Strasbourg, le 28 novembre 2024 CAI(2024)16rev2_FR

COMITE SUR L'INTELLIGENCE ARTIFICIELLE (CAI)

MÉTHODOLOGIE POUR L'ÉVALUATION DES RISQUES ET DES IMPACTS DES SYSTÈMES D'INTELLIGENCE ARTIFICIELLE DU POINT DE VUE DES DROITS HUMAINS, DE LA DÉMOCRATIE ET DE L'ÉTAT DE DROIT (MÉTHODOLOGIE HUDERIA)

www.coe.int/cai

Table des matières

Introduction

Qu'est-ce que l'HUDERIA ?

L'Évaluation des risques et des impacts des systèmes d'intelligence artificielle (IA) du point de vue des droits humains, de la démocratie et de l'État de droit (« l'HUDERIA ») est un guide qui fournit une approche structurée de l'évaluation des risques et des impacts des systèmes d'IA spécifiquement adaptée à la protection et à la promotion des droits humains, de la démocratie et de l'État de droit. Elle est destinée à jouer un rôle unique et essentiel à l'intersection des normes internationales en matière de droits humains et des cadres techniques existants sur la gestion des risques dans le contexte de l'IA.

L'HUDERIA peut être utilisée par les acteurs publics et privés pour aider à l'identification et à la réponse à apporter aux risques et aux impacts sur les droits humains, la démocratie et l'État de droit tout au long du cycle de vie des systèmes d'IA.

L'HUDERIA trouve son origine dans les travaux du Comité ad hoc sur l'intelligence artificielle (CAHAI) (2019-2021) et plus particulièrement de son Groupe d'élaboration des politiques, qui a chargé l'Alan Turing Institute, l'institut national du Royaume-Uni pour la science des données et l'IA, de préparer une proposition originale opérationnalisant les grandes lignes d'un modèle d'évaluation de l'impact sur les droits humains, la démocratie et l'État de droit. La Méthodologie HUDERIA a été adoptée par le Comité sur l'intelligence artificielle (CAI) du Conseil de l'Europe le 28 novembre 2024.

Relation avec la Convention-cadre

L'HUDERIA est un guide autonome, non juridiquement contraignant qui, en tant que tel, n'a pas d'effet juridique. Il ne revêt pas de caractère obligatoire et n'a pas vocation à être une aide à l'interprétation de la Convention-cadre du Conseil de l'Europe sur l'intelligence artificielle et les droits de l'homme, la démocratie et l'État de droit, ci-après dénommée « la Convention-cadre ». De nombreux cadres, politiques, orientations, normes ou outils existants ou futurs peuvent être utilisés en soutien à la gestion des risques et des impacts de l'IA, y compris l'HUDERIA.

Les Parties à la Convention-cadre ont la possibilité d'utiliser ou d'adapter le guide, en tout ou en partie, de développer de nouvelles approches de l'évaluation des risques ou d'utiliser ou adapter les approches existantes conformément à leurs lois applicables, à condition que les Parties respectent pleinement leurs obligations au titre de la Convention-cadre, en particulier la référence de base pour la gestion des risques et des impacts énoncé dans son Chapitre V.

Principaux objectifs de l'HUDERIA

L'HUDERIA vise à :

- aider à déterminer dans quelle mesure des activités de gestion des risques pour les droits humains, la démocratie et l'État de droit pourraient être nécessaires, et offrir une méthodologie pour l'identification, l'évaluation, la prévention et l'atténuation des risques et des impacts applicable à une vaste majorité de technologies et de contextes d'application de l'IA, et qui est réactive aux innovations et nouveaux cas d'utilisation de l'IA ;

- promouvoir la compatibilité et à l'interopérabilité avec les guides, normes, et cadres existants et futurs développés par les organisations ou organes compétents (tels que ISO, IEC, ITU, CEN, CENELEC, IEEE, l'OCDE, NIST), y compris le cadre de gestion des risques

liés à l'IA (*AI Risk Management Framework)* du NIST ainsi que la gestion des risques et les évaluations d'impact sur les droits fondamentaux au titre de la loi sur l'IA de l'UE).

Quelle est l'approche de l'HUDERIA ?

L'HUDERIA combine les connaissances contemporaines concernant les processus et mécanismes de gouvernance techniques et socio-techniques pouvant faciliter les activités responsables tout au long du cycle de vie des systèmes d'IA, et les procédures de diligence nécessaires à la protection et la promotion des droits humains, de la démocratie et de l'État de droit.

L'HUDERIA s'appuie sur des variables, des concepts et un langage bien connus pour l'évaluation des risques pour les droits humains (ampleur, portée, probabilité et réversibilité des effets négatifs potentiels sur les droits humains). Il vise à faciliter leur examen en fournissant des orientations supplémentaires compte tenu de la complexité socio-technique du cycle de vie de l'IA.

Approche socio-technique

L'HUDERIA adopte une approche socio-technique, qui considère tous les aspects du cycle de vie du système d'IA comme affectés par la relation interconnectée de la technologie, des choix humains et des structures sociales. Dans cette approche, la gestion des risques et des impacts des systèmes d'IA tient compte à la fois de leurs aspects techniques et des contextes juridique, social, politique, économique, culturel et technologique dans lesquels ils opèrent. Une telle approche promeut le développement d'une IA sûre, sécurisée et digne de confiance, à la fois performante et favorisant le respect des droits humains, de la démocratie et de l'État de droit.

Orientation générale et spécifique

L'HUDERIA offre une structure en combinant une orientation générale et spécifique et de la flexibilité en laissant une marge d'adaptation pour la mise en œuvre pratique.

Au niveau général, la **Méthodologie HUDERIA** décrit des concepts, processus et éléments de haut niveau guidant les activités d'évaluation des risques et des impacts des systèmes d'IA qui sont susceptibles d'avoir des impacts sur les droits humains, la démocratie et l'État de droit.

Au niveau spécifique, le **Modèle HUDERIA**[1] fournira des supports et des ressources (tels que des outils flexibles adaptés aux différents éléments du processus HUDERIA et des recommandations modulables) qui peuvent contribuer à la mise en œuvre de la Méthodologie HUDERIA. Ces ressources sont mentionnées tout au long du texte fourniront une bibliothèque de connaissances qui peut faciliter la prise en compte des risques et des impacts liés aux droits humains, à la démocratie et à l'État de droit, y compris dans d'autres approches de la gestion des risques.

Adaptabilité et flexibilité

La Méthodologie et le Modèle HUDERIA permettent tous les deux une adaptation aux différents contextes, besoins et capacités en fixant des buts, principes et objectifs tout en laissant une marge d'appréciation pour décider de la manière de les atteindre et en offrant un éventail d'options politiques et de gouvernance pouvant être adaptées aux contextes.

[1] A élaborer et adopter par le CAI en 2025

Approche graduée et différenciée

L'HUDERIA vise à établir une approche graduée et différenciée des mesures d'identification, d'évaluation, de prévention et d'atténuation des risques et des impacts, qui tienne compte de la gravité et de la probabilité d'occurrence des impacts négatifs sur les droits humains, la démocratie et l'État de droit, ainsi que des facteurs contextuels pertinents.

