CHARTES ÉTHIQUES ENCADRANT LES SYSTÈMES D'INTELLIGENCE ARTIFICIELLE: ANALYSE EXPLORATOIRE D'OUTILS ÉTHIQUES

Par **Michelle Albert-Rochette**, étudiante à la maitrise en droit, Université Laval
Sous la supervision de **Lyse Langlois** (Université Laval) et **Yves Boisvert** (ÉNAP)

Document de travail, version 4
Février 2022
ISBN : 978-2-925138-09-9

OBSERVATOIRE INTERNATIONAL
SUR LES IMPACTS SOCIÉTAUX
DE L'IA ET DU NUMÉRIQUE

INTRODUCTION

Au cours des dernières années, de nombreuses lignes directrices éthiques pour un développement responsable de l'intelligence artificielle (IA) ont été publiées par des acteurs en provenance d'une variété de secteurs. Ces lignes directrices se sont articulées autour d'un certain nombre de valeurs et de principes éthiques. Plusieurs publications ont évalué et comparé ces lignes directrices pour en faire émerger les convergences[1] et les lacunes[2]. Le présent rapport a une visée plus pragmatique qui va au-delà de l'identification et de la définition de principes ou de valeurs. Il s'inscrit dans la formalisation de l'éthique en faisant ressortir les façons de mettre en application, de façon concrète, les principes et les valeurs éthiques dégagés dans la littérature et en mettant en évidence les manières d'évaluer la mise en œuvre de ces derniers. Le rapport s'intéresse aussi à la formalisation de l'éthique des données vu la dépendance qu'entretient l'IA avec celles-ci.

Le rapport sert de document de travail et est divisé en deux sections :

SECTION A : Recension des modèles de développement éthique de l'intelligence artificielle

Cette section résume 19 modèles portant sur le développement éthique de l'IA ou sur l'utilisation responsable des données. Les modèles retenus s'inscrivent dans la formalisation de l'éthique, en ce qu'ils fournissent des recommandations pratiques relatives à la mise en œuvre de principes ou de valeurs éthiques, ou en ce qu'ils guident la réflexion de manière à permettre le repérage et la gestion de problèmes éthiques. Les modèles sont présentés 4 catégories : modèles gouvernementaux, modèles privés, modèles d'associations et d'organismes sans but lucratif (OSBL) et modèles académiques. Le contenu intégral de plusieurs des modèles résumés se trouve dans les annexes de cette section[3].

[1] Jessica Fjeld et al, "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI", Berkman Klein Center for Internet & Society, (2020), En ligne : <https://dash.harvard.edu/handle/1/42160420>; Anna Jobin, Marcello Ienca et Effy Vayena, "The global landscape of AI ethics guidelines" (2019) 1:9 Nat Mach Intell 389-399; Thilo Hagendorff, " The Ethics of AI Ethics: An Evaluation of Guidelines" (2020) 30:1 Minds Mach 99-120; Thilo Hagendorff, "The Ethics of AI Ethics: An Evaluation of Guidelines" (2020) 30:3 Minds Mach 457-461 (correction de la *Table 1* publiée dans 30:1 Minds Mach 99-120).

[2] Notamment : Thilo Hagendorff, *supra* note 1.

[3] Voir : *Annexes de la section A*. Le contenu de ces annexes est directement extrait des publications originales.

SECTION B : Chartes éthiques encadrant les systèmes d'intelligence artificielle

Cette section utilise la grille analytique conçue par Yves Boisvert (2021)[4] pour classer et analyser le contenu de 13 chartes éthiques visant à encadrer les systèmes d'IA. La grille est divisée en 4 catégories qui permettent d'évaluer la mise en œuvre des valeurs ou des principes éthiques dégagés dans les chartes retenues.

DOCUMENTS DE RÉFÉRENCE

Des documents de référence sont listés à la fin du rapport. Parmi ceux-ci se trouve le tableau ***Sites et publications recensant les outils et lignes directrices pour le développement éthique de* l'IA**[5]. Ce tableau a été mis à jour le 15 juin 2021 et vise à faciliter le suivi des derniers développements en matière de formalisation de l'éthique de l'IA.

OBJECTIFS

Le premier objectif était de sélectionner et de résumer des modèles prometteurs de développement éthique de l'IA ou d'utilisation responsable des données ayant dépassé les stades d'identification et de définition de principes ou de valeurs éthiques et offrant donc une formalisation de l'éthique. Il s'agissait de mettre en évidence les modèles qui proposent des recommandations concrètes pour intégrer les principes ou les valeurs éthiques, ou qui aident à repérer et à gérer des problèmes éthiques dans le développement, le déploiement ou l'utilisation de systèmes d'IA ou dans l'utilisation des données. La section A du rapport (*Recension des modèles de développement éthique de l'intelligence artificielle*) est consacrée à cet objectif.

Le second objectif était de proposer une matrice d'analyse de chartes éthiques encadrant les systèmes d'IA.

Le troisième objectif était de tester différentes chartes éthiques s'inscrivant dans la formalisation de l'éthique à l'aide de la matrice d'analyse proposée.

La section B du rapport (*Chartes éthiques encadrant les systèmes d'intelligence artificielle*) est consacrée aux second et troisième objectif.

[4] Voir : ANNEXE Q – Matrice de formalisation retenue – Grille analytique (Yves Boisvert, 2021).
[5] Voir : *Documents de référence*, à la page 107.

MÉTHODOLOGIE

L'élaboration du rapport s'est faite en trois étapes successives en adéquation avec les trois objectifs :

1. De septembre 2020 à décembre 2020, une recension des écrits a été effectuée. Cette recension était axée sur la sélection de modèles de développement éthique de l'IA ou d'utilisation responsable des données qui s'inscrivaient dans la formalisation de l'éthique. Les modèles retenus ont été résumés dans la section A du rapport (*Recension des modèles de développement éthique de l'intelligence artificielle*).

2. En janvier 2021, une grille analytique[6] a été proposée par Yves Boisvert. La grille sert de matrice d'analyse et est divisée en 4 catégories qui permettent d'évaluer la mise en œuvre des principes ou des valeurs éthiques dégagés dans les chartes éthiques :

    a. **Risques, enjeux ou dilemmes éthiques** (propres à un principe ou à une valeur éthique);
    b. **Meilleures pratiques** (pour gérer les risques, enjeux ou dilemmes éthiques identifiés);
    c. **Entraves** (à la mise en œuvre des meilleures pratiques);
    d. **Stratégies** (pour dénouer les impasses créées par les entraves).

3. De janvier 2021 à juin 2021, le contenu de 13 chartes éthiques portant sur l'encadrement de systèmes d'IA a été classé dans la grille conçue par Yves Boisvert (2021). La sélection des publications s'est faite à partir de deux critères d'inclusion cumulatifs et de cinq critères d'exclusion. La sélection n'est pas exhaustive. Les publications ont été classées par pays. Une même grille peut contenir deux chartes éthiques en provenance du même auteur[7].

[6]Voir : ANNEXE Q – Matrice de formalisation retenue – Grille analytique (Yves Boisvert, 2021).

[7]C'est la raison pour laquelle on retrouve les noms de 10 chartes éthiques dans la table des matières de la section B (*Chartes éthiques encadrant les systèmes d'intelligence artificielle*), mais 13 chartes éthiques classées et analysées dans les grilles. Par exemple, la grille pour l'Australie contient les informations tirées de *Australia's Ethics Framework: A Discussion Paper,* Gouvernement de l'Australie (2019) et de *Australia's AI Ethics Framework,* Gouvernement de l'Australie (2020).

Pour la section B du rapport, ont été <u>retenues</u> les publications :

a. Identifiant clairement plusieurs principes ou valeurs éthiques pour le développement, le déploiement ou l'utilisation de l'IA
   et
b. Proposant des recommandations concrètes clairement identifiées ou identifiables[8] pour mettre en œuvre ces principes ou valeurs éthiques ou pour remédier aux enjeux, risques ou dilemmes éthiques y étant associés.

Pour la section B du rapport, ont été <u>exclues</u> les publications :

a. Publiées par des entreprises;
b. Traitant uniquement d'enjeux, de risques ou de dilemmes éthiques sans proposer de recommandations pour gérer ces derniers;
c. Dont les recommandations sont trop générales ou abstraites;
d. Dont les recommandations relèvent uniquement du domaine technique;
e. Portant seulement sur un domaine précis d'utilisation de l'IA.

Chaque charte éthique retenue a été décortiquée de manière à en classer le contenu dans les cases appropriées de la grille proposée par Yves Boisvert (2021). Dans certains cas, des éléments auraient pu à la fois être classés dans la catégorie *Risques* et dans la catégorie *Entraves*. Dans le cadre du présent rapport, la différence de classement entre les deux relève de l'ordre et de la façon dont l'information était présentée dans les publications. Par exemple, si une publication abordait d'emblée le problème des « boîtes noires »[9], cet élément était classé dans *Risques* et les solutions correspondantes étaient placées dans *Meilleures pratiques*. Toutefois, si une publication présentait ce problème comme un frein à l'application de mesures d'explicabilité qui avaient été préalablement présentées, le problème des « boîtes noires » était plutôt classé dans *Entraves*, et les solutions correspondantes dans *Stratégies*. Par ailleurs, dans l'optique d'assurer une classification fidèle du contenu des chartes éthiques retenues, certaines cases des grilles peuvent

[8] Dans plusieurs publications, les recommandations n'étaient pas identifiées comme telles dans une section distincte, mais la formulation du texte permettait de conclure qu'il s'agissait de recommandations. Par exemple, une formulation du type « les organisations *devraient* faire … » était considérée comme une recommandation clairement identifiable.

[9] Terme faisant référence aux modèles d'apprentissage automatique dont la nature complexe rend difficilement explicables les fonctions, les processus d'apprentissage et les résultats. Par exemple : difficulté à savoir quelles données d'entrée ont contribué à une décision prise par un système d'IA.

reprendre textuellement des parties de celles-ci. Des titres et des sous-titres ont été ajoutés dans les cases des grilles pour faciliter la lecture. Plusieurs abréviations détaillées au début de la section B ont été utilisées. Lorsqu'une même grille contient le contenu de deux chartes éthiques du même auteur, un code de couleur précisé en légende dans le coin supérieur gauche indique la provenance du contenu classé. Finalement, plusieurs mots en anglais ont été conservés pour faire référence à des termes techniques spécifiques ou pour simplifier le repérage dans les publications retenues.

TABLE DES MATIÈRES DU RAPPORT :

A. RECENSION DES MODÈLES DE DÉVELOPPEMENT ÉTHIQUE DE L'INTELLIGENCE ARTIFICIELLE

[Septembre 2020 – décembre 2020]

1. Modèles gouvernementaux

1.1 Artificial Intelligence Ethics Framework for the Intelligence Community (*ODNI*, 2020)

Office of the Director of National Intelligence. "Artificial Intelligence Ethics Framework for the Intelligence Community", (2020), En ligne : <https://www.dni.gov/files/ODNI/documents/AI_Ethics_Framework_for_the_Intelligence_Community_10.pdf>.[10]

Ce guide met de l'avant les 10 objectifs qu'une technologie utilisant l'IA devrait respecter. Chaque objectif est accompagné d'une liste de questions guidant sa réalisation.

Les 10 objectifs que l'IA devrait respecter :

1. Be used when it is an appropriate means to achieve a defined purpose after evaluating the potential risks;
2. Be used in a manner consistent with respect for individual rights and liberties of affected individuals, and use data obtained lawfully and consistent with legal obligations and policy requirements;
3. Incorporate human judgment and accountability at appropriate stages to address risks across the lifecycle of the AI and inform decisions appropriately;
4. Identify, account for, and mitigate potential undesired bias, to the greatest extent practicable without undermining its efficacy and utility;
5. Be tested at a level commensurate with foreseeable risks associated with the use of the AI;
6. Maintain accountability for iterations, versions, and changes made to the model;
7. Document and communicate the purpose, limitation(s), and design outcomes;
8. Use explainable and understandable methods, to the extent practicable, so that users, overseers, and the public, as appropriate, understand how and why the AI generated its outputs;
9. Be periodically reviewed to ensure the AI continues to further its purpose and identify issues for resolution;
10. Identify who will be accountable for the AI and its effects at each stage and across its lifecycle, including responsibility for maintaining records created.

Les 10 objectifs et des exemples de questions correspondantes	
Objectifs	**Questions**
1. Purpose: Understanding Goals and Risks	• What benefits and risks, including risks to civil liberties and privacy, might exist when this AI is in use? Who will benefit? Who or what will be at risk?

[10] Document associé : Office of the Director of National Intelligence. *Principles of Artificial Intelligence Ethics for the Intelligence Community* (2020), En ligne : https://www.dni.gov/files/ODNI/documents/Principles_of_AI_Ethics_for_the_Intelligence_Community.pdf.

2. Legal Obligations and Policy Considerations Governing the AI and the Data	• How must data be stored, shared, retrieved, accessed, used, retained, disseminated, and dispositioned under the authority/agreement/contract, as well as relevant constitutional, statutory, and regulatory provisions?
3. Human Judgment and Accountability	• How may introducing an accountable human produce cognitive biases and/or confirmation bias?
4. Mitigating Undesired Bias and Ensuring Objectivity	• How complete are the data on which the AI will rely? Are they representative of the intended domain? How relevant is the training and evaluation data to the operational data and context? How does the AI avoid perpetuating historical biases and discrimination?
5. Testing Your AI	• Has the AI been evaluated for potential biased outcomes or if outcomes cause an inappropriate feedback loop? Have you considered applicable methods to make the AI more robust to adversarial attacks?
6. Accounting for Builds, Versions, and Evolutions of an AI	• As you refine the AI, how does the data you have used, the parameters and weights you have chosen, and the outputs ensure that this version or evolution is designed to achieve the authorized purpose?
7. Documentation of Purpose, Parameters, Limitations, and Design Outcomes	• Have you documented where the data came from and its downstream uses and sharability? The downstream uses and sharability of the AI?
8. Transparency: Explainability and Interpretability	• How are outputs marked to clearly show that they came from an AI?
9. Periodic Review	• How will the accountable human(s) address changes in accuracy and precision due to either an adversary's attempts to disrupt the AI or unrelated changes in the operational/business environment, which may impact the accuracy of the AI?
10. Stewardship and Accountability: Training Data, Algorithms, Models, Outputs of the Models, Documentation	• Who is ultimately responsible for the decisions of the AI and is this person aware of the intended uses and limitations of the analytic?

Voir ***Annexe A*** pour l'ensemble des questions associées aux 10 objectifs.

1.2 Data Ethics Framework (*Gouvernement du Royaume-Uni,* 2020)

Gouvernement du Royaume-Uni. "Data Ethics Framework", (2020), En ligne : *GOV.UK* <https://www.gov.uk/government/publications/data-ethics-framework>.

Résumé :

Le modèle vise une utilisation responsable des données. Il se divise en 3 principes éthiques et en 5 actions spécifiques. Des grilles d'auto-évaluation permettent d'attribuer un score variant entre 0 et 5 pour noter et suivre la progression de l'atteinte des 3 principes éthiques et des 5 actions spécifiques.

	Principles	Score 0	1	2	3	4	5
■	Transparency						
▲	Accountability						
●	Fairness						

	Specific actions	Score 0	1	2	3	4	5
1	Define public benefit and user need						
2	Involve diverse expertise						
3	Comply with the law						
4	Check the quality and limitations of the data						
4.1	Check the quality and limitations of the model						
5	Evaluate and consider wider policy implications						

Les 5 actions spécifiques :

Les 5 actions spécifiques se déclinent en plusieurs actions plus précises et/ou considérations éthiques, dont certaines visent directement à renforcer la mise en œuvre des 3 principes. Elles sont accompagnées de questions qui permettent de vérifier leur mise en œuvre.

Exemple :

Par exemple, l'action spécifique *Define and understand public benefit and user need* comprend l'action précise *Make your user need en public benefit transparent*, laquelle est directement liée à la mise en œuvre du principe de transparence. La question suivante permet de vérifier l'application de l'action précise : *How have you shared your understanding of the user need with the user?*

1.3 The Assessment List for Trustworthy Artificial Intelligence (ALTAI) (Groupe d'experts indépendants de haut niveau sur l'IA, 2020)

Groupe d'experts indépendants de haut niveau sur l'IA. "ALTAI - The Assessment List on Trustworthy Artificial Intelligence", (2020), En ligne : <https://futurium.ec.europa.eu/en/european-ai-alliance/pages/altai-assessment-list-trustworthy-artificial-intelligence>.

