Critère	Aide à l'évaluation	Évaluation
Avez-vous documenté les choix/data/paramètres/résultats des phases de conception et test/validation, via un guide méthodologique et un système de traçabilité data/paramètres/résultats par exemple ?	Des outils existent, tel ML Flow Tracking, qui permettent de suivre tout le cycle de vie d'un modèle, depuis la conception jusqu'à la mise en production	/5
Avez-vous prévu des mécanismes d'explication des résultats du modèle et analysé de leur influence possible sur les utilisateurs ?	Une couche d'intelligibilité du modèle via LIME ou SHAP est très souhaitable. L'analyse d'influence sur les utilisateurs peut être réalisée lors d'ateliers en fin de MVP et avant l'industrialisation	/5
Existe-t-il un modèle économique clair pour le système IA ? La valeur créée par ce système est-elle bien définie et identifiée par les parties prenantes ?	La valeur économique du traitement IA peut être par exemple de nature opérationnelle (économie d'ETP, focus des humains sur des tâches à haute valeur ajoutée, etc.) ou financière (plus de ventes, moins de churn, etc.)	/5
Les utilisateurs sont-ils explicitement informés de leur interaction avec un système IA ?	Une notification explicite importante, tout comme la possibilité d'un recours à un agent humain	/5
Existe-t-il des mécanismes pour la prise en compte des retours utilisateurs ?	Les retours des utilisateurs ne doivent pas être faits en fin de projet, mais intégrés dès la phase de conception, d'où l'importance de profils UX dans le développement de solutions	/5
Les parties prenantes (développeurs, utilisateurs, tiers) sont-elles informées à propos des limites, lacunes et risques potentiels du système IA ? Existe-t-il des risques perçus en décalage avec la réalité des limites et lacunes du système IA ?	Une bonne cartographie des risques selon les populations se trouve ici : https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/confronting-the-risks-of-artificial-intelligence	/5
Existe-t-il une information claire sur la finalité du système et ses scénarios d'utilisation ? Les risques humains associés à l'utilisation du système (biais de confirmation, fatigue cognitive, etc.) sont-ils bien documentés et partagés avec les parties prenantes ?	La finalité du système doit être bien documentée, ainsi que les choix méthodologiques et scientifiques réalisés et leur compatibilité avec cette finalité. Exemple : un système d'aide à la décision judiciaire basé sur les décisions passées peut renforcer ou déconstruire les biais sociaux, selon que la finalité est de donner au juge l'état de la pratique actuelle ou lui recommander la décision à prendre	/5

Diversité, non-discrimination et équité (5/7)

Critère	Aide à l'évaluation	Évaluation
Votre compréhension de la conception du système IA (data, algorithme) vous permet-elle d'identifier les biais possibles ? En particulier, avez-vous prévu des tests de représentativité des données, des tests spécifiques sur des populations à risque ? Dans l'idéal, avez-vous défini des contrôles de biais sur tout le cycle de vie du système depuis la conception jusqu'à l'utilisation ?	Un premier niveau de vérification porte sur la nature des données utilisées (sexe, ethnie, etc.). Un deuxième consiste en la réalisation de tests statistiques entre variables et sortie du modèle, avec possibilité de tests spécifiques sur des populations à risque, ex : comment mon modèle se comporte spécifiquement pour les personnes âgées ou de tel sexe ? Voir l'exemple récent de la carte de paiement Apple qui discriminait les femmes. En conséquence, on peut prévoir soit de changer le modèle ou les data, via des techniques de transport optimal par exemple, voir ici : *https://arxiv.org/abs/1806.03195*	/5
Existe-t-il des mécanismes de signalement en cas de biais, discrimination ou mauvaises performances ? Qui peut réaliser un tel signalement, de quelle manière, et à qui ?	Un simple formulaire dans l'interface utilisateur ou dans le processus d'exploitation du traitement peut suffire. Une bonne référence sur les types de biais se trouve ici : https://arxiv.org/pdf/1908.09635.pdf	/5
Avez-vous analysé la variabilité des décisions du système IA, ainsi que les causes probables de telles variations et leurs impacts sur les droits fondamentaux ?	La variabilité des décisions peut être mesurée sur le jeu de données historiques, mais également en production. L'impact sur les droits fondamentaux peut émerger par exemple via des discriminations dues à des variations systématiquement préjudiciables à une population donnée	/5
Quelle définition avez-vous choisi pour la notion d'équité ? Avez-vous comparé cette définition à d'autres possibles ? Avez-vous défini une manière de quantifier l'équité, et des mécanismes de contrôle/correction pour la garantir ?	Exemples de définition : égalité des taux de vrais positifs, des taux VP et FP, etc. Une bonne référence sur les définitions de l'équité se trouve ici : https://arxiv.org/pdf/1908.09635.pdf	/5
Votre système IA a-t-il été conçu pour être accessible à tous types d'utilisateurs ? Quelles vérifications de l'universalité du système ? Si le système n'est pas universel, est-ce pour des raisons bien précises et documentées ?	Un système universel s'entend comme accessible à tous types d'utilisateurs selon la finalité assignée au système, sans discrimination i.e. exclusion ne présentant aucun lien avec la finalité du système	/5
Avez-vous tenu compte des diversités d'opinions qui ont pu exister tout au long du parcours de conception du système ?	Toutes les opinions ne se valent bien évidemment pas, mais si des arbitrages entre opinions légitimes ont été réalisés, il est important de les documenter. Exemple : faut-il ou non intégrer les données open data INSEE pour « colorer » la base de données clients interne via de nouveaux attributs ? Si oui/non, pour quelles raisons ?	/5
Avez-vous évalué l'impact du système IA sur certaines personnes ou groupes, surtout s'il est disproportionné ?	Une bonne cartographie des risques selon les populations se trouve ici : https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/confronting-the-risks-of-artificial-intelligence	/5
Avez-vous prévu une participation des différentes parties prenantes (techniciens, utilisateurs, tiers) à la mise au point et l'utilisation du système ?	-	/5
Avez-vous préparé l'introduction du système IA au sein de l'organisation ?	La communauté data a un devoir de formation/acculturation vis-à-vis des membres de l'organisation impactés par l'introduction de traitements IA	/5