Contours de l'HUDERIA

La Méthodologie HUDERIA comprend quatre éléments :

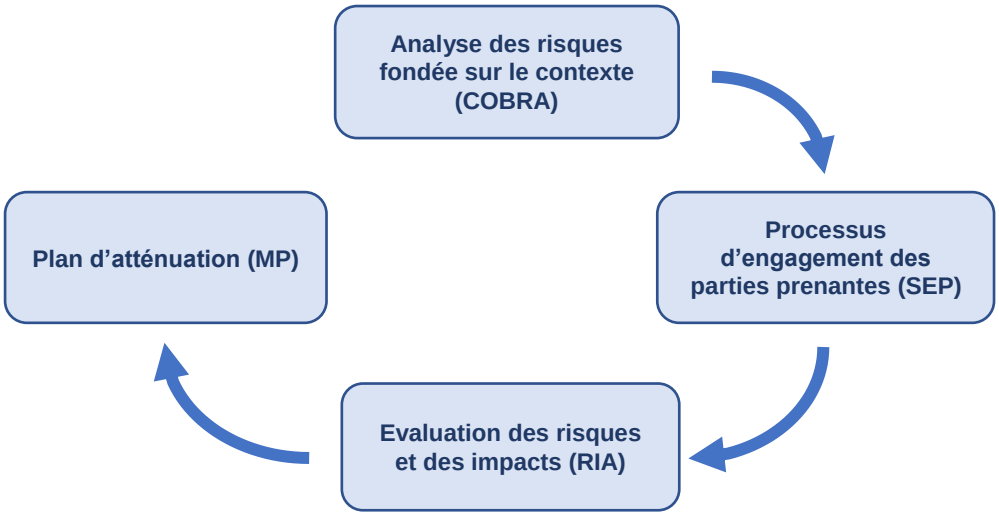
1. l'**Analyse des risques fondée sur le contexte** (COBRA) fournit une approche structurée de la collecte et de la cartographie des informations nécessaires pour identifier et comprendre les risques que le système d'IA pourrait poser pour les droits humains, la démocratie et l'État de droit, compte tenu de son contexte sociotechnique. Elle permet également de déterminer si le système d'IA est une solution appropriée au problème envisagé ;

2. le **Processus d'engagement des parties prenantes** (SEP) propose une approche pour permettre et rendre opérationnel l'engagement, le cas échéant, avec les parties prenantes pertinentes afin d'obtenir des informations sur les personnes potentiellement affectées et de contextualiser et corroborer les préjudices potentiels et les mesures d'atténuation ;

3. l'**Évaluation des risques et des impacts** (RIA) propose des étapes possibles pour l'évaluation des risques et des impacts liés aux droits humains, à la démocratie et à l'État de droit ;

4. le **Plan d'atténuation** (MP) propose des étapes possibles pour définir les mesures d'atténuation et de réparation, y compris l'accès aux recours et l'examen itératif .

S'il est logique d'exécuter l'élément COBRA en premier lieu, en fonction des besoins et des approches qu'elles ont décidé de suivre, il peut être choisi de modifier la séquence des éléments et/ou d'appliquer ou d'utiliser seulement certaines parties de la Méthodologie sur la base des approches existantes en matière de gouvernance de l'IA et des contextes, besoins et exigences spécifiques.



I. L’analyse des risques fondée sur le contexte (COBRA)

Introduction

La COBRA aide à l’identification des différents facteurs de risque - caractéristiques ou propriétés d'un système d'IA et de son contexte qui affectent la probabilité d'impacts négatifs sur les droits humains, la démocratie et l'État de droit. Ces facteurs ne doivent pas nécessairement être traités comme des causes d'impacts négatifs, mais plutôt comme des conditions qui sont corrélées à un risque accru de préjudices et qui doivent donc être anticipées et prises en compte dans les efforts de gestion des risques et d'atténuation des impacts. Les facteurs de risque sont classés en trois grandes catégories : le contexte d'application du système, son contexte de conception et de développement, et son contexte de déploiement[2].

L’examen des facteurs de risques vise à faciliter la cartographie des potentiels impacts négatifs sur les droits humains, la démocratie et l’État de droit. Les résultats de cette analyse de la cartographie des facteurs de risque et des impacts visent à déterminer l’étendue de l’approche concernant les éléments ultérieurs de l’HUDERIA, y compris en établissant la proportionnalité des activités HUDERIA ultérieures.

En outre, le résultat de cette analyse de la cartographie des facteurs de risque et des impacts pourrait aussi aider à identifier les contextes socio-techniques spécifiques qui, tout au long du cycle de vie du système, nécessitent une attention particulière en matière de gouvernance.

L’élément COBRA comprend quatre étapes :

1) le cadrage préliminaire ;

2) l’analyse des facteurs de risque ;

3) la cartographie des impacts potentiels sur les droits humains, la démocratie et l'État de droit ;

4) le triage.

Cadrage préliminaire

Objectifs

Le but principal de cette étape est de mener une recherche de fond préliminaire nécessaire pour guider les activités ultérieures d'identification des facteurs de risque et de cartographie des impacts.

Explications

Le processus COBRA commence par une recherche de cadrage préliminaire qui décrit l’objectif du système, les éléments principaux du système, des contextes dans lesquels il est destiné à être utilisé, du ou des domaines dans lesquels il fonctionnera, du degré d'intervention humaine, ainsi que de la nature et de la quantité des données qu'il traitera et sur lesquelles il sera formé, en notant toutes les vérifications qui ont déjà été effectuées pour

[2] Voir l’explication détaillée de ces domaines à la page 9.

évaluer les biais dans l'ensemble de données ou le modèle, identifie les personnes ou les groupes qui pourraient être affectés par le système ou l'affecter, en se concentrant sur les caractéristiques contextuelles pertinentes des personnes et des groupes identifiés, y compris les caractéristiques protégées et les facteurs de vulnérabilité, fournit un cadrage préliminaire des potentiels effets négatifs sur les droits humains, la démocratie et l'État de droit en explorant les domaines de préoccupation illustratifs[3]; et fournit une première cartographie des rôles et des responsabilités tout au long du cycle de vie du système d'IA[4].

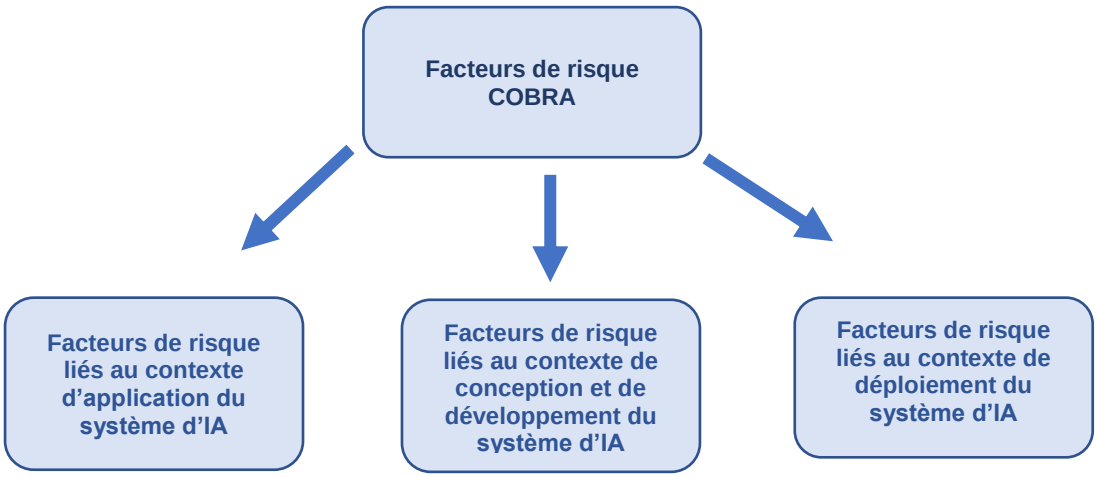
Cette activité préliminaire de cadrage peut s'appuyer sur des documents organisationnels (par exemple, l'étude d'opportunité, la preuve de concept ou la charte du projet), sur la collaboration et sur des recherches documentaires (si nécessaire). Cette activité préliminaire de cadrage ainsi que les éléments ultérieurs du processus HUDERIA devraient se dérouler, le cas échéant, au sein d'une équipe multidisciplinaire, composée d'experts possédant un éventail de spécialisations complémentaires[5] et de formations techniques et non techniques.