ALTAI

En 2019, le Groupe d'experts indépendants de haut niveau de la Commission européenne publiait les *Lignes directrices en matière d'éthique pour une IA digne de confiance*. Ce document comprenait la première version de la *Liste d'évaluation pour une IA digne de confiance* (ci-après : la *Liste*). La *Liste*, mise à jour en juillet 2020, permet d'évaluer la mise en œuvre de 7 exigences éthiques à l'aide d'une banque de questions spécifiques à chacune d'elles.

Les 7 principes directeurs pour une IA digne de confiance et des exemples de questions correspondantes

Principes	Questions *[1]La dernière version des questions de la* **Liste** *est seulement disponible en anglais.*
1. Action humaine et contrôle humain	• Could the AI system affect human autonomy by interfering with the end-user's decision-making process in any other unintended and undesirable way?
2. Robustesse technique et sécurité	• Did you consider whether the AI system's operation can invalidate the data or assumptions it was trained on, and how this might lead to adversarial effects?
3. Respect de la vie privée et gouvernance des données	• Depending on the use case, did you establish mechanisms that allow flagging issues related to privacy concerning the AI system?
4. Transparence	• In cases of interactive AI systems (e.g., chatbots, robo-lawyers), do you communicate to users that they are interacting with an AI system instead of a human?
5. Diversité, non-discrimination et équité	• Did you establish a strategy or a set of procedures to avoid creating or reinforcing unfair bias in the AI system, both regarding the use of input data as well as for the algorithm design?
6. Bien-être sociétal et environnemental	• Did you take measures that ensure that the AI system does not negatively impact democracy?
7. Responsabilité	• Did you establish mechanisms that facilitate the AI system's auditability (e.g. traceability of the development process, the sourcing of training data and the logging of the AI system's processes, outcomes, positive and negative impact)?

Voir ***Annexe B*** pour l'ensemble des questions associées aux 7 principes.

ALTAI TOOL

Un outil d'évaluation interactif a été développé à partir de la *Liste.* Un questionnaire semi-automatisé est proposé en ligne. Les questions réfèrent aux 7 principes éthiques. À la fin du questionnaire, un score entre 0 et 5 est attribué selon le degré de mise en œuvre des principes. Des recommandations sont aussi proposées. **L'outil interactif est disponible au** : https://altai.insight-centre.org/.

2. Modèles d'entreprises

2.1 A Framework for Systematically Applying Humanistic Ethics when Using AI as a Design Material (*Dent et al – PARC*, 2020)

Dent, Kyle, Richelle Dumond et Mike Kuniavsky, *A Framework for Systematically Applying Humanistic Ethics when Using AI as a Design Material*, SSRN, (Rochester, NY: Social Science Research Network, 2019), En ligne : <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3455518>.

Description

L'article est publié par des chercheurs du centre de recherche en informatique PARC[11]. Ils proposent un cadre d'examen éthique (aussi appelé *design methodology*) pour aider les concepteurs à identifier les effets négatifs d'une technologie d'IA résultant de leur travail et à réfléchir à des façons d'atténuer ces effets. Les risques identifiés doivent être communiqués à toutes les parties prenantes à un projet d'IA, y compris les utilisateurs. Les auteurs suggèrent que les effets potentiels d'une technologie d'IA devraient être considérés par un ensemble de questions, et non simplement par le biais d'une liste de normes éthiques. Le contexte d'utilisation propre à chaque technologie doit être pris en compte. Les auteurs proposent la mise en place d'un comité d'experts pouvant aider les concepteurs à considérer les enjeux éthiques dans leurs projets.

5 principes ont guidé la réflexion des auteurs pour l'élaboration du cadre d'examen éthique :

1. *Respect for individuals*
2. *Safety and security of person*
3. *Respect for individuals' right to privacy, protection of data, freedom of expression and participation in cultural life*
4. *Equality and fairness*
5. *Benefits to society must not come at the cost of others*

Contenu du cadre d'examen éthique :

Le cadre est divisé en deux sections : une liste de vérifications préliminaires (*Preliminary Checklist*) et des directives détaillées (*Comprehensive Guidelines*).

1) Liste de vérifications préliminaires (*Preliminary Checklist*)

But : éviter une évaluation éthique trop lourde lorsque la technologie ne le requière pas en évaluant sommairement les effets négatifs potentiels.

Exemples de questions :

- When the AI learns from existing data, ask yourself :
  Does the data contain individual personal attributes (especially protected attributes such as race, sex, gender identity, ability, status, socio-economic status, education level, religion, country of origin)? **Yes / No**

  Is it possible that any members of vulnerable populations (this could be any disadvantaged subsegment of an overall population, e.g. children, prisoners, refugees, people facing discrimination) might be recorded? **Yes / No**

[11] Palo Alto Research Center, Californie.

2) Directives détaillées *(Comprehensive Guidelines*)

Répondre « Oui » à au moins une question de la liste de vérifications préliminaires indique que les directives détaillées doivent être examinées.

But : par l'entremise d'un ensemble de questions et de recommandations correspondantes, susciter la discussion par rapport aux effets négatifs potentiels d'une technologie d'IA. Toutes les questions ne sont pas pertinentes selon la technologie visée. 6 thèmes sont couverts : *human data, social impact, environmental impact, physical interaction*, *miseuse or malicious intent* et *post deployment*. Certaines des questions ont été inspirées des principes de l'organisation FAT/ML.

Voir ***Annexe C*** pour l'ensemble des questions et recommandations.

Les 6 thèmes et des exemples de questions et recommandations correspondantes	
Thèmes	**Questions et recommandations**
Human data	***Will the benefits of your design extend to the entire population or could there be subgroups who are inadvertently excluded?*** In instances where there is a known risk of discrimination, include details of the application and its functions along with samples of the data and details about the data source. Is there a process to take effective action to mitigate discrimination?
Social impact	***AI decisions that could bias or alter societal norms*** Are the decisions produced by an algorithmic system explainable to the people affected by those decisions? Explanations help with the validation of results and build trust in the system.
Miseuse or malicious intent	***Are there guardrails in place to prevent off-label usage (the intentional or accidental misuse of the design)?***
Environmental impact	***Consider the energy impact of the computational resources necessary for the project. Is there a way to minimize that impact? Can the computation be handled differently to use less energy?*** Do the resources provisioned for the project match the requirements, or are they too excessive?
Physical interaction	***Does the design minimize potential risk?***
Post deployment	***Is there a plan for what to do if the project has unintended consequences?*** This may be part of a maintenance plan and should involve post-launch monitoring plans.

Autres considérations pertinentes de l'article

Pratiques de *design* éthiques	Il est impossible d'éviter complètement les conséquences négatives ou inattendues qu'une technologie peut avoir sur les individus qui interagissent avec elle, mais il est possible de mettre en place des pratiques de conception éthiques (*ethical design practices*) pour anticiper et atténuer ces conséquences.
AI Design governance model	Les listes de contrôle éthiques (*checklists*) sont insuffisantes. Elles ne vont pas assez loin et le développement éthique des systèmes d'IA ne passe pas seulement par l'élaboration de normes, mais aussi par la mise en place de processus, notamment ceux relatifs à l'examen de la conception (*design review process*) et à la collecte et transparence des données (*data collection process*). Un modèle de gouvernance pour la conception de l'IA (*AI Design governance model*) est nécessaire.
***Design* éthique et utilisateurs**	Le *design* éthique doit faire partie du processus de conception d'une technologie d'IA dès le départ. Les auteurs présentent l'éthique de l'IA comme un matériau de conception. Des équipes diversifiées sont nécessaires pour une technologie d'IA éthique. La conception d'un projet ne doit pas seulement se concentrer sur la convivialité d'une interface, mais doit aussi inclure les expériences des personnes par rapport auxquelles la technologie prendra des décisions. Les minorités doivent être considérées dans la conception – si elles sont perçues comme des cas d'exception, elles seront *de facto* marginalisées par la technologie.

2.2 Everyday Ethics for Artificial Intelligence (*IBM*, 2019)

IBM. "Everyday Ethics for Artificial Intelligence", (2019), En ligne : <https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf>.

Le document vise l'intégration de l'éthique dans la conception et le développement des technologies d'IA. Il se concentre sur 5 domaines d'intérêt éthique : *Accountability, Value Alignment, Explainability, Fairness* et *User Data Rights*.

Chaque domaine est abordé au moyen de 3 sections : recommandations pratiques (*Recommended actions to take*), actions à considérer (*To consider*) et questions à examiner (*Questions for your team*).

Pour le domaine ***Fairness***, une section présente la liste des principaux biais inconscients à considérer dans la conception et le développement d'une technologie d'IA.

Exemple :

Explainability		
Recommended actions to take	***To consider***	***Questions for your team***
Allow for questions. A user should be able to ask why an AI is doing what it's doing on an ongoing basis. This should be clear and up front in the user interface at all times.	Ensure an AI system's level of transparency is clear. Users should stay generally informed on the AI's intent even when they can't access a breakdown of the AI's process.	Which segments of our AI decision processes can be articulated for users in an easily digestible and explainable fashion?

Voir ***Annexe D*** pour les 4 autres domaines d'intérêt éthique identifiés.

2.3 Guidelines for Human-AI Interaction (*Amershi et al – Microsoft*[12]*,* 2019)

Référence de l'article : Amershi, Saleema et al, "Guidelines for Human-AI Interaction" (2019), En ligne : <https://www.microsoft.com/en-us/research/publication/guidelines-for-human-ai-interaction/>.

Informations supplémentaires et synthèse graphique : https://www.microsoft.com/en-us/research/project/guidelines-for-human-ai-interaction/

Les auteurs proposent 18 lignes directrices relatives à la conception (*design*) pour l'interaction humain-IA[13]. Le but est de centrer la conception sur l'humain (*human-centered design*) et de renforcer la confiance des utilisateurs. Les 18 lignes directrices représentent une synthèse de plus de 150 recommandations de conception liées à l'IA provenant de sources académiques et industrielles. Ces lignes directrices ont été validées par plusieurs cycles d'évaluation, dont une étude d'utilisateurs auprès de 49 concepteurs (*design practicionners)* les ayant testées sur 10 catégories de produits d'IA. Cette étude a permis aux auteurs de reformuler certaines lignes directrices sur la base des commentaires reçus.

Les lignes directrices ne doivent pas être utilisées comme une simple liste de vérifications, mais doivent plutôt permettre aux personnes qui les appliquent de réfléchir à d'autres enjeux potentiels. Certains cas justifieront de pondérer les lignes directrices, d'en ajouter d'autres ou d'en retirer.

Les lignes directrices sont divisées en 4 sections : *Initially, During interaction, When wrong* et *Over time*.

[12] **Autres initiatives de Microsoft pour le développement responsable de l'IA :**

Responsible bots : 10 guidelines for developers of conversational bots (novembre 2018) : **voir section 2.5**.

Fairness Checklist (mars 2020) : liste de contrôle détaillée pour le principe éthique de *Fairness* conçue en collaboration avec 48 praticiens de manière à répondre à leurs besoins.

Référence de l'article **:** Madaio, Michael et al, "Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI" (2020), En ligne : <https://www.microsoft.com/en-us/research/publication/co-designing-checklists-to-understand-organizational-challenges-and-opportunities-around-fairness-in-ai/>.

Lien de la *Fairness Checklist* seule : https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE4t6dA

[13] ***Partnership on AI***

Partneship on AI propose un cadre sur ce même thème composé de 36 questions pour comprendre la relation entre l'humain et les techniques d'IA de manière à mieux répondre aux enjeux de transparence, de confiance (*Trust*) de responsabilité et d'autonomie. Le cadre est divisé en 4 sections : *Nature of Collaboration, Nature of situation, AI System Characteristics et Human Characteristics*. Référence : Partnership on AI, "Human-AI Collaboration Framework & Case Studies", (2019), En ligne : *The Partnership on AI* <https://www.partnershiponai.org/human-ai-collaboration-framework-case-studies/>.

Exemple :

Sections	Lignes directrices d'IA
Initially	**Make clear how well the system can do what it can do**. Help the user understand how often the AI system may make mistakes.
During interaction	**Make clear how well the system can do what it can do**. Help the user understand how often the AI system may make mistakes.
When wrong	**Make clear why the system did what it did**. Enable the user to access an explanation of why the AI system behaved as it did.
Over time	**Encourage granular feedback**. Enable the user to provide feedback indicating their preferences during regular interaction with the AI system.

Voir ***Annexe E*** pour l'ensemble des lignes directrices.

2.4 Responsible AI by Design in Practice (*Benjamins* et *al* – *Telefonica*, 2019)

Benjamins, Richard, Alberto Barbado et Daniel Sierra, "Responsible AI by Design in Practice" (2019) arXiv:190912838 [cs], En ligne : <http://arxiv.org/abs/1909.12838>.

Les auteurs présentent le modèle de leur entreprise (Telefonica) pour une IA responsable. Le modèle comporte des principes éthiques et une méthodologie pour leur mise en œuvre (*Responsible AI by Design*). Les principaux défis rencontrés pour la mise en œuvre des principes et les sujets nécessitant des recherches supplémentaires sont discutés par les auteurs. Le modèle s'adresse aux organisations privées qui prévoient utiliser l'IA.

Les 4 principes sélectionnés

1) *Fair AI*
2) *Transparent and Explainable AI*
3) *Human-centric AI*
4) *Privacy and Security by Design*

La méthodologie de mise en œuvre des principes – *Responsible AI by Design*

La méthodologie comprend 5 « ingrédients » :

1) *The AI principles setting the values and boundaries*

2) *A set of questions and check points, ensuring that all AI principles have been considered in the creation process*
3) *Tools that help answering some of the questions, and help mitigating any problems identified*
4) *Training, both technical and non-technical*
5) *A governance model assigning responsibilities and accountabilities*

Les principes, les questions et les stratégies correspondantes (« ingrédients » 1-2-3-4)

Principes	Questions	Mise en œuvre via
Fair AI	Does your data set contain sensitive variables?	***Training**** Explain what is sensitive personal data as defined by law.
	Does any of the variables strongly correlate with sensitive variables?	***Technical tool**** Check correlations between all variables in data set and vizualize result.
Transparent and Explainable AI	Could the user think that s/he interacts with a person rather than with your system?	***Training*** This may require building in certain features during design.
	Is it possible to understand how the algorithm had reached its conclusions?	***Technical tool*** Define what transparency level is required according to the profile of the users of the system and use an available solution to generate local or global explanations.
Human-centric AI	Is there a possibility that your P&S has a negative impact on Human Rights?	***Training*** General training on Human Rights and examples of P&S that impact them in both positive and negative ways.
Privacy and Security by Design	Does your AI system use personal data?	***Training*** Explain what is personal data, psydonymized data and anonymous data.
	Is the system robust against attacks that seek to exploit weaknesses in it and manipulate the outputs?	***Technical tool*** Before passing a model tu production, it is recommended to assess its vulnerabilities to these attacks with the tools available.

***Training :** puisque l'utilisation de l'IA dans les organisations est récente, un nouveau « matériel de formation » propre aux questions de transparence et d'explicabilité doit être développé. Pour les questions non spécifiquement liées aux enjeux d'IA, il est possible de se référer au matériel de formation existant dans l'organisation.

***Technicals tools :** des outils techniques sont nécessaires pour garantir une IA explicable. Les auteurs en proposent dans *Table 3* et sous les rubriques *Fair AI, Transparent and Explainable AI* et *Privacy and Security by Design* (p. 4-5-6).

Voir ***Annexe F*** pour l'ensemble des questions et les façons de les mettre en œuvre.

Le modèle de gouvernance

A governance model assigning responsabilities and accountabilities

- Le modèle de gouvernance est « agile » et vise à changer la culture pour créer une IA responsable dès la conception. Il mise sur la sensibilisation sans imposer trop de contrôle.
- L'organisation devrait faire une campagne de sensibilisation pour expliquer ce qu'est l'IA, pourquoi et comment elle est utilisée dans l'entreprise et quels sont les défis qui y sont rattachés. Les notions suivantes devraient être introduites : principes, méthodologie, programme de formation et outils.

- Les personnes les plus proches de la conception et du développement des produits et services d'IA devraient être formées en premier. La formation pourrait ensuite être étendue au reste de l'organisation.
- La formation est technique et non technique selon le cas.
- La responsabilité est le plus possible déléguée aux personnes responsables des produits et des services concernés par l'utilisation de l'IA.
- Le modèle est utilisé pour ce qui relève de l'IA. Pour les questions de confidentialité et de sécurité (*privacy and security manners*), les modèles de gouvernance existants au sein d'une entreprise peuvent être utilisés.