Bien-être sociétal et environnemental (6/7)

Critère	Aide à l'évaluation	Évaluation
Avez-vous défini des mesures de l'impact environnemental du système IA, ainsi que des stratégies de réduction de cet impact sur tout le cycle de vie du système ?	Plusieurs niveaux de mesure d'impact sont possibles : lors de la prédiction, lors de la conception, les modifications de comportements grâce au traitement IA, etc. Il n'existe pas d'outil simple de mesure, ex : les proxy comme la consommation CPU par fonction sont pollués par la gestion séquentielle imbriquée des tâches dans un processeur	/5
Avez-vous évalué les risques d'attachement ou d'incompréhension sur la nature du système par les utilisateurs ?	Eviter autant que possible d'humaniser l'algorithme dans la communication réalisée auprès des utilisateurs. Toujours être clair sur la nature simulée de tout comportement ou émotion humains	/5
Avez-vous une bonne compréhension des incidences sociales du système (emplois, compétences), et quelles mesures sont prises pour pallier les incidences négatives ?	Selon la nature du système (conseil, décision, etc.), il est généralement possible d'établir une équivalence ETP des gains de temps réalisés. L'utilisation des ETP économisés est avant tout un choix de la direction	/5
Avez-vous évalué les incidences du système IA au-delà de l'utilisateur final, des tiers qui peuvent par exemple être indirectement impactés ?	Exemple : des traitements IA de scoring crédit et fraude qui accélèrent la souscription en ligne pour des crédits ou des cartes de paiement ont un impact indirect sur les commerciaux en agence et l'atteinte de leurs objectifs	/5

Responsabilité (7/7)

Critère	Aide à l'évaluation	Évaluation
Avez-vous réalisé une analyse des risques sur toute la chaîne opérationnelle du système IA, et facilité l'audit des mesures de réduction des risques identifiés ?	Les risques doivent être identifiés à chaque étape du traitement : stockage des données (risque de fuite), transformation (risque d'erreur), entraînement (risque de biais), etc. Les mesures de réduction de risque peuvent être techniques (tests data par exemple) ou des processus métier (solution alternative au traitement IA en cas de défaillance)	/5
Votre personnel est-il formé sur les responsabilités associées à la conception et l'utilisation de tels systèmes, notamment sur les aspects juridiques et éthiques ? Existe-t-il un comité d'éthique pour discuter des pratiques globales du personnel sur les sujets IA ?	-	/5
Quelles possibilités de signalement (vulnérabilités, risques, biais) par des tiers (par exemple fournisseurs, consommateurs, distributeurs/vendeurs, travailleurs, etc.) ?	Idem, un simple formulaire peut suffire	/5
Avez-vous documenté les arbitrages d'intérêts et/ou de valeurs réalisés, ainsi que le processus de prise de décision ?	Ces arbitrages peuvent intervenir à tout moment du projet : lors de la sélection des données, lors de l'élimination de variables, lors du choix du modèle global optimal (qui peut être sous-optimal pour certaines populations). Dans l'idéal, les choix réalisés ont un impact mesuré et documenté	/5
Existe-t-il des voies de recours pour les utilisateurs, et en sont-ils informés ?	Les voies de recours doivent être bien définies et systématiquement aboutir à une prise de décision humaine en dernière instance	/5