Analyse des facteurs de risque

Objectifs

Le but principal de cette étape est de collecter les informations pertinentes concernant les facteurs de risque liés au contexte prévu d'application du système, son contexte de conception et de développement, et son contexte de déploiement. Ces facteurs faciliteront la cartographie des potentiels impacts négatifs sur les droits humains, la démocratie et l'État de droit et l'évaluation ultérieure des variables de risque principales : gravité (ampleur, portée et réversibilité)[6] et probabilité.

Explications



[3] Les **Ressources COBRA E (Domaines de préoccupation illustratifs du point de vue des droits humains, de la démocratie et de l'État de droit)** [à élaborer et adopter par le CAI en 2025] fournissent un outil qui pourrait être utilisé pour conduire ou informer cette évaluation.
[4] La section **Rôles et responsabilités** du Modèle HUDERIA fournira des conseils relatifs à cet aspect de la Méthodologie.
[5] L'expertise pertinente pourrait comprendre, le cas échéant, les questions de droits humains, de protection de la vie privée et des données personnelles, la science des données, la gestion des ensembles de données, la sécurité, les risques liés à l'IA, et les tests, l'évaluation, la vérification, et la validation de l'IA.
[6] Conformément à la Convention-cadre, au Haut-Commissariat des Nations unies aux droits de l'homme et aux Principes directeurs des Nations unies relatifs aux entreprises et aux droits de l'homme, le terme « gravité » s'entend, aux fins de l'HUDERIA, comme une combinaison des variables que sont l'ampleur, la portée et la réversibilité.

Les systèmes d'IA sont conçus, développés et utilisés dans une grande variété de contextes et de nombreuses façons différentes, rendant importante l'évaluation holistique de divers facteurs liés au contexte d'application du système, son contexte de conception et de développement, et son contexte de déploiement.

Le **contexte d'application du système** d'IA[7] comprend des informations relatives au secteur et au domaine d'application du système, aux environnements juridique et réglementaire dans lequel le système est développé et utilisé, l'objectif visé par le système, et d'autres détails pertinents du contexte d'application du système, tels que e tout héritage connu de biais de discrimination.

Le **contexte de conception et de développement du système d'IA[8]** comprend les caractéristiques techniques pertinentes du système. Il peut s'agir de limitations connues du système, de considérations liées à la collecte, à l'enrichissement, au stockage, à l'utilisation et au retrait des données, ainsi que de considérations liées à l'algorithme ou au modèle lui-même. Parmi les considérations particulièrement pertinentes, on peut citer les caractéristiques techniques liées à la vie privée et à la protection des données, au biais et à la discrimination, ainsi qu'à l'explicabilité et à l'interprétabilité.

Enfin, le **contexte de déploiement du système d'IA[9]** comprend des facteurs qui régissent la manière dont les risques potentiels peuvent se manifester et être gérés en pratique, comme les mesures qui seront prises pour protéger la vie privée et les données personnelles, atténuer les biais dommageables, assurer une formation adéquate, se prémunir contre les utilisations involontaires, et garantir la responsabilité et la conformité juridique.

Cartographie des impacts potentiels sur les droits humains, la démocratie et l'État de droit

Objectifs

L'étape de cartographie identifie les personnes ou groupes potentiellement affectés et conduit une évaluation initiale des variables de risque principales - gravité (ampleur, portée, réversibilité) et probabilité. La cartographie aide à informer les éléments ultérieurs de la Méthodologie, et l'étendue de l'intervention de gouvernance et des mesures d'atténuation qui pourraient être appropriées (voir la section Triage ci-dessous). L'analyse des principales variables de risque est essentielle pour obtenir une vue d'ensemble claire et structurée des endroits où les menaces sont les plus susceptibles de se produire et de leur impact potentiel.

Explications

Les **Ressources COBRA E[10] et F[11]** pourraient être utilisées pour identifier les domaines d'application potentiellement sensibles et les domaines de préoccupation potentiellement pertinents liés aux droits humains, à la démocratie et à l'État de droit[12].

---

[7] Les **Ressources COBRA A (Liste des facteurs de risque liés au contexte d'application du système)** [à élaborer et adopter par le CAI en 2025] fournissent un outil qui pourrait être utilisé pour conduire ou informer cette évaluation.

[8] Les **Ressources COBRA B (Liste des facteurs de risque apparaissant dans les contextes de conception et de développement du système)** [à élaborer et adopter par le CAI en 2025] fournissent un outil qui pourrait être utilisé pour conduire ou informer cette évaluation.

[9] Les **Ressources COBRA C (Liste des facteurs de risque liés au contexte de déploiement du système)** [à élaborer et adopter par le CAI en 2025] fournissent un outil qui pourrait être utilisé pour conduire ou informer cette évaluation.

[10] A élaborer et adopter par le CAI en 2025

[11] A élaborer et adopter par le CAI en 2025

[12] Dans les **ressources COBRA**, les références aux droits humains tels qu'ils sont énoncés dans les divers instruments internationaux relatifs aux droits humains sont incluses à titre d'illustration. Ces références ne

A l'aide des informations collectés lors des étapes précédentes :

a) déterminer si le système fonctionnera à proximité d'activité(s) (prise de décision ou actions) pouvant produire des impacts sur des personnes affectées dans les secteurs/domaines[13] ;

b) identifier et énumérer les domaines de préoccupation pertinents et, avec cela à l'esprit, répondre à la question de savoir si le système pourrait avoir des impacts potentiels ou réels sur des droits humains spécifiques, la démocratie et l'État de droit[14] ;

c) pour chaque impact potentiel ou réel identifié, décrire la nature de l'impact potentiel ou réel[15], en tenant compte des impacts différentiels sur les personnes et les groupes affectés en fonction des caractéristiques contextuelles pertinentes, y compris les caractéristiques protégées et les facteurs de vulnérabilité ;

L'analyse de ces points fournira des informations pour l'évaluation initiale des principales variables de risque - la gravité (ampleur, portée, réversibilité) et la probabilité - qui aident à déterminer le risque et à choisir la bonne approche pour les éléments ultérieurs de la Méthodologie, ce qui contribuera à garantir que les interventions de gouvernance et les mesures d'atténuation sont alignées sur les besoins tout au long du cycle de vie du système d'IA.

Les résultats de cette analyse peuvent également aider à identifier les possibilités d'utiliser le système d'IA pour soutenir des actions positives qui font progresser les droits humains, y compris la promotion et la garantie de la non-discrimination.

Détermination du niveau de risque

Les variables suivantes peuvent être employées pour indexer le niveau de risque de chacun des impacts négatifs potentiels sur les droits humains, la démocratie et l'État de droit qui ont été identifiés à la suite de l'exercice de cartographie :

1) L'**ampleur**[16] des impacts négatifs potentiels (c'est-à-dire l'importance du préjudice potentiel) ;

2) La **portée** des effets négatifs potentiels (y compris le nombre de personnes affectées, les caractéristiques protégées ou la vulnérabilité des individus ou des groupes et la durée des impacts) ;

3) La **réversibilité**[17] des impacts négatifs potentiels est l'information sur la possibilité de réparer ou de rétablir les personnes affectées dans leur situation antérieure à l'impact ou dans une situation équivalente.

s'appliquent qu'aux États qui sont Parties à ces instruments. Chaque État est censé appliquer ses propres lois et obligations internationales.