2.5 Responsible bots: 10 guidelines for developers of conversational AI (*Microsoft*, 2018)

Microsoft, "Responsible bots: 10 guidelines for developers of conversational AI" (2018), En ligne : <https://www.microsoft.com/en-us/research/publication/responsible-bots/>.

L'article présente 10 lignes directrices pour des *conversational bots* responsables et qui renforcent la confiance des utilisateurs. Les lignes directrices s'articulent autour des thèmes suivants : éthique, sécurité, sûreté, inclusion, transparence et responsabilité (*accountability*).

Les 10 lignes directrices et des exemples d'explications et de recommandations spécifiques :

Lignes directrices	Recommandations et explications spécifiques
1) Articulate the purpose of your bot and take special care if your bot will support consequential use cases	Assess whether the bot's intended purpose can be performed responsibly.
2) Be transparent about the fact that you use bots as part of your product or service	It should be apparent to the user that they are not having an interaction with another person.
3) Ensure a seamless hand-off to a human where the human-bot exchange leads to interactions that exceed the bot's competence	Respect individual engagement preferences, particularly if your bot deals in consequential matters.
4) Design your bot so that it respects relevant cultural norms and guards against misuse	To reduce the possibility of conflicting with those values and cultural norms, limit the surface area for norms violations. For example, if your bot is designed to take pizza orders, limit it to that purpose only, so that it does not engage on topics such as race, gender, religion, politics and the like.
5) Ensure your bot is reliable	Provide a feedback mechanism. Users will feel more comfortable with bots if they can provide feedback on their operation (and feedback is essential in any event, as with all product development work). Bots should actively ask for feedback. Set expectations as to whether the user will get any response to feedback provided.

6) Ensure your bot treats people fairly	Systematically assess the data used for training your bot.
7) Ensure your bot respects user privacy	Collect no more personal data than you need, limit access to it and store it for no longer than needed.
8) Ensure your bot handles data securely	Ensure the integrity of your training data. All AI systems must be able to distinguish between maliciously introduced data (which must be purged) and data that is merely rare, yet valid and potentially important.
9) Ensure your bot is accessible	Have people with disabilities test your bots.
10) Accept responsibility	Developers are accountable for the bots they deploy. If you are developing a bot that your organization will deploy, you should recognize that you are fully responsible for its operation and how it affects people.

Voir ***Annexe G*** pour l'ensemble des lignes directrices

2.6 The Box: Dynamics of AI principles (*AI Ethics Lab,* 2020)

AI Ethics Lab. "Tool: The Box", (2020), En ligne : *Toolbox: Dynamics of AI Principles* <http://aiethicslab.com/the-box/>.

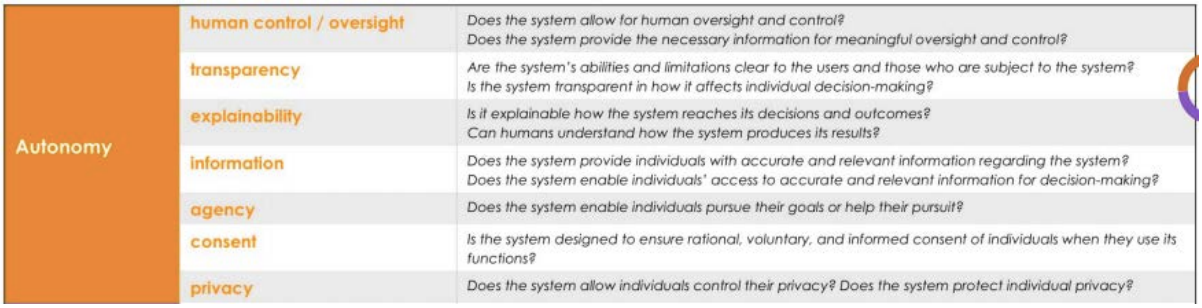
L'outil se divise en 3 principes éthiques fondamentaux (*Autonomy, Harm-Benefit, Justice*). Ces principes se déclinent en « principes instrumentaux », lesquels sont accompagnés de questions permettant d'évaluer leur mise en œuvre.

Un score (*None, Low, Medium, High*) est attribué à chaque « principe instrumental ».

THE BOX – by AI Ethics Lab

		none	minimum	medium	maximum	N/A
Autonomy	human control / oversight					
	transparency					
	explainability					
	information					
	agency					
	consent					
	privacy					
Harm-Benefit	accuracy / reliability					
	security					
	safety					
	well-being					
	impact					
	efficiency					
Justice	distribution of burden & benefit					
	equality / non-discrimination					
	protecting the vulnerable					
	accountability					
	contestability					

Exemples de questions pour le principe *Autonomy* :



Autonomy	human control / oversight	*Does the system allow for human oversight and control?* *Does the system provide the necessary information for meaningful oversight and control?*
	transparency	*Are the system's abilities and limitations clear to the users and those who are subject to the system?* *Is the system transparent in how it affects individual decision-making?*
	explainability	*Is it explainable how the system reaches its decisions and outcomes?* *Can humans understand how the system produces its results?*
	information	*Does the system provide individuals with accurate and relevant information regarding the system?* *Does the system enable individuals' access to accurate and relevant information for decision-making?*
	agency	*Does the system enable individuals pursue their goals or help their pursuit?*
	consent	*Is the system designed to ensure rational, voluntary, and informed consent of individuals when they use its functions?*
	privacy	*Does the system allow individuals control their privacy? Does the system protect individual privacy?*

Voir ***Annexe H*** pour l'ensemble des questions associées à chaque principe et pour un exemple d'interprétation des résultats.

3. Modèles d'associations et d'OSBL

3.1 AI Ethics Framework (*Digital Catapult*, 2018)

Digital Catapult. "AI Ethics Framework", (2018), En ligne : *Machine Intelligence Garage* <https://www.migarage.ai/ethics/ethics-framework/>.

Digital Catapult est une agence gouvernementale britannique pour l'innovation dans les industries numérique et logicielle.

Le cadre éthique se compose de 7 concepts, lesquels sont accompagnés d'une liste de questions permettant de guider leur mise en application pratique. Les concepts se rapprochent de ceux développés par le *Groupe d'experts indépendants de haut niveau sur l'intelligence artificielle* de la Commission européenne de 2019 et des principes de l'OCDE et de Beijing sur l'IA.

Les 7 concepts et des exemples de questions correspondantes	
Concepts	**Questions**
1. **Clear benefits**	• How can the products or services be monitored and tested to ensure they meet [the] goals, purposes and intended applications?
2. **Know and manage the risks**	• How can it be known whether a bias has been created or reinforced with the system?
3. **Use data responsibly**	• Have potential biases in the data been examined, well-understood and documented and is there a plan to mitigate against them?

4.	**Be worthy of trust**	• Who is accountable if things go wrong? Are they the right people? Are they equipped with the skills and knowledge they need to take on this responsibility?
5.	**Diversity, Equality and inclusion**	• Are potential biases in the data and processes are examined, well-understood and documented and is there a plan to mitigate against them?
6.	**Transparent communication**	• Does the company communicate clearly, honestly and directly about any potential risks of the product or service being provided?
7.	**Business model**	• Do users have a clear idea of how the data will be used, including any future linking/sale of the data?

Voir ***Annexe I*** pour l'ensemble des questions associées aux 7 concepts.

3.2 AI Ethics Label (*AIEI Group, Association VDE*, 2020)

AIEI. "AI Ethics Impact Group: From Principles to Practice – VDE", (2020), En ligne : <https://www.ai-ethics-impact.org/en>.

Le *AI Ethics Label* :

FIGURE 1 **The AI Ethics Label with six selected values**

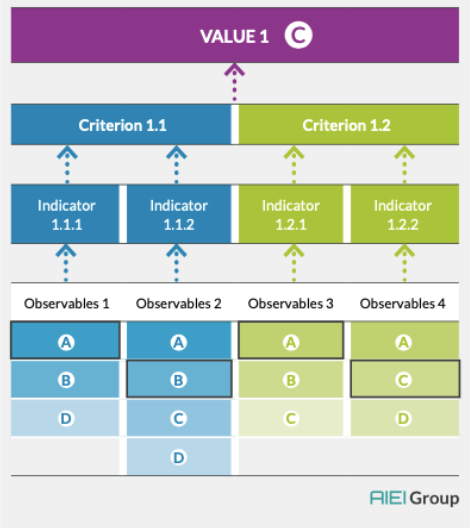
Le modèle[14] vise à attribuer des notes (A, B, C, D, etc.) à 6 valeurs éthiques de manière à avoir une représentation visuelle globale des caractéristiques éthiques propres à un système utilisant l'IA.

L'approche VCIO (valeurs, critères, indicateurs et éléments observables) permet d'attribuer ces notes.

L'approche VCIO :
L'atteinte d'une **valeur** est définie selon certains **critères**. Des **indicateurs** (sous forme de questions) permettent de vérifier l'atteinte de ces critères. Les éléments **observables**, associés à des notes (A, B, C, D, etc.), quantifient dans quelle mesure les indicateurs sont atteints.

La note attribuée à un élément observable devient celle de l'indicateur correspondant. En prenant la note la plus basse des indicateurs associés à un critère, on obtient la note pour le critère en question. En prenant la note la plus basse des critères caractérisant une valeur, on obtient la note propre à la valeur éthique visée, laquelle est transposée sur le AI Ethics Label.

Voir ***Annexe J*** pour un exemple d'attribution de note à la valeur *Transparency* sur le AI Ethics Label à l'aide de l'approche VCIO.



Conflits de valeurs : un conflit peut survenir lorsque qu'une valeur et ses indicateurs ne peuvent être réalisés qu'en brimant une autre valeur. Ces conflits sont pris en compte dans l'approche d'évaluation : ils peuvent être résolus en fonction du contexte d'application de la technologie en question, et des points de vue du régulateur ou de l'utilisateur. Les façons de résoudre les conflits de valeurs sont présentées à la page 18 de la publication.

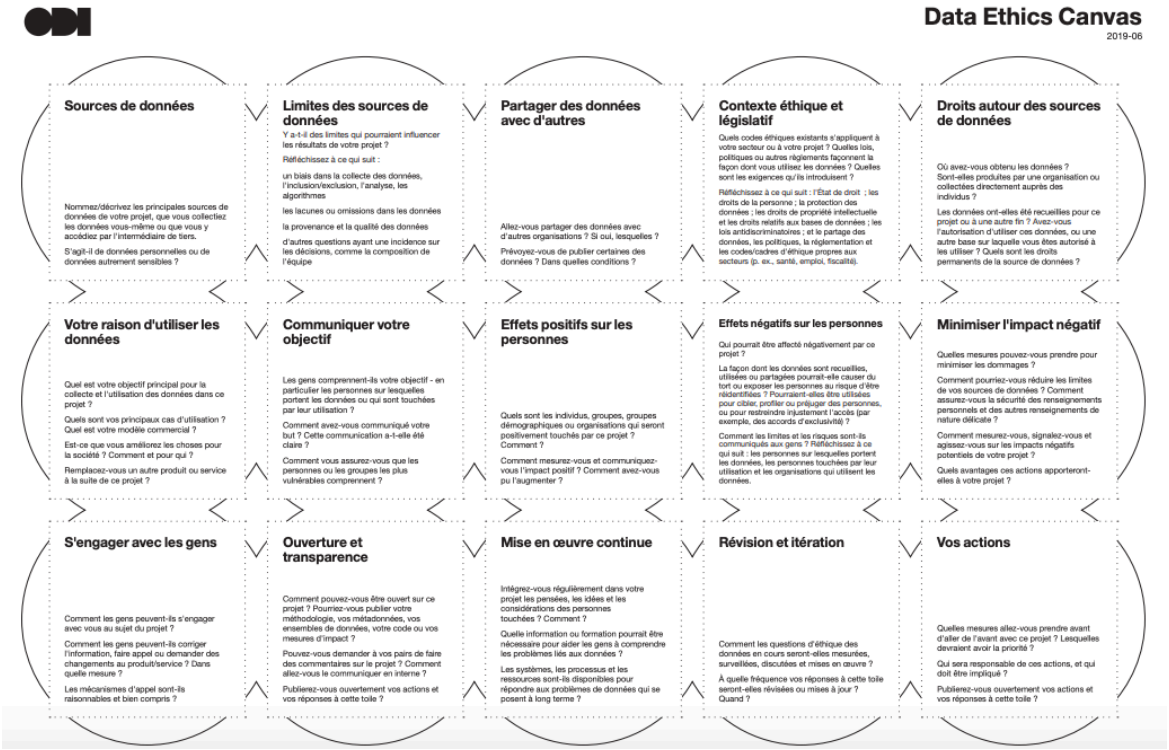
[14] Note : les auteurs proposent aussi une **matrice de risques** qui complète le modèle et qui n'est pas détaillée dans le cadre du présent rapport. Voir ***Annexe J*** et consulter plus spécifiquement les pages 35-40 de la publication *AI Ethics Impact Group: From Principles to Practice – VDE* (2020).

3.3 Data Ethics Canvas[15] (*Open Data Institute*, 2019)

Open Data Institute. "The Data Ethics Canvas", (2019), En ligne : <https://theodi.org/article/data-ethics-canvas/>.

Le *Open Data Institute* est un organisme sans but lucratif basé au Royaume-Uni.

Le modèle prend la forme d'une toile divisée en 15 sections permettant d'identifier et d'aborder les questions éthiques d'un projet utilisant des données[16]. Par exemple, la section « Limites des sources de données » invite à réfléchir aux biais dans la collecte des données et à la provenance de ces dernières. La section « Effets négatifs sur les personnes » propose de réfléchir aux personnes touchées par l'utilisation des données et aux organisations qui les utilisent, et la section « S'engager avec les gens » s'intéresse à la manière dont les personnes peuvent corriger l'information ou demander des changements relatifs au produit ou au service qu'elles utilisent.



Voir ***Annexe K*** pour la version originale du modèle (version 2019).

[15] **Une version mise à jour du modèle a été publiée en 2021**. Voir : https://theodi.org/article/the-data-ethics-canvas-2021/#1563365825519-a247d445-ab2d.

[16] Voir aussi l'initiative de ***Dataethics.eu*** qui a développé des principes éthiques, des directives et un court questionnaire pour l'intégration de l'éthique dans un projet impliquant des données. Les principes de *Dataethics* : *The Human being at the centrer, Individual Data control, Transparence and Explainability, Accountability, Equality*. Consulter : Dataethics, "Data Ethics Principles" (2017), En ligne : <https://dataethics.eu/data-ethics-principles/> et : Dataethics.eu, "Data Ethics Impact Assessment" (2021), En ligne : <https://dataethics.eu/wp-content/uploads/dataethics-impact-assessment-2021.pdf>.

3.4 Éthique et numérique : un référentiel pratique pour les acteurs du numérique (Cigref et Syntec Numérique, 2018)

Cigref et Syntec Numérique, « Éthique & numérique : un référentiel pratique pour les acteurs du numérique », (2018), En ligne : *Cigref* <https://www.cigref.fr/ethique-numerique-un-referentiel-pratique-pour-les-acteurs-du-numerique>.

Le Cigref (*Club informatique des grandes entreprises françaises*) et Syntec Numérique proposent un cadre de référence pratique axé sur l'éthique du numérique qui se destine aux entreprises et à leurs partenaires. Il vise à les outiller et à les sensibiliser aux enjeux éthiques du numérique en catégorisant les questions éthiques pertinentes des points de vue des utilisateurs et des concepteurs, et en fournissant des solutions pratiques.

Cadre de référence : questions éthiques

Le cadre de référence propose 3 questions préliminaires et des questions et recommandations éthiques plus spécifiques propres à 3 catégories de l'éthique du numérique.

Les 3 questions préliminaires :

1) L'éthique du numérique fait-elle l'objet d'un traitement spécifique dans l'entreprise (y a t-il des comités dédiés, des programmes de sensibilisation portant exclusivement sur l'éthique du numérique, de l'IA etc.)?
2) L'éthique du numérique fait-elle partie des enjeux de gouvernance globale de la transformation numérique?
3) Le sujet de l'éthique du numérique est-il défi ni clairement et différencié des sujets de conformité / du travail des directions juridiques?

Les 3 catégories de l'éthique du numérique

1) L'éthique *by design*
    - S'intéresse à la phase de conception
    - Vise particulièrement les développeurs, les designers numériques et les chefs de projets.

2) L'éthique des usages
    - Se questionne sur les usages du numérique faits notamment par les utilisateurs, les collaborateurs et les partenaires.

3) L'éthique sociétale
    - Se questionne sur les impacts du numérique sur la société.

Des exemples de questions et de recommandations spécifiques à chaque catégorie :

Catégories	Thèmes	Questions	Recommandations
Éthique *by design*	Protection de la vie privée et des données personnelles	La corrélation de données issues de diverses sources induit-elle la production d'informations personnelles (dans le cadre de projets big data et d'IA notamment) ?	Mettre en place un dispositif qui mesure la personnalisation des données à la sortie du traitement
	Éthique algorithmique et IA	La logique de fonctionnement des algorithmes déployés en intelligence artificielle peut-elle être expliquée ?	Avoir une politique d'explicabilité des systèmes, sur l'ensemble de la chaîne (provenance des données, explication du raisonnement suivi)
Éthique d'usage	Éthique avec les utilisateurs	Des moyens sont-ils proposés aux utilisateurs de services personnalisés pour gérer leurs paramètres ?	Donner la possibilité aux utilisateurs de paramétrer facilement la gestion de leurs données personnelles, et d'opérer des choix éclairés
	Éthique partenariale	Existe-t-il une politique permettant de vérifier, dans l'assemblage des solutions numériques entre divers partenaires, que le process dans son ensemble est éthique ?	Faire appel à des tiers de confiance, certifications et/ou labels, démontrant l'engagement éthique de chaque partie prenante
Éthique sociétale	Empreinte écologique et sociétale des solutions	L'impact sociétal des projets (origine des matériaux, bonnes pratiques des partenaires, etc.) est-il pris en compte ?	Réaliser une étude d'impact sociétal des projets
	Économie de l'attention et bulles informationnelles	Les risques de biais cognitifs humains sont-ils pris en compte dans la conception des solutions numériques ?	S'assurer que les applications et solutions numériques n'ont pas été conçues de manière à manipuler volontairement l'utilisateur par l'exploitation de biais cognitifs

Voir ***Annexe L*** pour l'ensemble des questions et des recommandations spécifiques à chaque catégorie.

3.5 Responsible AI Design Assistant (*AI Global*[17], 2020)

Référence de l'outil : AI Global, "Responsible AI Design Assistant"[18], (2020), En ligne : <https://oproma.github.io/rai-trustindex/>.

Informations supplémentaires :

Casovan, Ashley, "Creating a Responsible AI Trust Index: A unified assessment to assure the Responsible Design, Development, and Deployment of AI", En ligne : *AI Global* <https://ai-global.org/2020/04/28/creating-a-responsible-ai-trust-index-a-unified-assessment-to-assure-the-responsible-design-development-and-deployment-of-ai/>.

AI Global, "Responsible AI Design Assistant_Use case testing guidev1 - Google Slides", (2020), En ligne : <https://docs.google.com/presentation/d/1EDPhyRhIsiOrujLcHQv_fezXfgOz4Rl7a8lyOM_guoA/edit#slide=id.p1>.

Description :

L'outil consiste en un questionnaire à compléter en ligne qui permet d'aider les personnes qui conçoivent, développent et mettent en œuvre des systèmes utilisant l'IA à le faire de manière responsable en anticipant plus facilement les risques éthiques liés à 5 dimensions d'une IA fiable (*trusted AI program*) : *Accountability*, *Explainability and Interpretability*, *Data Quality*, *Bias and Fairness* et *Robustness*. L'outil peut être utilisé à toutes les phases d'un projet et cible les grandes comme les petites entités. Il est accompagné d'un guide de l'utilisateur qui comprend des définitions et des informations supplémentaires utiles pour compléter l'évaluation. Le *Responsible AI Design Assistant* a été conçu par un groupe multidisciplinaire d'experts comprenant notamment des ingénieurs, des éthiciens, des décideurs publics, des experts en sécurité et des avocats.

Contexte :

Étant donné l'abondance de principes éthiques mis de l'avant ces dernières années et la multiplication de stratégies pour gérer le développement de l'IA de manière responsable, les auteurs de l'outil souhaitaient mettre en place une évaluation facile d'utilisation intégrant les meilleurs principes, politiques et pratiques. Ils mentionnent avoir élaboré la liste de questions de l'outil en compilant le contenu des documents les plus cités et référencés (livres blancs et publications d'académiques, d'organisations de normalisation, de gouvernements, de groupes de réflexion et d'entreprises du secteur privé).

[17] **AI Global a changé de nom en 2021 pour *Responsible Artificial Intelligence Institute* (RAII)**. Le lien vers le nouveau site web de l'organisation est le suivant : https://www.responsible.ai/.

[18] Des modifications ont été apportées à l'outil en 2021. Désormais, **6 dimensions éthiques** d'une IA fiable sont examinées : *Organization Maturity, Accountability, Data, Fairness, Interpretability, Robustness*.
Le lien vers la nouvelle version de l'outil est le suivant : https://designassistant.responsible.ai/.

Exemple de questions relatives aux 5 dimensions d'une IA fiable[19] :
Voir ***Annexe M*** pour la liste complète des questions.

Les 5 dimensions d'une IA fiable et des exemples de questions		
Dimensions d'une IA fiable	**Questions**	**Choix de réponses**
Accountability	**Is there a process in place for determining if the automated activity will be flagged for human oversight?**	• System is only used to assist a decision maker • System is replacing a decision that would otherwise be made by a human and no judgement or discretion is required • System is replacing a decision that would otherwise be made by a human and judgement or discretion is required
Bias and Fairness	**Are the objectives of the system clear to the users?**	• Yes, objectives are clear • Yes, clear documentation about these objectives have been provided to the intended user • No
	How is the privacy and intimacy of an individual or group protected in both the development and implementation of your system?	• The collection of private information only takes place if a user consents or for authorized surveillance purposes. • Data related to personal thoughts and emotions are not used in situations where the system could case harm, especially in circumstances where moral judgements (eg. lifestyle choices) could be made. • Users are able to disconnect or stop sharing information with the system at any point in time. • Only users have the ability to set profile preferences, changes to the these preferences can easily be done at any time. • Access to personal information is limited to only individuals who require it for the direct functioning of the system. • Individuals have the ability to access their personal data including, but not limited to, the collection, use, and sharing of this data at any time. • Individuals have the ability to donate their personal data to research organizations. • Data integrity is assured. The system does not use private data to imitate or alter a persons appearance, voice, or other individual characteristics in order to damage one's reputation or manipulate other people. • The system does not curtail people's real or percieved liberty.
Explainability and Fairness	**Is it possible to discover how your system renders a decision or performs a function?**	• System is transparent, it is possible to know for certain how and why the system made a particular decision, or in the case of a robot, acted the way it did. • System is opaque, it is possible through post-decision or post-action, to analyze through various processes (eg. counterfactual or repeatability testing) to draw an accurate conculsion on how the decision was made or the action was taken. • System is a black box, but detailed records of the design processes and decision making have been kept throughout the entire process. • System is a black box, it is possible to make best guesses on why a decision was rendered or a decision was taken, but it's not certain.

[19] Selon les dimensions éthiques considérées dans la version 2020 de l'outil.
Le lien vers la version 2021 de l'outil est le suivant : https://designassistant.responsible.ai/.

Robustness	**What safeguards have you put in place to ensure your system is robust enough to deal with the edge cases and extreme scenarios (eg. load inputs, adversarial attacks) to adequately mitigate erroneous outcomes to the best extent possible?**	• None • Have ensured that there is a mitigation plan in place for any individual, group, or organization who has an incentive to make the system misbehave. • The unintended consequences resulting in a mistake from the system have been assessed and mitigated to the best extent possible. • A rigorous threat model to understand all possible attack vectors has been implemented. • The system has been tested against adversarial attacks. • Third party advesarial testing of the system was completed. • Ongoing research is being conducted to ensure the latest tools are being applied. • Preventative and precautionary measures have been taken.
Data Quality	**How is the data being collected, used, and stored being managed?**	• Through a third-party data service, terms and conditions are unknown. • Through a third-party data service, terms and conditions ensure users data should be protected from theft, misuse, or data corruption

4. Modèles académiques

4.1 An Ethical Framework for Good AI Society: Opportunities, Risks, Principles, and Recommendations (*Floridi et al,* 2018)

Floridi, Luciano et al. “AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations” (2018) 28:4 Minds & Machines 689–707.

Les auteurs proposent 5 principes éthiques devant guider le développement de l’IA. Ces principes résultent d’une synthèse de 47 principes préalablement mis de l’avant par des organisations réputées et multipartites. Les 4 premiers principes réfèrent aux principes traditionnels de bioéthique. Les auteurs proposent le nouveau principe *Explicability*, propre à l’IA, pour intégrer à la fois les notions d’intelligibilité et de responsabilité.

Les 5 principes : *Beneficience, Non-maleficience, Autonomy, Justice* et *Explicability*.

*Ces 5 principes sont associés aux 7 facteurs éthiques mis en évidence dans *Floridi et al* (2020) (voir section 4.5).

4.2 Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications (*Ryan* et *Stahl*, 2020)

Ryan, Mark et Bernd Carsten Stahl. "Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications" (2020), Journal of Information, Communication and Ethics in Society, En ligne : <https://doi.org/10.1108/JICES-12-2019-0138>.

Les auteurs compilent l'ensemble des exigences normatives éthiques existantes de manière à fournir des instructions et des conseils pouvant guider la conception et le développement éthique de l'IA. Pour ce faire, ils s'appuient sur 10 principes éthiques dégagés après l'analyse de 91 ensembles de lignes directrices éthiques. Ces 10 principes se déclinent en plusieurs catégories, lesquelles sont accompagnées des exigences normatives existantes. Les auteurs identifient la provenance de ces exigences normatives (ex : *IEEE*, *AI Now Institute*, *Internet Society*, etc.).

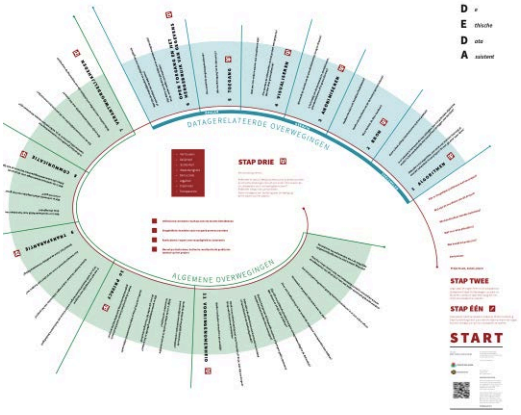
Les 10 principes éthiques et des exemples d'exigences normatives existantes	
Principes	**Exigences normatives et provenance**
1. Transparency ***Transparency, explainability, explicability, understandability, interpretability, communication, disclosure, showing***	AI organisations should document how their AI makes certain decisions and be able to reproduce them for audits (SIIA, 2017).
2. Justice and fairness ***Justice, fairness, consistency, inclusion, equality, equity, non-bias, non-discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge***	There should be close attention paid to the training data used, potential human biases and bias derived from the results of algorithmic processes (Cerna Collectif, 2018).
3. Non-maleficience ***Non-maleficience, security, safety, harm, protection, precaution, prevention, integrity***	The effects of [the] systems must be reviewed on an ongoing basis (Algo.Rules, 2019).
4. Responsability ***Responsibility, accountability, liability, acting with integrity***	There needs to be clear and concise allocation of responsibilities within the organisation using AI, and the creation of potential scenarios and ways to deal with harms when they occur (EGE 2018; FATML, 2016).
5. Privacy ***Privacy, personal and private information***	Users should have control and access to data stored about them (IEEE, 2019).
6. Beneficience ***Benefits, beneficience, well-being, peace, social good, common good***	AI organisations should use data retrieved for the benefit of their customers and society (OP, 2019).
7. Freedom and autonomy ***Freedom, autonomy, consent, choice, self-determination, liberty, empowerment***	The use of personal data must be clearly articulated and agreed upon before its use (UNDG, 2017).
8. Trust ***Trustworthiness***	Organisations can cultivate trust by demonstrating the security of their AI (Intel, 2017) and guard the data retrieved from these systems in a responsible way (Unity Blog, 2018).
9. Sustainability ***Sustainability, environment (nature), energy, resources (energy)***	Resource use and environmental impact should be held in importance in the life cycle impact assessment of AI (COMEST/UNESCO, 2017).

10. Dignity	AI needs to be developed and used in a way that makes it clear to the user that they are interacting with AI and not another human being (EGE, 2018).
11. **Solidarity** ***Solidarity, social security, cohesion***	Democratic values should not be jeopardised as a result of AI use and citizens should receive accurate and impartial information without interference or manipulation for political purposes (EGE, 2018).

4.3 DEDA : Data Ethics Decision Aid (*Utrecht Data School* et *Université d'Utrecht*, 2020)

Utrecht Data School et Université d'Utrecht. "Data Ethics Decision Aid (DEDA)", (2020), En ligne : *Utrecht Data School* <https://dataschool.nl/en/deda/>.

Le modèle permet de repérer les problèmes éthiques dans les projets impliquant des données dans le but de faire une utilisation et une gestion responsables des données et des algorithmes. Sous forme de spirale, il met en évidence différentes étapes à compléter successivement. Le centre de la spirale marque la fin du parcours.



Le modèle se compose de deux principales sections. La première, ***Data related considerations*** se divise en trois sous-sections : *Collecting*, *Using* et *Storing*. Chaque sous-section aborde des thèmes différents, ces derniers étant accompagnés de questions permettant de réfléchir aux considérations éthiques pertinentes. La deuxième section, ***General considerations***, s'intéresse à d'autres thèmes, eux aussi accompagnés d'une liste de questions.

Section 1 : Data related considerations		
Sections	**Thèmes**	**Exemples de questions correspondantes**
Collecting	Algorithms Source	Is there someone within the team that can explain how the algorithm in question works?
Using	Anonymization Visualization	Should the data be anonymized, pseudonymized or generalized?
Storing	Access Sharing, reusing and repurposing	Who has access to the data and under what conditions?

Section 2 : General considerations	
Thèmes	**Exemples de questions correspondantes**
Responsibility	Who is ultimately responsible for the project?

Communication	What communication strategies are there for cases in which something goes wrong, and who is responsible for them?
Transparency	Do citizens have the opportunity to raise objections to the results of the project?
Privacy	Do the data provide insight into the personal lives of citizens?
Bias	Is there a risk that your project could contribute to discrimination?

Voir ***Annexe N*** pour l'ensemble des questions propres à chaque thème.

4.4 FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity (*Arnold et al*, 2019)

Arnold, Matthew et al, "FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity" (2019) arXiv:180807261 [cs], En ligne : <http://arxiv.org/abs/1808.07261>.[20]

Inspirés par la pratique des déclarations de conformité du fournisseur (*SDoC – supplier's declarations of conformity*), les auteurs proposent l'idée des fiches d'informations (*FactSheets on AI services*) comme un mécanisme permettant d'accroitre la confiance dans les services d'IA par une plus grande transparence. Ces fiches contiennent des informations sur l'utilisation prévue, les performances, la sûreté, la sécurité et la provenance des données de la technologie d'IA. Les fiches incluent aussi des informations sur la manière dont la technologie a été créée, entraînée et déployée, les scénarios dans lesquels elle a été testée, la manière dont elle répond à des scénarios non testés et des directives spécifiant les tâches pour lesquelles elle ne devrait pas être utilisée.

Le contenu de la fiche d'information est modulable selon le contexte d'utilisation et se divise en plusieurs catégories, lesquelles sont associées aux « éléments de confiance de l'IA » identifiés par les auteurs (*Elements of Trust in AI Systems*) : *Statement of purpose*, *Basic performance*, *Safety, Security* et *Lineage*.

*Les pages 17-23 et 24-31 de l'article montrent deux exemples concrets de *FactSheets*.

[20] Des initiatives semblables ont été proposées par Gebru et al, *Datasheets for datasets* (2018); Bender et Friedman, *Data statements for NLP : Toward mitigating system bias and enabling better science* (2018); Holland et al, *The dataset nutrition label: A framework to drive higher data quality standards* (2018); Mitchell et al, *Model cards for model reporting* (2019). L'article de *Arnold et al* s'en distingue notamment en ne s'intéressant pas seulement à la question des données (niveau des composantes), mais au résultat final (niveau fonctionnel) de la technologie d'IA.

Éléments de confiance de l'IA et des exemples de questions :

Éléments de confiance de l'IA	Questions correspondantes
Statement of purpose	▪ What is the intended use of the service output? ▪ How is the input provided? By whom? How is the output returned?
Basic performance	▪ Which datasets was the service tested on? (e.g., links to datasets that were used for testing, along with corresponding datasheets) ▪ In addition to the service provider, was this service tested by any third party?
Safety	**General** ▪ Are you aware of possible examples of bias, ethical issues, or other safety risks as a result of using the service? ▪ Were the possible sources of bias or unfairness analyzed? ▪ Where do they arise from: the data? the particular techniques being implemented? other sources? ▪ Is there any mechanism for redress if individuals are negatively affected? **Explainability** ▪ Are the service outputs explainable and/or interpretable? **Fairness** ▪ For each dataset used by the service : Was the dataset checked for bias? What efforts were made to ensure that it is fair and representative?
Security	▪ List applications or scenarios for which the service is not suitable
Lineage	▪ Were the datasets used for training built-for-purpose or were they repurposed/adapted? Were the datasets created specifically for the purpose of training the models offered by this service?