[13] Les **Ressources COBRA F** pourraient être utilisées pour identifier des secteurs/domaines potentiellement sensibles.

[14] Les **Ressources COBRA E** pourraient être utilisées pour identifier des domaines de préoccupation relatifs aux droits humains, à la démocratie et à l'État de droit.

[15] Pour la clarté de l'évaluation, il convient de prendre en compte à la fois les impacts négatifs ou restrictifs et les impacts bénéfiques, améliorants ou autrement positifs produits par les systèmes d'IA, étant donné que divers problèmes de partialité et de discrimination peuvent se poser en ce qui concerne les systèmes qui produisent les deux types d'impacts.

[16] Le terme « ampleur » est parfois appelé « sévérité » dans le contexte de l'évaluation des risques.

[17] Le terme « réversibilité » est parfois appelé « rémédiabilité » dans le contexte de l'évaluation des risques.

4) La **probabilité** des impacts négatifs potentiels.

Les équipes concernées devraient passer en revue chacun des impacts potentiels qui ont été identifiés et considérer pour chaque domaine de préoccupation lié aux droits humains, à la démocratie et à l'Etat de droit, et chaque groupe affecté, l'ampleur, la portée, la réversibilité et la probabilité des effets négatifs potentiels ou réels. La législation ou la politique nationale peut fournir des définitions plus détaillées qui peuvent être utilisées pour éclairer cette détermination des risques et pour déterminer des approches appropriées et proportionnées pour les activités ultérieures d'HUDERIA (par exemple, l'engagement des parties prenantes).

Il pourrait être envisagé d'établir une méthode pour combiner ces variables afin de permettre l'étalonnage du risque et la détermination d'approches appropriées et proportionnées pour les activités ultérieures de HUDERIA (par exemple, l'engagement des parties prenantes), ainsi que l'étendue et la profondeur des interventions de gouvernance et des mesures de gestion et d'atténuation des risques. Cela peut impliquer la formulation de méthodes quantitatives ou semi-quantitatives de calcul des risques, de matrices de risques ou de procédures plus qualitatives ou fondées sur des règles. Tout mécanisme d'étalonnage des risques résultant de la combinaison de ces variables aux fins de l'évaluation d'HUDERIA peut prendre en compte les éléments suivants :

- en ce qui concerne les droits humains, les effets de faible portée et à fort impact ainsi que les effets de grande portée et à faible impact sur chaque personne affectée ;

- en ce qui concerne la démocratie et l'État de droit, le mécanisme pourrait prendre en compte en particulier, les effets durables et de grande portée sur les personnes, les institutions et la société en général.

Triage

Objectifs

L'objectif principal de cette étape est de s'appuyer sur les informations recueillies dans le cadre des activités COBRA précédentes, et ainsi de :

- faciliter l'identification et le triage des systèmes qui présentent un risque significatif, de sorte que la Méthodologie HUDERIA ne soit pas onéreuse pour les systèmes d'IA à risque minime ou faible ;

- de procéder à une évaluation initiale visant à déterminer si le système d'IA devrait être développé ou déployé, en se fondant sur la question de savoir si les avantages du développement ou du déploiement du système d'IA l'emportent sur les risques qu'il comporte, compte tenu notamment de ses impacts potentiels sur les droits humains, la démocratie et l'État de droit, et si l'utilisation du système d'IA est incompatible avec le respect des droits humains, de la démocratie et de l'État de droit.

Approche adaptable du triage

Les activités préalables de cette étape permettent de dresser un premier tableau du profil de risque du système d'IA.

Les informations recueillies peuvent, par exemple, être suffisantes pour déterminer qu'il est peu probable que le système ait un quelque impact sur les droits humains, la démocratie ou l'État de droit, rendant ainsi superflus les éléments suivants d'HUDERIA. Une conclusion similaire pourrait être tirée si les impacts identifiés sont insignifiants ou peu probables. Si les impacts identifiés conduisent à la décision de ne pas développer ou déployer un système d'IA

parce qu'il est considéré comme incompatible avec le respect des droits humains, de la démocratie et de l'État de droit, les éléments suivants de l'HUDERIA sont également inutiles. Enfin, dans les cas où des impacts potentiels graves sont identifiés, un éventail de stratégies de gestion des risques et de réponses (y compris le **processus d'engagement des parties prenantes** et diverses autres activités de gouvernance du projet) peuvent être justifiées. Pour répondre à cette complexité, l'HUDERIA ne prescrit pas d'orientations détaillées pour ajuster les efforts de gestion des risques, mais présente simplement des éléments proposés qui pourraient être appliqués sur la base du risque d'impacts négatifs potentiels sur les droits humains, la démocratie et l'État de droit. Différentes approches de la détermination des étapes de gestion des risques basées sur les impacts potentiels et réels identifiés - ou une combinaison d'entre elles - peuvent être appliquées en fonction du cadre ou de l'environnement réglementaire national spécifique, du secteur, du système et du contexte (par exemple, approches fondées sur des seuils, des scénarios, la proportionnalité, la dynamique ou des approches spécifiques au contexte).

La décision finale d'utiliser une méthode qualitative, quantitative, mixte ou autre est laissée à la discrétion des autorités ou, le cas échéant, des équipes de projet d'IA responsables du système.

"Questions zéro"

Pour déterminer si les avantages de la construction ou du déploiement du système d'IA, y compris les avantages sociaux supplémentaires pouvant résulter de l'utilisation du système au-delà de son objectif premier, l'emportent sur les risques compte tenu des facteurs de risque et des impacts potentiels identifiés, il convient de considérer les points suivants :

- si l'utilisation du système est appropriée compte tenu de la nature du problème que le projet d'IA tente de résoudre ;

- la mesure dans laquelle les technologies et les processus existants déjà en place pour résoudre le problème considéré sont mieux placés pour le faire, compte tenu du profil de risque et des impacts négatifs potentiels du système envisagé, en mettant l'accent, le cas échéant, sur tout risque marginal ajouté par l'introduction de l'IA dans le contexte actuel ;

- la mesure dans laquelle le système envisagé sera en mesure de répondre aux besoins et aux attentes des opérateurs ;

- la mesure dans laquelle les effets du système prospectif seront équitables pour tous les groupes affectés ;

- la mesure dans laquelle la qualité et la représentativité des données actuellement ou potentiellement disponibles sont suffisantes pour que le système prospectif soit efficace et sûr, et qu'il évite raisonnablement les biais potentiellement dommageables ;

- la mesure dans laquelle des ressources (humaines et matérielles) suffisantes sont disponibles et capables de répondre aux exigences techniques et pour mener à bien des actions techniques et de gouvernance afin d'atténuer de manière adéquate les risques identifiés ; et

- les contextes d'utilisation potentiels du système et les risques de mauvaise utilisation ou d'abus, y compris à travers son déploiement au-delà de l'objectif prévu.

II. Processus d'engagement des parties prenantes (SEP)

Introduction

La possibilité de conduire cette étape peut être considérée afin d'améliorer la qualité des informations pour l'élément suivant de l'HUDERIA - l'**Evaluation des risques et des impacts** - en intégrant les points de vue des personnes potentiellement affectées identifiées, y compris celles en situation de vulnérabilité.