Voir ***Annexe O*** pour les éléments d'une *FactSheet* et les questions correspondantes.

4.5 How to Design AI for Social Good : Seven Essential Factors (*Floridi et al*, 2020)

Floridi, Luciano et al, "How to Design AI for Social Good : Seven Essential Factors" (2020) 26 :3 Sci Eng Ethics 1771–1796, En ligne : <https ://link.springer.com/article/10.1007/s11948-020-00213-5>.

Résumé

En raison de la compréhension limitée de ce que constitue concrètement une IA « socialement bonne » (*AI for Social Good*, ci-après : AI4SG), les auteurs proposent des lignes directrices basées sur 7 facteurs éthiques essentiels et non hiérarchisés à considérer dans la conception (*design*) et le déploiement de l'IA. Ils identifient aussi les meilleures pratiques correspondantes. Certains des facteurs sont entièrement nouveaux parce que propres à l'IA. Chaque facteur est lié au principe éthique de *Beneficience* (perçu comme une condition préalable à l'AI4SG) et à au moins un des 4 autres principes éthiques de l'IA définis dans *Floridi et al (2018*) : *Nonmaleficence, Justice, Autonomy* et *Explicability*. Les 7 facteurs identifiés sont davantage liés à des considérations de *design* spécifiques à l'AI4SG et sont plus facilement opérationnalisables que les 5 principes généraux.

Les facteurs et les pratiques correspondantes ne doivent pas être intégrés à un projet en toutes circonstances, mais il est essentiel que chaque pratique identifiée soit proactivement considérée et qu'elle ne soit écartée que s'il existe une raison claire, démontrable et moralement défendable de la rejeter. Les 7 facteurs ne sont pas en eux-mêmes suffisants pour une AI4SG, mais l'examen attentif de chacun d'eux est requis.

La conclusion de l'article s'intéresse aux manières de balancer les 7 facteurs éthiques.

Les 7 facteurs éthiques et les principes éthiques correspondants :
*Le principe de *Beneficience* est lié à tous les facteurs.

1) Falsifiability and incremental deployment (***Nonmaleficience***)
2) Safeguards against the manipulation of predictors (***Nonmalificience***)
3) Receiver-contextualised intervention ***(Autonomy***)
4) Receiver-contextualised explanation and transparent purposes ***(Explicability***)
5) Privacy protection and data subject consent (***Nonmaleficience, autonomy***)
6) Situational fairness (***Justice***)
7) Human-friendly semanticisation (***Autonomy***)

Exemple : les 7 facteurs éthiques, un résumé de leur définition et les meilleures pratiques correspondantes :

Voir ***Annexe P*** pour un résumé des recommandations propres à chacun des 7 facteurs.

Facteurs éthiques	Description **Le degré d'application de chaque facteur varie selon le contexte*	Meilleures pratiques correspondantes
Falsifiability and	▪ La fiabilité (*trustworthiness*) est essentielle pour une AI4SG et la « falsiabilité » (*falsiability*) est nécessaire pour l'améliorer.	▪ Les concepteurs d'AI4SG devraient identifier les exigences falsifiables et les tester

incremental deployment	▪ La falsiabilité implique de pouvoir tester des « exigences critiques », c.à.d. des exigences sans lesquelles quelque chose ne pourrait ou ne devrait pas fonctionner. *EX : La sécurité est une exigence critique. En ce sens, elle doit être falsifiable sans quoi le système ne sera pas considéré comme fiable.*	progressivement du « laboratoire au monde extérieur » (selon un *incremental deployment*).
Safeguards against the manipulation of predictors	▪ Lié aux risques de la manipulation des données et du recours excessif à des indicateurs non causaux, c.à.d. le recours à des données qui sont *corrélées* à un phénomène mais qui ne sont pas *causales*. ▪ Ces risques peuvent mener à des résultats inéquitables qui violent le principe de *Justice*.	▪ Adopter des garanties (*safeguards*) qui (i) garantissent que les indicateurs non causaux ne biaisent pas indûment les interventions, et (ii) limitent, le cas échéant, la connaissance de la manière dont les intrants affectent les extrants des systèmes d'IA (pour éviter la manipulation des données).
Receiver-contextualised intervention	▪ Les systèmes d'IA doivent intervenir dans le respect de l'autonomie des utilisateurs.	▪ Les concepteurs devraient créer des systèmes de prise de décision (*decision-making systems*) en consultant les utilisateurs interagissant avec ces systèmes et qui sont impactés par eux; en tenant compte des caractéristiques des utilisateurs, des objectifs et des effets d'une *intervention*; et dans le respect des droits des utilisateurs d'ignorer ou de modifier les *interventions*.
Receiver-contextualised explanation and transparent purposes	▪ Étroitement lié à l'explicabilité et à la transparence. ▪ Les objectifs derrière une technologie devraient être transparents et ses opérations et résultats devraient être explicables. ▪ L'explication varie selon la notion que l'on tente d'expliquer et selon la personne à qui elle est adressée. ▪ Pour maximiser la confiance, l'explication devrait inclure des informations sur : le système en général, sa logique, et les raisons de sa décision. ▪ La transparence des objectifs est nécessaire, surtout dans des contextes où l'explicabilité des opérations et des résultats est difficile ou non souhaitable.	▪ L'explication doit remplir le but explicatif souhaité et être appropriée eu égard au système et aux personnes qui reçoivent l'explication. ▪ Les objectifs du système doivent être connus par défaut.
Privacy protection and data subject consent	▪ En lien avec le consentement des utilisateurs à l'utilisation de leurs données personnelles. ▪ Le niveau et le type de consentement varient selon le contexte (seuil de consentement).	▪ Les concepteurs doivent respecter le seuil de consentement établi pour le traitement des ensembles de données (*datasets*) de données personnelles.
Situational fairness	▪ Les données d'un système d'IA peuvent être biaisées, ce qui peut impacter la prise de décision d'une technologie d'IA et créer des résultats injustes pour certains individus. ▪ Dans les interactions homme-machine (ex : *chatbots*), l'utilisation de l'IA requière une compréhension par le système des groupes auxquels appartiennent les utilisateurs et de leurs caractéristiques.	▪ Les concepteurs devraient supprimer dans les ensembles de données les variables qui ne sont pas pertinentes pour un résultat, à moins que leur intégration soutienne l'inclusivité, la sécurité ou d'autres considérations éthiques. (Supprimer trop de données pourrait supprimer des facteurs importants pour l'inclusivité.)

Human-friendly semanticisation	▪ Certaines tâches méritent d'être déléguées à un système d'IA, d'autres non.	▪ Les concepteurs ne devraient pas empêcher les gens de « sémantiser » (c.à.d. de donner un sens à quelque chose et de le comprendre).

B. CHARTES ÉTHIQUES ENCADRANT LES SYSTÈMES D'INTELLIGENCE ARTIFICIELLE

Matrice d'analyse : grille analytique conçue
par Yves Boisvert, 2021

[Juin 2021]

Abréviations utilisées

App(s)	Application(s) d'IA
Dév.	Développement
DF	Droits fondamentaux
Fct	Fonctionnement
HITL	*Human-in-the-loop*
MEO	Mettre en œuvre / mise en œuvre
ML	*Machine learning*
P/r	Par rapport
SAAD	Systèmes d'acquisition et d'archivage des données personnelles
SIA	Système(s) d'intelligence artificielle
TX	Traiter / traitement

Cette section utilise comme matrice d'analyse la grille conçue par Yves Boisvert (2021)[21] pour tester différentes chartes éthiques encadrant les systèmes d'IA. La grille est divisée en 4 critères permettant d'analyser la mise en œuvre des principes ou des valeurs éthiques dégagés dans les chartes retenues :

a. **Risques, enjeux ou dilemmes éthiques** (propres à un principe ou à une valeur éthique);
b. **Meilleures pratiques** (pour gérer les risques, enjeux ou dilemmes éthiques identifiés);
c. **Entraves** (à la mise en œuvre des meilleures pratiques);
d. **Stratégies** (pour dénouer les impasses créées par les entraves).

Des notes complémentaires à certaines des chartes présentées se trouvent en ***Annexe R***.

[21] Voir : ANNEXE Q – Matrice de formalisation retenue – Grille analytique (Yves Boisvert, 2021).

1. PUBLICATIONS INTERNATIONALES

Légende Bleu : informations tirées des *Lignes directrices en matière d'éthique pour une IA digne de confiance,* Commission européenne (2019)	**1.1 Assessment List for Trustworthy Artificial Intelligence (ALTAI), Groupe d'experts de haut niveau sur l'IA, Commission européenne (2020)**						
	7 exigences pour une IA digne de confiance (fondées sur 4 principes éthiques)						
	1. Action humaine et contrôle humain (**Comprend : DF, action humaine et contrôle humain)*	**2. Robustesse technique et sécurité** (**Comprend : résilience aux attaques et sécurité, plans de secours et sécurité générale, précision, fiabilité et reproductibilité*)	**3. Respect de la vie privée et gouvernance des données** (**Comprend : respect de la vie privée, qualité et intégrité des données et accès aux données*)	**4. Transparence** (**Comprend : traçabilité, explicabilité et communication*)	**5. Diversité, non-discrimination et équité** (**Comprend : absence de biais injustes, accessibilité et conception universelle, et participation des parties prenantes*)	**6. Bien-être sociétal et environnemental** (**Comprend : durabilité et respect de l'environnement, impact social, société et démocratie*)	**7. Responsabilité** (**Comprend : auditabilité, réduction au minimum des incidences négatives et communication à leur sujet, arbitrages et recours*)
	Principes éthiques correspondants						
	Respect de l'autonomie humaine	**Prévention de toute atteinte**		**Explicabilité**	**Équité**	**Équité** **Prévention de toute atteinte**	**Équité**
Enjeux, risques ou dilemmes	**DROITS FONDAMENTAUX**[22] **Risque** : Entrave aux droits fondamentaux : *- Discrimination à l'endroit de certaines personnes;* *- Menace aux droits de l'enfant;* *- Atteinte au droit à la protection des données;* *- Atteinte à la liberté d'expression, d'information et d'association.*	**RÉSILIENCE AUX ATTAQUES ET SÉCURITÉ** **Risques :** **1.** Effets dommageables en cas de défauts de conception, de pannes, d'attaques ou d'utilisation inappropriée ou malveillante du SIA : *Ex :* *Altération de la prise de décision ou arrêt du SIA en cas d'attaque;* *Corruption des données du SIA si interventions malveillantes ou si*	**Risques** **1.** Atteinte à certains droits : - Droit à la vie privée; - Droit à l'intégrité physique, mentale et/ou morale; - Droit à la protection des données. **2.** Utilisation des données à des fins discriminatoires. **3.** Perte de confiance des utilisateurs résultant d'une telle utilisation.	**Risques :** **1.** Perte de confiance des utilisateurs si absence d'explication. **2.** Contestation d'une décision impossible si absence d'explication. **3.** Confusion sur la nature de l'interaction (*SIA se présentant comme un humain*). **4.** Reproduction d'erreurs si ignorance de la/des raison(s) pour	**Risques :** **1.** Discrimination et préjudices directs/indirects involontaires à l'encontre de certains groupes ou personnes en raison d'ensembles de données biaisés et/ou ou de la façon dont les SIA sont mis au point. Risque d'exacerbation des préjugés et de la marginalisation. **2.** Exploitation intentionnelle de préjugés.	**Risques** **1. Incidences sociales négatives résultant de l'utilisation des SIA** - Détérioration des compétences sociales; - Détérioration du bien-être physique et mental; - Modification fondamentale des milieux de travail et risque déqualification de la main-d'œuvre. **2. Impacts négatifs sur l'environnement**	**Risques** - Absence de mécanismes de responsabilisation pour répondre aux situations où des impacts injustes ou négatifs se produisent *Et donc :* - Absence de possibilité adéquate de réparation; - Incapacité de rendre compte des actions ou des décisions qui contribuent au résultat du SIA et de réagir

[22] Dans la ALTAI (2020), les droits fondamentaux font l'objet d'une section à part (*FRIA : Fundamental Rights Impact Assessment*) qui précède l'évaluation des 7 exigences pour une IA digne de confiance. Dans les *Lignes directrices en matière pour une IA digne de confiance* (2019), la section sur les droits fondamentaux se trouve sous le titre « Action humaine et contrôle humain ».

	ACTION HUMAINE ET AUTONOMIE **Risques :** **1.** Incidence du SIA sur le processus de prise de décision humaine. **2.** Manipulation du comportement des utilisateurs par le SIA via des mécanismes parfois difficilement détectables (exploitation de processus subconscients). **3.** Atteinte à l'autonomie si une décision entrainant des effets juridiques (ou autres effets d'importance considérable) est fondée exclusivement sur un TX automatisé. **4.** Confusion sur la nature de l'interaction (SIA ou personne). **5.** Dépendance excessive des utilisateurs envers le SIA et/ou création d'un attachement disproportionné envers le SIA. **CONTRÔLE HUMAIN** **Risque** : Surveillance inadéquate du SIA.	*exposition à des situations imprévues;* *Décisions erronées et/ou préjudices physiques en cas de procédures de sécurité insuffisantes.* **PRÉCISION** **Risques** : **1.** Conséquences dommageables résultant d'un faible niveau de précision – *ex : jugement incorrect, décisions ou recommandations erronées, survenance de risques imprévus.* **2.** Effets dommageables si un SIA invalide les données et les hypothèses (*assumptions*) à partir desquelles il a été entraîné. **3.** Perte de confiance des utilisateurs dans le SIA si leurs attentes en matière de performances du SIA ne correspondent pas à ses performances réelles. **FIABILITÉ ET REPRODUCTIBILITÉ** **Risques :** **1.** Conséquences dommageables en cas de faible fiabilité ou reproductibilité du SIA – *ex : préjudices involontaires*	**4.** Dysfonctionnement du SIA si les données collectées sont biaisées ou si elles comportent des erreurs. **5.** Modification du comportement du SIA s'il est alimenté avec des données malveillantes.	laquelle/lesquelles un SIA a rendu une décision x. **5.** Entrave aux DF en cas d'impossibilité de remplacer l'interaction IA – utilisateur par une interaction humain – utilisateur. **Enjeu** : Dans certains cas, + d'explicabilité = – de précision et vice versa.	**3.** Accès impossible, inéquitable ou restreint à un SIA *(ex : en raison de l'âge, du sexe, d'un handicap, des caractéristiques ou des capacités d'une personne.)* **4.** Entrave à la participation active en société de certaines personnes.	**3. Incidence négative sur la société et la démocratie** – *ex : atteinte au processus démocratique, amplification de fausses nouvelles, menace aux délibérations humaines, etc.)*	aux conséquences d'un tel résultat.