L'engagement des parties prenantes, tel qu'il est défini dans la Méthodologie HUDERIA, peut prendre diverses formes. Le niveau de participation des personnes affectées devrait être informée par les facteurs de risques et impacts potentiels et réels identifiés lors de l'étape COBRA. L'implication des parties prenantes tout au long du cycle de vie du système d'IA peut également offrir une variété d'avantages supplémentaires tels que la promotion de la transparence, l'instauration de la confiance et l'amélioration de la facilité d'utilisation et les performances du système d'IA.

Explication

Le SEP comporte cinq étapes clés[18] : Analyse des parties prenantes, Réflexion sur la positionnalité, Définition des objectifs de l'engagement, Détermination de la méthode d'engagement, Lancement et mise en œuvre.

Analyse des parties prenantes

L'analyse des parties prenantes identifie les groupes de parties prenantes susceptibles d'être affectés, ou d'affecter, les activités menées dans le cadre du cycle de vie de système. Cette analyse[19] doit permettre d'évaluer les intérêts relatifs, les droits, les vulnérabilités et avantages potentiels et existants des parties prenantes identifiées, ainsi que l'importance des groupes de parties prenantes identifiés. A cette étape, considérer d'inclure de façon constructive les points de vue de ceux qui :

1) sont exposés de manière disproportionnée aux risques liés à l'utilisation du système ;
2) sont particulièrement vulnérables aux éventuels préjudices ; ou
3) ont une capacité particulièrement limitée à influer sur la manière dont le système est conçu (par exemple, groupes actuellement ou historiquement marginalisés, défavorisés ou sous-représentés ou personnes en situation de vulnérabilité, ou ayant des besoins spécifiques).

Réflexion sur la positionnalité

L'étape suivante du SEP consiste à réfléchir au point de vue positionnel vis-à-vis des parties prenantes affectées, le but étant de reconnaître les limites des perspectives des utilisateurs

[18] Le processus décrit dans cette section est de nature illustrative, la décision finale concernant le processus de participation des parties prenantes étant laissée à la discrétion des autorités ou, le cas échéant, des équipes de projet d'IA responsables du système.
[19] Les **Ressources SEP A (Liste de questions pour l'évaluation de l'importance des parties prenantes)** [à élaborer et adopter par le CAI en 2025] fournissent des questions et outils détaillés pour guider l'identification des parties prenantes pertinentes.

d'HUDERIA et d'identifier les points de vue manquants qui permettraient d'améliorer l'évaluation des impacts potentiels et réels du système.

En fonction des facteurs de risques et des impacts potentiels identifiés lors de la phase COBRA, cela pourrait comprendre une évaluation autodéterminée des utilisateurs d'HUDERIA de leurs caractéristiques démographiques, leur éducation et leur formation, leur parcours socioéconomique, mais aussi le contexte à l'échelle de l'équipe et de l'organisation.

Les principales questions auxquelles les utilisateurs d'HUDERIA doivent réfléchir lorsqu'ils entreprennent cette étape de la Méthodologie sont les suivantes :

- Dans quelle mesure mes caractéristiques personnelles, mes identifications de groupe, mon statut socio-économique, mes antécédents en matière d'éducation, de formation et de travail, la composition de mon équipe et mon cadre institutionnel représentent-ils des sources de pouvoir et d'avantages ou des sources de marginalisation et de désavantages ?

- Comment cette position influence-t-elle ma capacité et celle de mon équipe à identifier et à comprendre les parties prenantes affectées et les impacts potentiels du système d'IA ?

En fonction des facteurs de risque identifiés au cours du processus COBRA, les utilisateur d'HUDERIA devraient également envisager de faire appel à des parties prenantes externes ou à des consultants disposant d'une expertise spécifique, par exemple en matière de droit des droits humains, en ce qui concerne les impacts potentiels et réels du système sur les droits humains.

Définition des objectifs de l'engagement

La définition d'objectifs clairs pour l'engagement des parties prenantes vise à créer une compréhension aisée de *comment* et *pourquoi* des activités d'engagement sont menées. Ces objectifs doivent permettre l'engagement inclusif, éclairé et constructif des personnes potentiellement affectées[20].

Détermination de la méthode d'engagement

Pour déterminer la/les méthode(s) appropriée(s)[21] à employer pour l'engagement des parties prenantes, il est nécessaire d'évaluer les besoins des personnes potentiellement affectées en prenant en considération, le cas échéant, les résultats du processus COBRA et d'autres facteurs pertinents tels que les contraintes en matière de ressources, les difficultés à aller au-devant des groupes isolés ou socialement exclus, les contraintes de moyens comme les difficultés dues à la fracture numérique ou au déficit d'information, les délais, etc.

Les critères suivants peuvent servir d'orientation pour l'élément SEP :

1) l'**engagement** - la participation significative des personnes affectées ou potentiellement affectées est intégrée dans les éléments pertinents du processus ;

2) l'**égalité et l'interdiction de la discrimination** - les processus d'engagement et de consultation sont inclusifs, sensibles au genre et tiennent compte des besoins des personnes

[20] Les **Ressources SEP B (Liste des facteurs déterminant les objectifs et les niveaux d'engagement des parties prenantes)** [à élaborer et adopter par le CAI en 2025] fournissent des questions détaillées indicatives et des options d'engagement des parties prenantes.

[21] Les **Ressources SEP C (Exemples de méthodes d'engagement et de questions pertinentes)** [à élaborer et adopter par le CAI en 2025] fournissent de possibles exemples de méthodes d'engagement et une liste de questions pertinentes qui pourront aider à la détermination des groupes de parties prenantes appropriés.

et des groupes présentant des caractéristiques protégées ou susceptibles d'être vulnérables ou marginalisés ;

3) l'**autonomisation** - la prise en compte du caractère approprié en fonction de l'âge et les besoins en termes d'accessibilité, et le renforcement des capacités des personnes et des groupes présentant des caractéristiques protégées ou risquant d'être vulnérables ou marginalisés est entrepris afin de garantir leur participation significative ;

4) la **transparence** - assurer le partage d'informations pertinentes et intelligibles entre les parties prenantes à intervalles réguliers, mettre à la disposition des parties prenantes participantes des informations sur le système d'IA qui permettent une compréhension globale des potentiels implications et impacts sur les droits humains, le cas échéant, communiquer publiquement les conclusions d'HUDERIA et les plans de gestion des impacts (plans d'action) ; et

5) la **responsabilité** - la responsabilité de la mise en œuvre, du contrôle et du suivi des mesures d'atténuation est attribuée à des entités, des personnes ou des fonctions particulières au sein de l'organisation.

Mise en œuvre

Une fois les quatre activités précédentes menées à terme, les processus d'engagement proportionné peuvent être menés. Ils devraient être en phase avec les résultats de l'analyse des parties prenantes, la réflexion sur la positionnalité et les objectifs et méthodes d'engagement, et être documentés de manière appropriée.

III. L'Evaluation des risques et des impacts

Introduction

L'objectif de l'évaluation des risques et des impacts est de fournir des évaluations détaillées des impacts potentiels et réels que les activités menées au cours du cycle de vie d'un système d'IA pourraient avoir sur les droits humains, la démocratie et l'État de droit.

Conformément au triage effectué lors de l'étape COBRA, réaliser l'évaluation des risques et des impacts est particulièrement important pour les systèmes d'IA qui pourraient poser un risque significatif pour les droits humains, la démocratie et l'État de droit. Après le triage de l'analyse COBRA, cette étape des processus peut n'être nécessaire que pour certains systèmes d'IA, en particulier ceux qui sont considérés comme présentant des risques importants pour les droits humains, la démocratie et l'État de droit.

L'évaluation des risques et des impacts vise à :

- réexaminer, contextualiser et corroborer les préjudices potentiels et réels identifiés lors de la COBRA ;

- identifier et analyser d'autres préjudices potentiels et réels en se livrant à une réflexion approfondie afin d'identifier des lacunes dans la complétude et l'exhaustivité des préjudices précédemment énumérés ;

- évaluer les variables de risques en matière d'échelle, de portée, de réversibilité et de probabilité des impacts négatifs potentiels, afin de mieux évaluer les risques pour être ensuite hiérarchisés, gérés et atténués .

L'évaluation des risques et des impacts s'inscrit dans le prolongement de l'identification des facteurs de risques contextuels pour les droits humains, la démocratie et l'État de droit et la cartographie des impacts potentiels pour les droits humains, la démocratie et l'État de droit menée dans la COBRA et des potentiels éclairages apportés par le SEP, afin de traiter les impacts potentiels et réels du système d'IA.

Cette évaluation est conduite de manière constructive à travers un processus en deux étapes qui permettent d'élaborer un Plan d'atténuation des impacts et la mise en place d'Accès à des voies de recours lors de l'étape suivante de la Méthodologie HUDERIA.

Explications concernant les questions et pistes de réflexion pour l'évaluation des risques et des impacts

Introduction

L'évaluation des risques et des impacts dans le contexte de l'HUDERIA est organisée en deux étapes.

Lors de la première étape, l'accent est mis sur l'identification des impacts potentiels et réels et, plus spécifiquement, sur « comment » les impacts potentiels et réels identifiés lors des étapes COBRA et SEP pourraient se produire, ce qui permet une approche plus ouverte et exploratoire et facilite une analyse plus approfondie des contextes spécifiques, de la portée, de l'ampleur et de la réversibilité des impacts, en particulier en ce qui concerne les individus en situation de vulnérabilité ou les groupes vulnérables.

Lors de la seconde étape, l'évaluation des variables de risque que sont l'ampleur, la portée, la réversibilité et la probabilité des impacts potentiels ou réels identifiés est conduite. Un examen approfondi de ces variables en fonction du contexte permet de hiérarchiser les mesures d'atténuation en différenciant la gravité des impacts du système d'IA.

Ampleur

L'ampleur d'un impact négatif potentiel et réel se réfère à la sévérité de la conséquence attendue du préjudice potentiel.

La considération de la gravité de tout préjudice potentiel devrait inclure une réflexion quant aux différentes manières et aux différentes mesures dans lesquelles les personnes ou les groupes (en particulier ceux qui possèdent des caractéristiques qui pourraient les rendre plus vulnérables à l'impact négatif) pourraient subir ce préjudice.

Les délibérations concernant l'ampleur prennent en considération les questions supplémentaires suivantes :

1) Pour chaque impact négatif potentiel et réels identifié, existe-t-il des personnes ou des groupes qui possèdent des caractéristiques susceptibles de les rendre plus vulnérables à l'impact ? Dans l'affirmative, quelles sont ces caractéristiques et ceux qui les possèdent pourraient-ils subir le préjudice de manière plus aiguë ou plus grave que les autres ?

2) Pour chaque impact négatif potentiel et réel identifié, quels sont les personnes ou les groupes qui pourraient subir l'impact le plus grave préjudice considéré ?

Les réponses à ces questions joueront par la suite un rôle important lors de la phase de planification des mesures d'atténuation, lorsque la réparation et la hiérarchisation des potentiels préjudices seront envisagées.

Portée

La portée d'un effet négatif potentiel et réel concerne l'estimation du nombre de personnes affectées et de la durée des effets.

Les estimations de la portée des impacts négatifs potentiels et réel identifiés sont examinées une à une, en accordant une attention particulière aux niveaux d'exposition au préjudice de certains groupes de personnes affectées ainsi qu'aux impacts cumulés ou agrégés du système sur les personnes et groupes de personnes affectés actuels et futurs.

Les délibérations sur la portée peuvent comprendre la considération des questions suivantes :

1) Pour chaque impact potentiel et réel identifié, existe-t-il des groupes présentant des caractéristiques susceptibles de les rendre plus vulnérables à des niveaux plus élevés

d'exposition[2223] à l'impact en question ? Dans l'affirmative, à quel niveau d'exposition ces groupes pourraient-ils être confrontés ?

2) Pour chaque impact potentiel et réel identifié, réfléchissez à la durée globale des impacts du système d'IA sur le droit ou le domaine de préoccupation (dans le cas de la démocratie ou de l'Etat de droit) considéré. Le système a-t-il des impacts cumulés ou agrégés sur les personnes affectées et les générations futures de personnes affectées qui auraient pour conséquence d'étendre les effets de l'impact au-delà de la portée déjà identifiée ?

Quelques questions générales à considérer lors de l'évaluation des impacts cumulés ou agrégés pourraient inclure :

- Y a-t-il un risque que la mise à disposition et l'utilisation du système contribuent à produire, plus largement, des effets négatifs sur les droits humains, la démocratie ou l'Etat de droit lorsque le déploiement du système est coordonné avec (ou se produit en même temps que) d'autres systèmes qui remplissent des fonctions ou des objectifs analogues ?

- Y a-t-il un risque que la mise à disposition et l'utilisation du système reproduisent, renforcent ou accroissent des préjudices hérités, enracinés dans l'histoire sociale, ou des caractéristiques inhérentes qui pourraient avoir des effets induits pour les individus et les groupes concernés ?

- Peut-on considérer que la mise à disposition et l'utilisation du système contribuent, plus largement, à des effets négatifs agrégés (par exemple sur l'environnement et la santé publique) si l'on met en parallèle le déploiement du système et l'existence d'autres systèmes susceptibles d'avoir des impacts comparables ?

Réversibilité

Comme expliqué précédemment, la réversibilité se réfère à l'information sur le degré de réparabilité ou de restauration que les personnes potentiellement affectées concernées peuvent attendre comme suite aux efforts déployés pour surmonter l'impact négatif considéré, et de replacer les personnes affectées dans une situation au moins identique ou équivalente à celle dans laquelle elles se trouvaient avant l'impact considéré. À l'instar des considérations sur l'ampleur des impacts potentiels, la compréhension de la réversibilité d'un préjudice dépend de la connaissance que l'on a du contexte spécifique de ce préjudice et des personnes affectées qui y sont soumises. Pour déterminer le degré de réversibilité d'un impact négatif potentiel, il convient de réfléchir aux efforts à fournir pour surmonter le préjudice et (éventuellement) y mettre fin.

Les membres de différents groupes peuvent nécessiter des niveaux d'effort variés pour surmonter des impacts négatifs, en fonction de leur âge, de leur position dans la société et

[22] Les **Ressources SEP A et B** [à élaborer et adopter par le CAI en 2025] fournissent des questions détaillées **(Liste de questions pour l'évaluation de l'importance des parties prenantes)** et la description des formats d'engagement des parties prenantes **(Liste des facteurs déterminant les objectifs et les niveaux d'engagement des parties prenantes)** niveaux d'engagement des parties prenantes pouvant aider les équipes de projet à déterminer les groups de parties prenantes particulièrement pertinents et les objectifs d'engagement.