		2. Reproductibilité possible seulement dans des contextes ou conditions spécifiques. **Enjeu** : Objectivité VS bons résultats – *L'apprentissage de méthodes nouvelles ou inhabituelles par le SIA pour obtenir de bons résultats peut porter atteinte à sa fonction objective.*					
Meilleures pratiques	**DROITS FONDAMENTAUX** **Analyse d'impact** Entreprendre une analyse d'impact relative aux DF préalablement au développement du SIA. **Mécanismes** Mettre en place des processus pour tester et surveiller, pendant les phases de dév., de déploiement et d'utilisation du SIA : - La discrimination potentielle (biais); - Les dommages potentiels du SIA pour les enfants; - Les atteintes potentielles à la liberté d'expression, d'information et d'association. Mettre en place des processus pour traiter et corriger les atteintes identifiées. **Commentaires** Mettre en place des mécanismes	**SÉCURITÉ** **Certification** Certifier le SIA pour la cybersécurité ou le rendre conforme à des normes de sécurité spécifiques. **Prévention des risques** - Évaluer les formes potentielles d'attaques auxquelles le SIA pourrait être vulnérable; - Considérer différents types de vulnérabilités (*ex : data poisoning, model evasion, model inversion)* et identifier les conséquences possibles; - *Red-teaming, pentesting*; - Pour chaque cas d'utilisation spécifique, définir les risques, les mesures de risque et les niveaux de risque; - Mettre en place un processus pour mesurer et évaluer en permanence les risques; - Informer les utilisateurs des risques potentiels.	**CONFIDENTIALITÉ** **Signalement** Mettre en place des mécanismes permettant de signaler les problèmes liés à la confidentialité du SIA. **Protection des données** - DPIA (*Data Protection Impact Assessment*) - Désigner un délégué à la protection des données qui participe à la conception et au dév. du SIA. - Mettre en place des mécanismes de surveillance pour le TX des données, incluant : • La limitation de l'accès aux données au personnel qualifié qui justifie du besoin d'accéder à des données à caractère personnel; • Des mécanismes pour enregistrer l'accès aux	**EXPLICABILITÉ** **Explication** Expliquer la décision prise par le SIA aux utilisateurs. **Exigence** Dès qu'un SIA a une incidence importante sur la vie de personnes, il doit être possible d'exiger une explication appropriée du processus de décision du SIA au moment opportun et selon le destinataire. **Consultation** Sonder en permanence les utilisateurs pour vérifier s'ils comprennent la/les décision(s) du SIA. **TRAÇABILITÉ** Mettre en place des mesures pour assurer la traçabilité du SIA pour tout son cycle de vie (*EX : pour retracer les données + règles utilisées par le SIA pour arriver à une décision x)*. Documenter, selon les normes les plus strictes,	**BIAIS** **Phase de collecte des données** Supprimer les biais détectables et discriminatoires. Tenir compte de la diversité et de la représentativité des utilisateurs dans les données, via par ex : • L'évaluation et la mise en place de processus pour tester et surveiller les biais potentiels durant tout le cycle de vie du SIA. **Stratégie** Établir une stratégie pour éviter de créer/renforcer un biais injuste dans le SIA, tant en ce qui concerne l'utilisation des données d'entrée que pour le *design* de l'algorithme. **Éducation** Mettre en place des initiatives d'éducation et de sensibilisation pour aider les concepteurs et les développeurs d'IA – *sensibilisation au fait qu'ils peuvent involontairement injecter des biais ds le SIA.*	**IMPACT SOCIAL** **Milieu de travail** Avant d'introduire l'IA dans un milieu de travail : - Informer et consulter les travailleurs concernés et leurs représentants; - Adopter des mesures pour faire en sorte que les impacts du SIA sur le travail humain soient bien compris; - Vérifier que les travailleurs comprennent le fonctionnement du SIA; - Fournir des possibilités de formation si le SIA requière de nouvelles compétences; - Mettre en place des mesures pour contrer les risques de déqualification. **Incidences sociales négatives en général** Contrôler et examiner minutieusement les effets des IA sociaux. **ENVIRONNEMENT** Mettre en place des mécanismes pour évaluer	**AUDITABILITÉ** **Audit indépendant** Les applications affectant les DF devraient faire l'objet d'un audit indépendant. **Auditabilité facilitée** Mettre en place des mécanismes qui facilitent l'auditabilité du SIA (*ex : traçabilité du processus de développement du SIA)* **Auditeurs externes** S'assurer que le SIA peut être vérifié par des auditeurs externes lorsque requis. **GESTION DES RISQUES** Prévoir des directives externes ou des processus d'audit par des tiers pour surveiller les préoccupations éthiques et les mesures de responsabilité; Organiser une formation sur les risques qui informe également sur le cadre juridique potentiel applicable au SIA;

	permettant de recevoir des commentaires externes concernant les SIA susceptibles d'avoir une incidence négative sur les DF. **ACTION HUMAINE ET AUTONOMIE** **Prise de décision et comportement** Mettre en place une procédure pour éviter que le SIA affecte par inadvertance l'autonomie humaine et prendre des mesures pour atténuer le risque de manipulation des utilisateurs. **Information** Informer suffisamment les destinataires qu'une décision, un résultat, un contenu (etc.) provient d'une décision algorithmique et/ou qu'ils interagissent avec un SIA. **Dépendance** Mettre en place des procédures pour minimiser le risque de dépendance au SIA. Prévoir d'avance des mesures pour gérer les conséquences négatives potentielles qui pourraient résulter d'un attachement disproportionné au SIA. **Formation** Donner aux utilisateurs les connaissances et outils nécessaires pour comprendre le SIA et interagir avec lui dans	**Mises à jour** - Fournir des mises à jour; - Informer les utilisateurs de la durée de la couverture de sécurité et des mises à jour. **Pannes** Planifier la résistance aux pannes. **Réexamen** Développer un mécanisme pour évaluer quand le SIA doit faire l'objet d'un nouvel examen de sa robustesse technique et de sa sécurité. **PRÉCISION** **Données** Mettre en place des mesures pour garantir que les données utilisées pour développer le SIA sont à jour, de haute qualité, complètes et représentatives de l'environnement dans lequel le SIA sera déployé. **Surveillance** Surveiller et documenter la précision du SIA à travers des étapes prédéterminées. **Communication** Mettre en place des processus pour vérifier que le niveau de précision du SIA attendu des utilisateurs leur est correctement communiqué.	données et l'ajout de modifications; • L'utilisation de mesures de *privacy-by-design (ex : cryptage, pseudonymisation, etc.)* **DONNÉES** **Qualité des données** Tenir compte du fait que les données collectées peuvent être biaisées et/ou contenir des imprécisions, des erreurs ou des inexactitudes avant d'utiliser un ensemble de données pour entraîner un SIA. **Intégrité des données** Tester et documenter à chaque étape les processus et les ensembles de données utilisés, autant pour les SIA développés à l'interne que ceux acquis de l'extérieur. **Utilisation des données** Garantir le respect de la vie privée et la protection des données tout au long du cycle de vie d'un SIA – *Données concernées : infos initialement fournies par l'utilisateur et infos générées au sujet de l'utilisateur au cours de ses interactions avec le SIA.*	les ensembles de données et les processus permettant à un SIA de rendre une décision. Faire la même chose pour les décisions rendues par l'IA. **COMMUNICATION** Informer les utilisateurs qu'ils interagissent avec un SIA et leur permettre de s'opposer à l'interaction au profit d'une interaction humaine. Mettre en place des mécanismes pour informer les utilisateurs de l'objectif, des critères et des limites des décisions d'un SIA, de ses risques, et de ses limitations techniques. Fournir du matériel de formation approprié et des *disclaimers* aux utilisateurs sur la manière d'utiliser correctement le SIA.	**Conception** Mettre en place des procédures de contrôle pour analyser de manière claire et transparente la finalité, les contraintes, les exigences et les décisions du SIA. **Recrutement** Recruter des personnes issues de contextes et cultures différents. **Signalement** Mettre en place un mécanisme permettant de signaler les problèmes liés aux préjugés, à la discrimination ou aux mauvaises performances du SIA. **ACCESSIBILITÉ** ***Design* universel** Envisager des principes de *design* universel pour répondre aux besoins du plus large éventail d'utilisateurs selon les normes existantes d'accessibilité pertinentes. Veiller à ce que les principes de *design* universel soient utilisés à chaque étape du dév. **Consultation** Consulter les utilisateurs cibles dans le dév. du SIA. **PARTICIPATION DES PARTIES PRENANTES** **Consultation** Consulter les personnes susceptibles d'être directement ou indirectement affectées par le SIA (utilisateurs et autres).	l'impact environnemental du dév., du déploiement et/ou de l'utilisation du SIA *(ex : qté d'énergie utilisée et émissions de carbone).* Définir des mesures pour réduire l'impact environnemental du SIA tout au long de son cycle de vie. **SOCIÉTÉ ET DÉMOCRATIE** Évaluer l'impact sociétal de l'utilisation du SIA au-delà de l'utilisateur *(ex : individus indirectement affectés ou société au sens large)*; Mettre en place des mesures pour garantir que le SIA n'a pas d'impact négatif sur la démocratie. *L'utilisation des SIA devrait faire l'objet d'une attention particulière dans les situations mettant en jeu le processus démocratique.	Envisager de créer un comité d'examen éthique de l'IA (ou un mécanisme similaire) pour discuter de l'ensemble des pratiques de responsabilité (*accountability*) et d'éthique; Établir un processus pour discuter et surveiller en permanence l'adhésion du SIA à la ALTAI; Établir un processus permettant à certains tiers (ex : *fournisseur, utilisateur, vendeur, etc.*) de signaler les vulnérabilités, risques ou biais potentiels du SIA. Offrir une protection adéquate aux entités qui signalent des préoccupations légitimes concernant un SIA (*ex : ONG, syndicats, etc.*) **RÉPARATION** Mettre en place des mécanismes de réparation de type *redress by design* pour les applications qui peuvent nuire aux individus

	une mesure satisfaisante. **Évaluation** Permettre aux utilisateurs de procéder à une évaluation du SIA ou de le contester d'une manière appropriée. **CONTRÔLE HUMAIN** **Mécanismes** Recourir à des mécanismes de gouvernance tels que les approches : → ***Human-on-the-loop*** *– capacité d'intervention humaine dans la conception et la surveillance du fct du SIA;* → ***Human-in-command*** *– capacité de contrôle de l'activité globale du SIA et faculté de décider quand/comment utiliser le SIA dans une situation X – ex : décision d'ignorer une décision du SIA* ***NOTE** : L'approche **human-in-the-loop** (où l'humain intervient dans chaque cycle de décision du système) est souvent ni possible ni souhaitable.* **Formation surveillance** Donner une formation spécifique sur la façon d'exercer la	**FIABILITÉ ET REPRODUCTIBILITÉ** **Vérification** - Vérifier l'atteinte d'objectifs par le SIA via un processus bien défini; - Mettre en place des méthodes de vérification et de validation pour évaluer et garantir différents aspects de la fiabilité et de la reproductibilité du SIA et clairement documenter et opérationnaliser ces processus. **Plan de secours** Mettre en place des garanties permettant le déclenchement d'un plan de secours en cas de problème (*ex : changement de procédure ou demande d'instructions à un humain avant de poursuivre son action*). Tenir compte des scénarios où une intervention humaine n'est pas immédiatement possible.	Mettre en place des garanties assurant aux citoyens que leurs données ne seront pas utilisées à leur encontre à des fins discriminatoires, de manière illicite ou injuste. **Consentement** Permettre le retrait du consentement, le droit d'opposition et le droit à l'oubli. **Normes** « Aligner » le SIA sur les normes pertinentes (*ex : ISO*) ou sur des protocoles largement adoptés propres à la gestion quotidienne des données.		Envisager un mécanisme pour inclure la participation du plus large éventail possible de parties prenantes. **Mécanismes à long terme** Mettre en place des mécanismes à plus long terme pour la participation des parties prenantes. Solliciter régulièrement des commentaires, même après le déploiement.		

	surveillance du SIA aux personnes concernées. ***Detection & response*** Mettre en place des mécanismes *"detection and response*" pour les effets indésirables du SIA. **Mesures spécifiques** Mettre en place des mesures de surveillance et de contrôle propres à la nature d'autoapprentissage du SIA. **Interruption** Mettre en place un "*stop buton"* ou une procédure pour interrompre en toute sécurité une opération lorsque requis. **Autorités publiques** Veiller à ce que les autorités publiques puissent exercer un contrôle via des mécanismes de contrôle qui varient selon le risque et le domaine d'application du SIA. **Proportionnalité** Approfondir les essais et renforcer la gouvernance proportionnellement au degré de contrôle humain en vigueur.						
Entraves	-	**Précision** Situations où il n'est pas possible d'éviter des	-	**Black boxes** Situations où il n'est pas possible d'expliquer pourquoi le SIA a généré	**Biais** Phase de collecte des données : certains biais	-	-

		prévisions inexactes occasionnelles.		un résultat ou une décision, **ou encore** quels facteurs d'entrée (*input factors*) y ont contribué.	discriminatoires ne sont pas détectables. **Absence de définition commune d'équité**		
Stratégies	-	**Précision** Indiquer le niveau de probabilité de ces erreurs.	-	**Black boxes** Mettre en place d'autres mesures d'explicabilité *(ex : traçabilité, auditabilité et communication transparente sur les capacités du SIA)*.	**Définition de l'équité** Consulter les groupes pertinents pour la définition correcte de l'équité *(ex : personnes âgées, personnes handicapées)*. Adopter une définition commune et la mettre en œuvre à toutes les phases du processus de mise en œuvre du SIA.	-	-

	1.2 Recommandation du Conseil sur l'intelligence artificielle, OCDE (2019)				
	5 principes d'une IA digne de confiance				
	Croissance inclusive, développement durable et bien-être	**Valeurs centrées sur l'humain et équité**	**Transparence et explicabilité**	**Robustesse, sûreté et sécurité**	**Responsabilité**
Enjeux, risques ou dilemmes	**Enjeux** Réalisation de résultats bénéfiques des SIA pour les individus et la planète (ex : renforcement des capacités et de la créativité humaines); Inclusion des populations sous-représentées; Réduction des inégalités (économiques, sociales, de genre); Préservation des milieux naturels.	**Enjeu** Respect par les acteurs de l'IA de l'état de droit, des droits de l'homme et des valeurs démocratiques tout au long du cycle de vie des SIA – *liberté, dignité, autonomie, protection de la vie privée et des données, non-discrimination, égalité, diversité, équité, justice sociale, droits des travailleurs*.	**Enjeux** Compréhension générale des SIA; Sensibilisation des parties prenantes à leurs interactions avec des SIA; Compréhension du résultat d'un SIA par les personnes concernées; Possibilité pour les personnes subissant les effets néfastes d'un SIA d'en contester les résultats.	**Risques** Fct non convenable des SIA posant un risque de sécurité démesuré dans des conditions d'utilisation normales ou défavorables, ou en cas d'utilisation abusive; Atteinte à la vie privée; Compromission de la sécurité numérique et de la sûreté; Biais.	**Enjeu** Responsabilité pour le bon fct des SIA et pour le respect des principes.
Meilleures pratiques[23]	**Compte tenu de l'état de l'art et du contexte :**				
	-	**Garanties, mécanismes** Instituer des garanties et des mécanismes (ex : attribution de la capacité de décision finale à l'humain).	**Contestation** Fournir des infos claires et facilement compréhensibles sur les facteurs et la logique ayant servi aux prévisions, recommandations ou décisions.	**Traçabilité** Veiller à la traçabilité des : - Ensembles de données; - Processus; - Décisions prises au cours du cycle de vie des SIA pour - Analyser les résultats produits par ces SIA;	**Responsabilité** Tenir responsables les acteurs de l'IA en prenant en considération leurs rôles.