[23] Le terme « niveau d'exposition » est ici compris comme la proportion d'un groupe qui est affectée négativement par un système d'IA. Lorsque seule une petite fraction du groupe est impactée, les membres ont un faible niveau d'exposition, et lorsque qu'une très grande fraction du groupe est impactée, les membres ont un niveau d'exposition élevé. Par exemple, les membres d'un groupe caractérisé par un faible statut socio-économique peuvent avoir un niveau d'exposition élevé aux impacts négatifs potentiels d'un modèle d'IA utilisé pour l'attribution de prestations publiques.

des circonstances du préjudice (les groupes vulnérables et marginalisés disposant souvent de moins de résilience que les groupes dominants, privilégiés ou majoritaires).

Probabilité

L'évaluation de la probabilité d'un risque consiste à estimer la vraisemblance qu'un risque donné se produise, en se fondant, le cas échéant, sur un jugement qualitatif, une analyse quantitative et une compréhension du contexte.

La détermination du niveau de probabilité d'un risque implique une large analyse des conditions contextuelles et opérationnelles et est généralement déterminée par le niveau (type, quantité et qualité) des informations indiquant que le risque est susceptible de se matérialiser. Cela garantit que les évaluations des risques sont fondées à la fois sur des données et sur des avis d'experts, ce qui facilite la hiérarchisation et l'atténuation des risques potentiels.

Résultat de l’évaluation des risques et des impacts

Après avoir répondu aux questions et réfléchi aux pistes de réflexion sur l’identification et l’évaluation des impacts négatifs potentiels et réels, la prévention des impacts ainsi que la hiérarchisation et la planification des mesures d’atténuation peuvent être amorcées. Ce processus de planification des mesures d’atténuation et de mise en place d’accès à des voies de recours est traité dans la section qui suit.

IV. Plan d'atténuation

Introduction

Une fois que les impacts négatifs potentiels et réels ont été identifiés et évalués, un plan d'atténuation devrait être élaboré et une réflexion concernant la mise en place d'accès aux voies de recours pour les personnes potentiellement affectées devrait avoir lieu, le cas échéant.

Cette partie du processus HUDERIA spécifie les actions et les processus visant à traiter les impacts négatifs potentiels et réels en ;

- formulant des mesures d'atténuation ;

- élaborant un plan d'atténuation en fonction de la gravité et de la probabilité des préjudices identifiés ;

- le cas échéant, mettre en place l'accès à des voies de recours pour les personnes potentiellement affectées et les autres parties concernées.

Explications

Cadrage et priorisation

La planification diligente de la prévention et de l'atténuation des risques et des impacts débute par une étape de cadrage et de priorisation. Grâce aux contributions des personnes affectées associées le cas échéant, il faudrait passer en revue chaque impact négatif potentiel et réel identifié et cartographier les relations mutuelles et les interdépendances entre ces impacts ainsi que les facteurs de risque sociaux environnants identifiés lors de l'étape COBRA (par exemple, vulnérabilités et précarité spécifiques au contexte).

Lorsqu'il est nécessaire de prioriser les mesures de prévention et d'atténuation (par exemple, lorsque des retards dans la prise en charge d'un préjudice potentiel ou la vulnérabilité spécifique d'un individu ou groupe affecté risquent de réduire sa réversibilité), les décisions devraient être dictées par la prise en compte la probabilité et de la gravité relative des impacts examinés.

Obligations juridiques

Une considération importante dans l'élaboration d'un plan d'atténuation est que des obligations juridiques concernant le respect des droits humains, de la démocratie et de l'Etat de droit telles que définies dans le droit international et national applicable, devraient être prises en compte à cette étape du processus HUDERIA en considérant si, et le cas échéant, comment, les impacts négatifs potentiels peuvent être atténués et les impacts négatifs réels peuvent être traités.

La disponibilité et l'efficacité des recours juridiques, y compris la restauration ou l'indemnisation en tant que recours juridiques, sont déterminées par le droit international et national applicable.

Hiérarchie des mesures d'atténuation

Pour déterminer l'éventail des mesures qui peuvent être prises pour prévenir ou atténuer les impacts négatifs potentiels, une approche structurée appelée « hiérarchie des mesures d'atténuation » (éviter, atténuer, restaurer, compenser) peut être utilisée.

Au cours des phases initiales de conception, au début du cycle de vie du système d'IA, les impacts à examiner n'auront pas encore eu lieu. À ce stade, les options d'atténuation « éviter » et « atténuer » sont donc plus pertinentes. En revanche, lors des itérations ultérieures du suivi, réexamen et de la réévaluation (c'est-à-dire en phase de déploiement), des impacts négatifs pourraient s'être produits. Les options « restaurer » et « compenser » auront donc toute leur place, aux côtés des options « éviter » et « atténuer ».

Les différentes options de la hiérarchie des mesures d'atténuation sont les suivants :

EVITER	ATTENUER	RESTAURER	COMPENSATE
Procéder à des changements dans les processus de conception, de développement et de déploiement qui précèdent la production et l'utilisation du système d'IA, ou dans le système lui-même, dès le début, pour éviter les impacts négatifs. Attention : éviter ne signifie pas ignorer.	Mettre en œuvre des mesures au niveau des processus de conception, de développement ou de déploiement qui précèdent la production et l'utilisation du système d'IA, ou procéder à des changements dans le système lui-même, afin de réduire l'impact négatif le plus possible.	Procéder à des changements pour replacer les personnes affectées dans une situation au moins identique ou équivalent à celle dans laquelle ils se trouvaient avant le préjudice.	Compensation en nature ou par d'autres moyens, lorsque faisable et d'autres mesures d'atténuation ne sont pas envisageables ou pas efficaces.
Niveau 1	Niveau 2	Niveau 3	Niveau 4
Mesure la plus recommandée	…		Mesure la moins recommandée

L'utilisation de l'expression « hiérarchie des mesures d'atténuation » suggère que la priorité soit donnée, dans un premier temps, à l'évitement total des impacts négatifs potentiels et réels, puis à leur réduction et à leur réparation. Il convient également de noter qu'aux derniers stades du cycle de vie du système d'IA, lorsque les options de restauration et de compensation sont plus pertinentes, plusieurs options d'atténuation peuvent être pertinentes (c'est le cas par exemple d'une personne affectée qui doit être restaurée dans ses droits en même temps que des mesures immédiates sont aussi prises pour réduire le plus possible les préjudices ultérieurs).

En tout état de cause, les décisions concernant la ou les mesures de prévention et/ou d'atténuation à prendre devraient être guidées par des considérations priorisant la protection des droits humains, de la démocratie et de l'Etat de droit, et les choix visant à éviter et à atténuer les impacts négatifs devraient être favorisés par rapport aux choix visant à compenser ou à indemniser les personnes potentiellement affectées pour les préjudices qu'ils ont subis.

Compte tenu de l'ensemble des informations obtenues à ce stade du processus HUDERIA, il existe une possibilité de réexaminer les questions zéro. Cette information pourrait également être utile pour informer les discussions visant à savoir si les activités du cycle de vie du système d'IA en question (en cours de développement ou déjà utilisé en cas de révision itérative) sont conformes aux droits humains, à la démocratie et à l'État de droit.

Accès à des voies de recours

Les mesures visant à traiter des impacts négatifs ne se limitent pas aux recours juridiques. Ces impacts peuvent être traités à l'aide d'autres mesures d'atténuation telles que celles définies dans les politiques, les orientations ou d'autres instruments.