[23] Applicables aux « acteurs de l'IA », c.à.d. : « les parties jouant un rôle actif dans le cycle de vie d'un système d'IA, y compris les organisations et les individus qui déploient ou exploitent l'IA. »

				- TX les demandes d'information. **Gestion des risques** Appliquer une approche systématique de la gestion du risque de façon continue et à chaque phase du cycle de vie des SIA.	
	5 recommandations pour les décideurs (élaboration des politiques nationales et coopération internationale) pour tendre vers une IA digne de confiance : • Investir dans la R&D en matière d'IA ; • Favoriser l'instauration d'un écosystème numérique pour l'IA Ex : technologies et infrastructures numériques, mécanismes de partage des connaissances en matière d'IA, fiducies de données. • Façonner un cadre d'action favorable à l'IA Ex : envisager le recours à l'expérimentation pour tester les SIA dans un environnement contrôlé. • Renforcer les capacités humaines et préparer la transformation du marché du travail Ex : doter les personnes des compétences nécessaires pour utiliser et interagir efficacement avec les SIA, mettre en place des programmes de formation pour assurer une transition équitable des travailleurs, soutenir les personnes affectées par les suppressions de postes. • Favoriser la coopération internationale au service d'une IA digne de confiance.				
Entraves	-	-	-	-	-
Stratégies	-	-	-	-	-

2. ALLEMAGNE

	2.1 From Principles to Practice : An interdisciplinary Framework to operationalise AI Ethics, AIEI Group, (2020)					
	6 valeurs éthiques					
	Justice ***(Comprend : fairness & non-discrimination)***	**Environmental sustainability**	**Accountability**	**Transparency** ***Comprend : explicabilité et interprétabilité***	**Privacy**	**Reliability**
Enjeux, risques ou dilemmes	**Risques** **1. Discrimination envers certains groupes résultant de** : - *Training data* biaisées; - Biais involontaires des ingénieurs; - Implantation de biais résultant de contextes d'utilisation particuliers; - Manque de « diligence raisonnable » et de réflexion dans le processus de développement; -Prise en considération, dans la prise de décision algorithmique, de facteurs qui ne devraient pas jouer de rôle (*ex : âge, origine, handicap, sexe, etc.*) -Biais de sélection des données (manque d'exhaustivité). **2. Atteinte à la justice sociale résultant de :** -Mauvaises conditions de travail dans le processus de dév. d'un SIA– *le dév. d'un SIA requiert l'annotation manuelle d'ensembles de données => nécessite de recourir à des entreprises d'étiquetage spécialisées, particulièrement en Asie, qui n'offrent pas de salaire minimum et où les*	**Enjeux** **1. Préservation des conditions de vie des générations futures** **2. Précision VS empreinte carbone** *-La puissance et la précision d'un SIA dépendent du nombre de ressources de calculs dont le modèle d'IA dispose, mais + de ressources de calculs = + de* training data *traitées = consommation d'énergie augmentée = empreinte carbone augmentée.* **Risque** : **Élimination inadéquate du matériel informatique obsolète.**	**Enjeu** : **difficulté d'attribution de responsabilité / de définition des agents responsables** **Risques associés** : **1.** En cas de problème, impossibilité d'identifier précisément un coupable en particulier; **2.** *Responsability diffusion*; **3.** *Reponsability flight*.	**Enjeu : + de transparence peut équivaloir à – de confidentialité.** **Risques** **L'absence de transparence compromet** : - « L'autorégulation » : possibilité, pour les personnes affectées un SIA, d'ajuster le comportement décisionnel de l'IA envers elles de manière éclairée; - L'identification et la correction de violations de droits; -Le débat social; - La construction de relations de confiance.	**Risques** **1. Atteinte à la confidentialité des informations** via l'utilisation de l'IA dans des domaines sensibles à la confidentialité (*ex : surveillance de masse, soins de santé, marketing, etc.*) **2. Risque inhérent à l'IA :** les applications d'IA posent elles-mêmes un problème de confidentialité en raison de leur dépendance à une quantité massive de données.	**Risques** **1. Problèmes liés à l'absence de fiabilité** L'absence de fiabilité contribue à l'absence de prévisibilité, à la perte de confiance dans le SIA et à la survenance de préjudices individuels ou sociétaux. **2. Atteinte à la confidentialité, à l'intégrité et à la disponibilité de l'information** Des pannes, des accidents ou des attaques (internes ou externes) peuvent compromettre la cybersécurité, et donc compromettre la confidentialité, l'intégrité et la disponibilité de l'information.

	droits des travailleurs ne sont pas respectés.					
Meilleures pratiques	**DÉTECTION DE BIAIS POTENTIELS** - Analyser les ***training data*** de manière à assurer la parité démographique et l'égalité des chances; - Examiner, sur une base régulière, le ***input design*** (capteurs + interface utilisateur) et les ***input data****;* - **Rechercher des effets discriminatoires** implicites ou explicites dans les exigences, les objectifs et les définitions des tâches (*task definitions*). Effectuer cette recherche de manière continuelle *(c.à.d. : pas seulement après des modifications de l'app ou de son environnement, et pas seulement de manière périodique.)* - Prendre en compte les potentiels processus **d'auto-renforcement** de façon périodique; - Porter une attention particulière aux effets discriminatoires causés par la **conception** du *data output* et revoir ce *design* de façon périodique *(c.à.d. pas uniquement lors du processus de développement et de MEO de l'app)*; - Éviter la collecte de ***proxies*** et mettre en place une procédure spéciale de vérification d'éventuels *proxies* de données sensibles. Faire cette vérification de façon continuelle *(c.à.d. pas uniquement lors de modifications de l'app et pas uniquement de façon périodique)*. **JUSTICE SOCIALE** - Évaluer les **conditions de travail** via des mécanismes d'évaluation internes et externes. **DÉTECTION ET PRÉVENTION DES BIAIS**	**ENVIRONNEMENT** -Mettre en place des **infrastructures économes** en ressources pour les TI – *principalement via la construction de centres de données écoénergétiques, et par le dév. de modèles d'apprentissage automatique moins énergivores.* -Avoir recours à des processus de **certification** de l'empreinte carbone. -Faire preuve de **transparence** en ce qui concerne la consommation d'énergie d'un SIA; -En ce qui concerne l'élimination du matériel **obsolète** : "In this context, a **right to repair** can improve the situation."	**ATTRIBUTION DE LA RESPONSABILITÉ ORGANISATIONNELLE** - Dans un **contrat**, clarifier les responsabilités des acteurs. - **Revoir** et mettre à jour l'attribution des responsabilités de façon régulière et permanente (*c.à.d. pas seulement après des changements significatifs dans l'app ou de son environnement)*. - Dans des cas de **responsabilité partagée**, définir les rôles et obligations des responsables concernés via une documentation détaillée et accessible de manière à ce que les personnes concernées puissent connaitre leurs obligations et celles des autres. **MESURES TECHNIQUES** -Pour assurer la traçabilité interne, mettre en place des méthodes permettant d'**expliquer** les résultats de façon causale et d'observer les **influences** environnementales sur les SIA (*techniques to causally explain outputs and to observe environmental influences on AI systems*). - **Surveiller**, dans leur interaction avec leur environnement, les SIA qui ont une composante d'apprentissage. Effectuer cette surveillance via des techniques permettant d'expliquer les résultats de manière causale et d'observer les influences environnementales le SIA. **RESPONSABILITÉ INSTITUTIONNELLE** - Mettre en place des moyens d'**indemnisation** suffisants (*ex : police d'assurance*) pour les sinistres probables et moins probables;	**DIVULGATION DE L'ORIGINE DES ENSEMBLES DE DONNÉES** -Documenter **l'origine** des données – via un enregistrement complet de toutes les *training data* et des données d'exploitation et via un contrôle de versions des ensembles de données. *Cette documentation doit être faite par l'organisation (et non par des tiers). -Identifier les données qui sont utilisées par un SIA dans **chaque cas** donné *(et non pas de façon générale*). -Documenter et divulguer les caractéristiques des ***training data sets*** *– via des fiches techniques complètes.* **DIVULGATION DES PROPRIÉTÉS DE L'ALGORITHME / MODÈLE** - **Tester** suffisamment le modèle en théorie et en pratique • Le modèle doit être facilement inspectacle et testable afin de découvrir ses faiblesses potentielles. -Parmi les modèles envisagés, choisir le modèle le plus **intelligible**. **ACCESSIBILITÉ** -Envisager des **modes d'interprétabilité spécifiques** aux groupes cibles et les développer avec eux *(participation des groupes cibles)*. -Donner accès aux informations sur l'algorithme et le modèle utilisés. -S'assurer que le principe de fct du SIA est facilement compréhensible et interprétable – *via un modèle directement compréhensible, c.à.d. qui ne nécessite pas de se référer à des « modes d'interprétabilité ».*	**COLLECTE ET UTILISATION** - Collecter et utiliser les données personnelles à des **fins spécifiques** seulement. - **Mettre fin** au TX des données une fois l'objectif atteint. - N'utiliser les données à des fins autres que celles pour lesquelles elles ont été collectées que si les personnes concernées ont donné un **consentement explicite** à cet effet. - Permettre aux personnes concernées de **refuser l'utilisation** de leurs données – *ex : droit à la suppression de données, droit de rectifier et capacité de restreindre le TX des données.* **DÉVELOPPEMENT SIA** - Favoriser l'apprentissage avec des données **anonymisées** ou **pseudonymisées**. - Favoriser l'utilisation de procédures d'anonymisation et de pseudonymisation **fiables**. -Intégrer des normes de ***privacy-by-design***.	**SÉCURITÉ GÉNÉRALE** -Mettre en place une **évaluation** technologique approfondie pour révéler les menaces potentielles les plus importantes. **CYBERSÉCURITÉ** -Avoir recours à des **moyens techniques** (*ex : normes de cryptage de données, pare-feux, reconnaissance de logiciels malveillants, etc*.) et **non-techniques** (*ex : pratiques sociales, formation*).

	- Avoir recours à une **investigation externe** sur les sources d'erreurs par une institution indépendante; - Effectuer des **simulations** avant la MEO de l'app pour identifier les biais possibles *(*les simulations doivent être conçues pour des cas d'utilisation spécifiques / pas uniquement des « simulations générales de robustesse »)*. - Fournir une **documentation transparente** de l'ensemble des processus d'application de manière à rendre publics les mécanismes d'examen et les sources d'erreurs de façon suffisamment détaillée. - **Communiquer** publiquement les biais potentiels. **PROCÉDURES PARTICIPATIVES** - Favoriser **l'accessibilité** – *absence de restriction faite en fonction de catégories ou de critères protégés comme le sexe, l'origine, etc.)*. -Permettre à quiconque de *initiate an assessment of bias and processing* d'une plainte sans égard à la preuve d'une atteinte personnelle; - Inclure les *affected demographics* via la mise en œuvre, sur une base régulière, d'un processus de participation; - Définir et documenter de manière fiable les parties prenantes et les *affected demographics.* Rendre la documentation disponible aux tiers de confiance.		- Mettre en place un organe facilement accessible et publiquement annoncé (**ombudsman**) – *ex : pour permettre les questions, les plaintes*, *etc*. **DIVULGATION DES RESPONSABILITÉS ORGANISATIONNELLES** - Institutionnaliser une méthode permettant la **divulgation anonyme** d'informations aux parties concernées via un organe facilement accessible et ouvertement annoncé; -Définir les responsabilités à l'égard des tiers; - Mettre en place une journalisation complète (***comprehensive logging***) du processus de *design – portant sur toutes les* training data et les données d'exploitation entrantes, et sur le contrôle de versions des enregistrements de données. **Ce processus doit être réalisé par l'organisation, et non par des tiers.* **TOLÉRANCE À L'ERREUR** - Développer une culture où il est possible de discuter ouvertement des erreurs, mais sans que cela mène à une tolérance à l'erreur. **MÉCANISMES DE RESPONSABILITÉ ET DE RÉPARATION** -Concevoir les mécanismes de responsabilité et de réparation **indépendamment des principes juridiques** de causalité. -Rendre possible **l'appel** de décisions algorithmiques.	-Rendre accessibles les « **hyperparamètres** » (paramètres des méthodes d'apprentissage). -Mettre en place une autorité de **médiation** compétente pour régler les conflits propres à la transparence et attribuer suffisamment de pouvoirs à cette autorité.		
Entraves	-	-	-	**État de la technique** Impossible d'interpréter exactement pourquoi les réseaux de neurones profonds produisent certains résultats.	-	-
Stratégies	-	-	-	-	-	-

	2.2 Trustworthy use of artificial intelligence: priorities from a philosophical, ethical, legal, and technological viewpoint as a basis for certification of artificial intelligence, Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS (2020)					
	"6 AI-specific audit areas for trustworthy use of AI"					
	Autonomy and control	**Fairness**	**Transparency**	**Reliability**	**Security**	**Data protection**
Enjeux, risques ou dilemmes	**Risques** **Atteinte à l'autonomie des individus et des groupes sociaux :** - Confiance excessive des utilisateurs envers l'app; - Dév. de liens émotionnels utilisateurs – app; - Altération ou manipulation de la prise de décision des utilisateurs et influence sur les personnes dans le choix de leurs objectifs et des moyens pour les atteindre (alors que les apps d'IA ne peuvent que choisir les *moyens* pour atteindre des objectifs humains). **Utilisation incontrôlée** - Impossibilité pour les utilisateurs de contrôler l'app dans une mesure suffisante; - Changement constant du modèle ML (apprentissage continu) difficilement contrôlable et pouvant avoir des conséquences indésirables. *Ex : chatbot Tay*	**Risques** **TX discriminatoire injustifié et résultats injustes sur la base de l'appartenance à un groupe (sexe, couleur de peau, religion) résultant notamment de :** - Données historiques discriminatoires; - Sous-représentation de groupes dans les bases de données. **Absence de prise en compte de différences lorsque requis** (ex : système de reconnaissance vocale qui ne tient pas compte des accents.) **Enjeu** Dépendance des apps aux données (historiques) d'entraînement.	**Risques** Confusion sur le fait qu'une communication se produit avec une app; Utilisateurs non informés du contexte d'utilisation de l'app; Consentement ou refus non éclairé; Difficulté à comprendre les fonctions, les processus d'apprentissage et les décisions prises par certaines apps en raison de la nature complexe du modèle ML (*black box*) ou d'un manque de connaissances. **Enjeux** Les exigences d'interprétabilité, de traçabilité et de reproductibilité nécessitent une compréhension des processus internes de l'app; L'amélioration de l'app et la clarification des conflits requièrent la compréhension du comportement général de l'app par des experts; L'acceptabilité de l'app dépend de sa transparence. **Dilemme**	**Risques** Survenance de nouveaux types d'erreurs pouvant conduire à des situations critiques (surtout en ce qui concerne l'interaction homme – machine) pcq app d'IA non entraînée pour y répondre; Méthodes classiques pour tester les logiciels → échouent pcq impossible de décomposer les modèles ML en unités testables séparément (tests modulaires); Souvent impossible de trouver une formule pour caractériser les *inputs* fiables. **Enjeux** Exactitude des résultats; Estimation des incertitudes du modèle; Robustesse aux *inputs* malveillants (ex : attaques) et aux erreurs.	**Risques** Défaillance fonctionnelle, changement majeur dans le composant IA ou fuites d'informations non autorisées pouvant résulter de : - Attaques de l'app; - Dangers, accidents et erreurs dans l'utilisation. **Enjeux** Confiance dans les apps; Existence d'objectifs de sécurité spécifiques aux apps mais abstraits et non opérationnalisés (HLEG) **VS** existence de normes vérifiables et opérationnelles dans le domaine de la sécurité, mais non spécifiques aux apps d'IA.	**Risques** **Ingérence dans la sphère privée individuelle et commerciale (ex : secrets commerciaux);** - Failles de protection permettant un accès non autorisé aux données; - Accès aux données personnelles en l'absence de consentement / consentement vicié; - Identification possible de personnes même en l'absence de spécification des attributs correspondants ou de données personnelles. **Atteinte au droit à l'autodétermination informationnelle** **Enjeux** **« Liaison » de données** Défis pour la protection des données plus importants que dans les systèmes

	de Microsoft + apprentissage de phrases racistes.		Les *black boxes* = + de précision et de robustesse, mais - d'explicabilité. Les modèles plus simples = + d'explicabilité, mais – de précision et de robustesse.	<u>Autorisation d'apps</u> Détermination des apps qui devraient être autorisées. <u>Évaluation de la fiabilité</u> Détermination des *conditions* dans lesquelles l'app peut être classée comme fiable.		informatiques traditionnels, pcq les apps peuvent lier des données qui ne l'étaient pas auparavant + nouvelles méthodes de liaison des données rendues possibles avec l'apprentissage-machine. **Identification** Plus il y a de données liées, plus le risque que des personnes puissent être identifiées est grand.
Meilleures pratiques	**Évaluation** - Établir dans quelle mesure les utilisateurs peuvent développer une confiance excessive dans l'app, tisser des liens émotionnels ou voir leur prise de décision altérée ou manipulée de façon inacceptable par l'app. **Information** <u>Informer les utilisateurs</u> : - Des risques liés à une atteinte à leur autonomie; - De leurs droits et obligations, de leurs options d'intervention et des possibilités de plaintes/réclamations; - Du comportement de l'app d'IA. **Adapter ces informations aux personnes ayant des besoins particuliers;*	***Training data*** Utiliser des données d'entraînement représentatives des différents groupes visés. ***Outputs*** Améliorer le *output* du modèle ML ("an improvement in the output of the ML model comes into consideration as a suitable instrument for avoiding bias"). **Mesure de *fairness*** <u>Développer une mesure technique quantifiable d'équité</u> : Identifier les groupes qui ne devraient pas être désavantagés (personnes physiques ou morales);	**Information** Rendre disponibles les informations sur l'utilisation correcte de l'app; Informer les gens que la communication se produit avec une app; Informer les utilisateurs du but, groupe cible, risques potentiels de l'app – incluant ceux relatifs aux autres *audit areas* (*reliability*, *security*,*fairness*). **Décision** Communiquer les informations permettant aux utilisateurs de comprendre le résultat. Ne pas « submerger » les utilisateurs de détails non pertinents. **Documenter** Documenter le comportement général d'une app (au stade de son dév. et après).	**Évaluation de la fiabilité** Prendre en compte l'évaluation initiale des risques et les cadres éthique et juridique; Traduire les exigences en mesures quantitatives et en valeurs cibles; Évaluer la fiabilité pour une application d'IA <u>spécifique</u>; Décrire le champ d'application de l'IA le plus précisément possible pour s'assurer que les *training data* et les *test data* utilisées couvrent la qté attendue de *inputs* lors du fct de l'app. **Capacités humaines** Adapter la fiabilité de l'app aux capacités des personnes qui l'utilisent.	**Rassembler exigences** Rassembler les exigences des normes existantes essentielles pour la protection contre les attaques et les menaces contre les apps d'IA, et les <u>compléter</u> par des exigences d'IA spécifiques.	**Règlementation** Veiller à ce que les réglementations pertinentes en matière de protection des données (ex : GDPR) soient respectées. **Data protection officer** <u>*Dans un contexte de certification</u> Nommer un *data protection officer*; S'assurer que le SIA analyse et documente suffisamment les risques et les mesures propres à la protection des données pour « soutenir » le *data protection officer*. **Protection** Protéger efficacement les informations recueillies