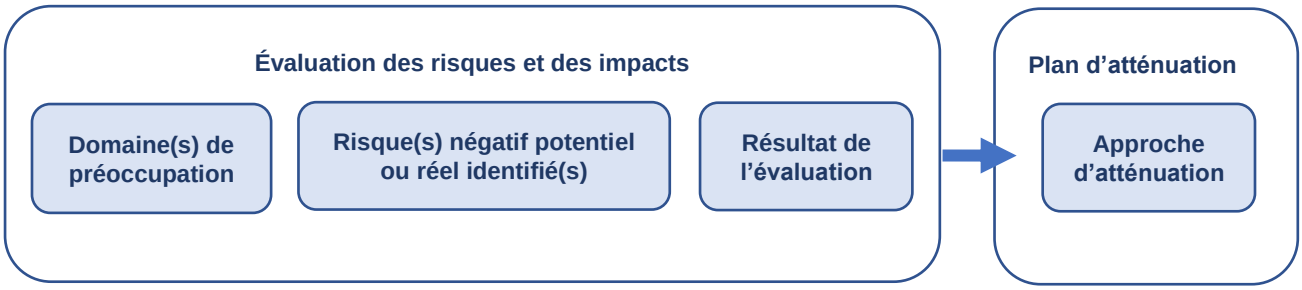
Lors de la mise en place de telles mesures, les sujets suivants devraient être abordés :

a) Si des mesures et mécanismes de responsabilité existantes sont en place en ce qui concerne les droits humains, de la démocratie et de l'État de droit. Il est essentiel que ces cadres existants soient appliqués au contexte des activités du cycle de vie des systèmes d'intelligence artificielle ;
b) la complexité technique, de la nature axée sur les données et de l'opacité relative de certains systèmes d'IA peuvent limiter leur transparence. Il peut en résulter une importante asymétrie dans l'accès, la compréhension ou le contrôle de l'information entre les différentes parties impliquées dans le cycle de vie du système d'IA. Des mesures visant à documenter et à fournir des informations sur le système d'IA et ses impacts aux personnes affectées peuvent faciliter la mise en place et l'accessibilité de voies de recours effectifs en cas d'impacts négatifs sur ces personnes ;
c) les informations fournies dans le cadre de ces mesures doivent être adaptées au contexte, claires et utiles, afin que les personnes puissent les utiliser efficacement pour exercer leurs droits dans les procédures liées aux décisions qui les concernent;
d) le cas échéant, il peut être nécessaire de fournir aux personnes affectées d'autres garanties et sauvegardes procédurales effectives, conformément au droit international et national applicable.

Résultat de cet élément du processus HUDERIA

Cet élément devrait permettre de décrire de manière claire les mesures et les actions visant à traiter les risques et les impacts identifiés, ainsi que les rôles et les responsabilités des différents acteurs impliqués dans l'atténuation, la gestion et le suivi. Le cas échéant, cet élément devrait également fournir un aperçu accessible des mécanismes et mesures de réparation à la disposition des personnes impactées.

En outre (voir la section sur la révision itérative ci-dessous), un plan est établi pour le suivi des efforts d'atténuation et pour la réévaluation itérative de ces efforts au cours des phases ultérieures du cycle de vie du système d'IA.



Révision itérative

Introduction

La réalisation de l'HUDERIA au début du cycle de vie d'un système d'IA n'est qu'une première étape, bien qu'essentielle, d'un processus d'évaluation et de réévaluation responsables plus long. Le processus de révision itérative garantit que l'évaluation des risques et des impacts demeure effective tout au long du cycle de vie du système d'IA. Il s'agit d'un processus continu qui offre, à intervalles réguliers, l'occasion d'identifier de nouveaux impacts et de mettre à jour le Plan d'atténuation.

Il est probable que les impacts du système d'IA évolueront au fil du temps, soit en raison de décisions prises en phase de production et de mise en œuvre du système d'IA, des applications contextuelles, soit en raison de changements externes dans l'environnement réel. Ces changements, qui peuvent concerner le cycle de vie des données, le développement et la conception du système d'IA, les processus de passation de marché, l'évolution des techniques d'IA, l'intégration ou l'opérationnalisation du système, les failles de sécurité, ainsi que les événements ou occurrences significatifs entraînant des conséquences préjudiciables ou involontaires, peuvent influencer les performances du système d'IA et/ou son impact sur les personnes et les groupes affectés.

De telles modifications appellent une révision destinée à s'assurer que les droits humains, la démocratie et l'État de droit sont respectés en permanence, tout au long du cycle de vie du projet d'IA. Une attention particulière devrait être accordée à la manière dont ces changements affectent la performance du système et son impact sur les personnes et les communautés.

Facteurs liés à la production, à la mise en œuvre et au déploiement

Les choix faits à tout moment pendant le cycle de vie du système ainsi que les évènements se déroulant lord du déploiement du système peuvent nécessiter la révision des décisions et évaluations antérieures, en particulier celles résultant du processus HUDERIA, d'où la nécessité d'une réévaluation, d'un réexamen et d'une révision.

Ces changements, en particulier ceux concernant le cycle de vie des données, le développement et la conception du système d'IA, les changements dans les techniques d'IA, l'intégration ou l'opérationnalisation du système, les vulnérabilités en matière de sécurité, ainsi que des évènements ou occurrences significatifs tels qu'ayant des conséquences préjudiciables ou inattendues, peuvent influencer la performance du système d'IA et/ou son impact sur les personnes et les groupes affectés. Les processus d'un modèle d'IA sont itératifs et fréquemment non linéaires, nécessitant des révisions et mises à jour fréquentes, le cas échéant.

Facteurs liés à l'environnement réel

Les changements qui surviennent dans les contextes sociaux, réglementaires, politiques ou juridiques durant lesquels le système est en cours de production ou d'utilisation peuvent avoir des effets sur la qualité de fonctionnement du système d'IA et sur la manière dont il impacte les droits des personnes et des groupes affectés.

De même, les réformes réglementaires et politiques, ainsi que les changements dans les méthodes d'enregistrement des données peuvent survenir au sein de la population concernée d'une manière qui affecte la question de savoir si les données utilisées pour entraîner le

modèle décrivent de manière correcte les phénomènes, les populations ou les facteurs connexes.

Dans le même ordre d'idées, des changements culturels ou comportementaux peuvent survenir au sein des populations affectées, ce qui modifie la distribution des données et entrave la performance d'un modèle, lequel a été entraîné sur des données recueillies avant ces changements. Toutes ces modifications des conditions contextuelles peuvent avoir un effet significatif sur la performance du système d'IA et sur la manière dont celui-ci agit impacte les personnes, groupes, communautés affectées et la société en général.

Mise en œuvre de la révision itérative

Si la Méthodologie HUDERIA offre une certaine flexibilité en ce qui concerne les modalités exactes, les seuils, les déclencheurs, les mécanismes de suivi et de gouvernance du processus de révision itérative, les principes suivants pourraient être pris en compte :

a) la révision continue de l'HUDERIA joue un rôle essentiel dans le maintien de son efficacité et de sa fiabilité ;

b) un plan est établi pour suivre les impacts et pour réévaluer l'HUDERIA à chaque phase du cycle de vie du projet jusqu'au retrait ou à la mise hors service du système ;

c) les processus utilisés pour la révision itérative devraient rester aussi réactifs que possible à la manière dont le système d'IA interagit avec ses environnements opérationnels et avec les personnes impactées (par exemple, les domaines d'application possibles du système d'IA, l'émergence de nouvelles formes d'utilisation abusive du système, etc.) ;

d) dans des contextes qui évoluent rapidement ou qui changent, il peut être nécessaire de procéder à des réévaluations plus fréquentes.