	Fournir les informations pertinentes de manière à ne pas « submerger » les utilisateurs.* **Consentement - Offrir aux utilisateurs plusieurs options d'utilisation (pas seulement une option oui/non); - Permettre aux utilisateurs de révoquer leur consentement à l'utilisation d'une app d'IA; - Permettre aux utilisateurs de désactiver complètement l'application. **Interventions** - Prévoir des options d'intervention sécurisées en cas de détection d'une menace pour l'autonomie d'un utilisateur. **Surveillance humaine** - Surveiller étroitement l'app.	Choisir une définition de d'équité; Quantifier la définition d'équité choisie. **Pour l'équité de groupe :* *<u>Exigence</u> : les résultats pour tous les groupes visés doivent être comparables (« probabilité de succès » semblable);* **Pour l'équité individuelle :* *<u>Exigence</u> : les mêmes personnes doivent être traitées de la même façon.* **Personnalisation des SIA** Ex : les systèmes de commande vocale doivent être personnalisables en fct des dialectes / accents spécifiques. **Vérifications** Lors du fct d'une app, vérifier régulièrement si les distributions de données (*training* data ET données générées lors du fct) correspondent ou divergent.	<u>→ Méthodes</u> : "logging, documentation, or archiving of the design, data, training, testing/validating the model as well as the embedded environment." **Black boxes** <u>Renonciation</u> Renoncer aux modèles *black boxes* pour certaines apps; <u>Processus ultérieurs</u> Expliquer les modèles *black boxes* par l'utilisation de processus *ultérieurs* comme : → Formation d'un modèle explicatif pour compléter le modèle ML *black box* (le modèle explicatif calcule quelles parties des *inputs* ont été décisives pour un résultat en particulier); → Analyse *LIME* (Local Interpretable Model-agnostic Explanations) – c.à.d. analyse du comportement *input/output* du modèle.	**Tests** Effectuer des tests spécifiques à l'app en question et vérifier régulièrement le fct de l'app; Remplacer les tests modulaires du modèle (unités vérifiables séparément) par des tests quantitatifs qui utilisent des données de test distinctes ayant la même distribution statistique que les *training data*. **Mécanismes correctifs** Mettre en place des mécanismes correctifs appropriés + un plan de secours si des faiblesses sont détectées dans le modèle.		pendant l'entraînement et l'exploitation de l'app. **Utilisation** Utiliser les données personnelles à des fins spécifiques et informer les personnes de cette utilisation. **Consentement** Obtenir le consentement du propriétaire légal pour : - Accéder aux données personnelles; - TX ultérieurement les données ou les divulguer à des tiers. Offrir aux personnes la possibilité de supprimer leurs données; Notifier les personnes de la finalité et de l'utilisation des données personnelles ou des données qui en découlent. **Mécanismes** Fournir des mécanismes adéquats de consentement, d'enquête, d'opposition et de révocation liés à l'utilisation des données personnelles. **Analyse de risque** Réaliser une analyse de risque pour examiner la possibilité d'identification des personnes; À cet effet, vérifier toutes les mesures prises pour rendre les données anonymes ou pour agréger les données

						contre le potentiel de réidentification. **Vérification** Vérifier qu'il n'y a pas de failles de protection qui pourraient permettre un accès non autorisé aux données.
	Développement de l'IA - Impliquer tous les acteurs qui sont directement ou indirectement affectés; - Effectuer une analyse des risques analysant les possibilités d'abus (*misuse*) ou de double usage; - Identifier dès le début : • Le domaine d'application; • L'objectif et la portée de l'application; • Les personnes affectées. L'app doit *by design* être construite de façon à pouvoir être auditée et testée.					
Entraves	-	-	**Secret commercial, sécurité** Sauvegarde des secrets commerciaux ou de la « sécurité sociale générale ».	-	-	-
Stratégies	-	-	**Exigences spécifiques à cette entrave** 1. Les applications d'IA qui affectent les droits et les intérêts de tiers doivent être transparentes (traçabilité du fct de l'app); *Ces applications peuvent exceptionnellement demeurer opaques si cela est proportionné lors de la mise en balance des intérêts en jeu.		-	-

3. AUSTRALIE

Légende : Mauve : informations tirées de *Australia's Ethics Framework: A Discussion Paper* (**2019**) Vert : commentaires du public (experts, parties prenantes) p/r à la publication de **2019**	**Australia's AI Ethics Framework, Gouvernement de l'Australie, (2020)**[24] **8 principes éthiques**							
	1. Human, social and environmental wellbeing	**2. Human-centred values**	**3. Fairness**	**4. Privacy protection and security**	**5. Reliability and safety**	**6. Transparency and explainability**	**7. Contestability**	**8. Accountability**
Enjeux, risques ou dilemmes	**Risque** : **Utilisation de SIA** : **1.** Nuisible pour les individus, la société ou l'environnement. **2.** Bénéfique pour certains humains seulement. *EX : exclusion des générations futures.*	**Enjeu** Impact des SIA sur les compétences cognitives, sociales et culturelles humaines. **Risques** **Entrave à une société équitable et démocratique** : Atteinte aux droits fondamentaux; Atteinte à l'autonomie individuelle; Absence de diversité; Impacts des SIA sur l'environnement. **Risques spécifiques par rapport à :**	**Enjeu** : **Utilisation de systèmes prédictifs** : risques de discrimination (*voir plus bas*) **VS** plusieurs avantages potentiels en cas d'utilisation appropriée, tels que : augmentation de la précision, de la reproductibilité et de l'efficacité, diminution des biais humains et diminution de la prise en compte de variables qui devraient être étrangères à une prise de décision (*EX : fatigue).* **Risques** : **1.** Absence **d'inclusivité**. **2. Accessibilité** restreinte ou inéquitable aux produits et services associés aux SIA. **3.** Atteinte aux **droits humains**.	**Risques** : **1. Atteinte à la vie privée** : **1.1 Absence de consentement ou consentement inadéquat** lorsque des données personnelles sont impliquées : **a.** Individu non adéquatement informé; *EX : ignorance du but pour lequel les données générées à partir des activités en ligne sont collectées.* **b.** Consentement non volontaire; **c.** Consentement non spécifique ou actuel (*current*), notamment en raison d'absence de dispositions législatives	**Enjeux** Fct fiable du SIA conformément à l'objectif prévu, précision et reproductibilité; Robustesse et sûreté dépendent de l'identification correcte de la responsabilité.	**Risques** : **1.** Individus non au courant qu'ils interagissent avec un SIA. **2.** Individus conscients qu'ils interagissent avec un SIA, mais sans savoir ce que le système fait et pourquoi. **3.** SIA impactant significativement des individus à leur insu ou sans publicité suffisante. *EX : utilisateurs affectés par l'utilisation d'un algorithme de police prédictive qui se sert de données accumulées sur eux sans qu'ils en soient conscients.*	**Risques** : **1.** Absence de processus ou inefficacité du processus pour permettre la contestation de l'utilisation ou du résultat d'un SIA qui a un impact significatif sur des individus. **2.** Difficulté de "opt-out" d'une application et des résultats qu'elle entraine. **3.** Difficulté d'obtenir une assistance humaine. **Enjeu** : Confiance du public.	**Risques** : **1. Difficulté d'attribution de la responsabilité** en cas de problème survenant à la suite d'une décision automatisée car : → L'application de principes légaux relativement au principe d'*accountability* pour les SIA est encore en développement = peu ou pas de précédents juridiques pour une application d'IA donnée; →Absence de lignes directrices cohérentes et universelles applicables dans différentes industries qui utilisent des technologies capables de prendre des

[24] Les principes *Generates net-benefits*, *Do no harm* et *Regulatory and legal compliance* définis dans *Australia's Ethics Framework: A Discussion Paper* (2019) ont été **remplacés** en 2020 par *Human, social and environmental wellbeing*, *Human-centred values* et *Reliability and safety*.

		1. Diversité Ignorance de certaines considérations importantes (uniquement perceptibles par certaines parties prenantes) en raison de l'absence d'un large éventail de perspectives dans le *design*, le dév., le déploiement et l'exploitation de SIA. **2. Autonomie individuelle** **2.1** Difficulté pour les personnes interagissant avec un SIA de garder un « contrôle » total et efficace sur elles-mêmes en raison d'actions comme la tromperie (*deception*), la manipulation injuste, la surveillance injustifiée ou le fait de ne pas maintenir « l'alignement » entre un objectif divulgué et une action réelle. **2.2** Atteinte à la « perception de la réalité » et au processus de prise de décision des individus. *EX : affichage ciblé qui impacte le contrôle de l'humeur et des émotions, manipulation d'informations en ligne, publicité ciblée, etc.* **2.3** Dépendance excessive aux SIA entraînant des conséquences nuisibles *EX : les biais d'automation (automation bias) entraînent une augmentation des erreurs par commission ou par omission (ignorer ou*	**4. Discrimination**[25] directe ou indirecte envers des individus, des communautés ou des groupes et violation des lois sur la discrimination résultant de : **4.1** Ensembles de données biaisés (notamment en raison des biais des programmeurs); **4.2** Ensembles peu précis de données (dont les *input data* ne sont pas raisonnablement exactes); **4.3** Utilisation de variables non discriminatoires, mais fortement corrélées à d'autres variables discriminatoires (*EX : quartier d'une personne*); **4.4** Utilisation de bases de données dont les données sont basées sur des pratiques biaisées (*EX : police + pratique du* stop-and-frisk = *cible de manière disproportionnée les hommes afro-américains = alimente les bases de données du SIA donné de police prédictive = + d'afro-américains scannés par le SIA* = feedback loop = *plus susceptibles d'être ciblés*); **4.5** « Alimentation » des SIA avec des données inappropriées (*EX : Tay, chatbot de Twitter*); **4.6** Difficulté du SIA à reconnaître et à filtrer les	spécifiques au droit à l'oubli; **d.** Incapacité à comprendre et à communiquer le consentement; **e.** Consentement d'un utilisateur pour l'utilisation des données d'un autre utilisateur (*EX : Cambridge Analatyca*). **1.2 Utilisation de données personnelles à des fins d'identification et de surveillance** *EX : applications de reconnaissance faciale*. **1.3 Atteinte au droit à la protection des données / violations des données personnelles *(data breach)*** résultant de : -Failles de sécurité dans le stockage et l'utilisation des données (ce qui peut mener à des erreurs humaines et à des attaques); -Absence de mesures ou de processus appropriés pour adéquatement protéger les données personnelles.		**4.** Conséquences négatives (*EX : résultats injustes, atteinte aux droits humains ou impossibilité de connaitre la cause d'un accident*) en raison de l'incompréhension des mécanismes de prise de décision automatisée. **Enjeux** : **1.** Des SIA transparents peuvent tout de même fonctionner avec un taux d'erreur élevé et des biais significatifs. **2.** L'absence de mesures de transparence empêche la responsabilité *(accountability)*. **3.** Le niveau de transparence requis pour que l'algorithme demeure *accountable* et qu'il fonctionne selon les lois et les normes sociales peut être difficile à établir. **4.** L'absence de mesures de transparence peut impacter : - La validation et la certification de l'IA;		décisions affectant de manière significative la vie humaine; →Absence de chaîne de responsabilité claire et de délimitation claire entre la responsabilité des utilisateurs et des programmeurs. **1.2 Difficulté d'attribution de la responsabilité** en ce qui concerne les algorithmes *open source* et l'utilisation de SIA au-delà de leur but initial.

[25] Entre autres : perpétuation d'injustices sociétales, traitement disparate de l'utilisation de SIA sur des groupes vulnérables et sous-représentés (*notamment en fct de l'âge, de la race, du sexe, du statut intersexe, de l'identité de genre et de l'orientation sexuelle*), actions de SIA avantageuses envers certains groupes mais désavantageuses envers d'autres et prédictions biaisées (dans le cas de SIA utilisés pour prédire le comportement humain).

		suivre le conseil d'un SIA) en raison de l'absence de remise en question des conseils donnés.	données offensantes (*EX : Tay – chatbot de Twitter*); **4.7** Utilisation de *indiscriminate data* qui mène à des résultats injustes.	**2. Conséquences nuisibles d'une gouvernance inadéquate des données et des violations aux données personnelles** : -Atteinte à la réputation, conséquences psychologiques, financières (très coûteux), juridiques, professionnelles, etc. **Enjeux** : **1. Définition** : pas de signification claire de ce que la confidentialité veut dire dans un monde numérique. **2. Portée des atteintes à la vie privée** : dév. de l'IA => atteintes à la vie privée de grande portée (*EX : Cambridge Analytica*). **3. Mesures non à jour et nouveaux risques propres à l'IA** : les mesures de confidentialité ne « suivent » pas les nouvelles capacités de l'IA – *ex : existence de règles pour la collecte et l'utilisation d'empreintes digitales, mais pas de règles spécifiques à la collecte d'infos biométriques.* **4. Proportionnalité :** les problèmes de		-Les enquêtes ("*regulators in the context of investigations*"); -La preuve et la prise de décision en contexte judiciaire.		

				gouvernance des données augmentent à mesure que le dév. de l'IA s'accélère. **5. Équilibre entre** : **5.1** La protection de la vie privée individuelle et des processus de consentement transparents **ET** l'encouragement des investissements et de **l'innovation** dans des nouvelles technologies qui nécessitent des ensembles de données riches. **5.2** La protection de la sécurité du public **ET** l'adoption de mesures de surveillance intrusives.				
Meilleures pratiques	**1. Objectifs** Clairement identifier et justifier les objectifs des SIA. **2. Impacts** Considérer tous les impacts (positifs et négatifs, internes et externes à l'organisation) des SIA sur le bien-être individuel, social et	**1. Droits fondamentaux** Considérer soigneusement les risques potentiels des SIA sur les droits humains. **Possibilité d'interférer avec certains droits que lorsque cela est nécessaire, raisonnable et proportionné.*	**1. Inclusivité** **1.1 Groupes vulnérables** Accorder une attention particulière aux groupes vulnérables, défavorisés ou protégés lors de la programmation de SIA pour que des situations similaires entraînent des résultats similaires, et que des situations différentes puissent produire des résultats différents. (*EX : SIA confronté*	**1.Données utilisées** Examiner s'il existe des formes de *input data* moins invasives/sensibles qui pourraient donner des résultats similaires ou meilleurs. **2. Consentement** **2.1** Traiter adéquatement le consentement au moment de la collecte des données.	**1. Mesures de sécurité** Mettre en place des mesures de sécurité proportionnées à l'ampleur des risques potentiels. **2. Surveillance et tests** Surveiller et tester les SIA pour s'assurer	**1. Divulgation et justifiabilité** **1.1** Informer les gens qu'ils interagissent avec un SIA. **1.2** Informer les gens lorsque l'utilisation d'un algorithme est susceptible de les affecter.	**1. Mécanismes** Mettre en place des mécanismes efficaces et accessibles qui permettent aux gens de contester l'utilisation ou le résultat d'un SIA. *****Accorder une attention particulière aux	**1. Responsabilité** Mettre en place une chaîne de responsabilité claire pour les décisions prises par un système automatisé; Les personnes responsables des différentes phases du cycle de vie du SIA devraient être

	environnemental, et ce pour tout leur cycle de vie. **3. Préoccupations mondiales** Encourager le dév. de SIA qui aident à TX des préoccupations mondiales *(EX : objectifs de développement durable des Nations Unies.)*	**2. Diversité** Embaucher des personnes d'origines, de cultures et de disciplines variées. S'assurer que les *training data sets* sont inclusifs et représentatifs. Consulter diverses parties prenantes. Assurer une surveillance humaine des SIA pour garantir une gestion efficace des risques liés à la diversité et à l'inclusion. **3. Autonomie** **3.1 Dépendance et erreurs** a. Développement SIA ▪ Prendre en compte les biais d'automatisation lors du dév. et examiner attentivement la manière dont les SIA interagissent avec les opérateurs humains. ▪ S'assurer que les recommandations fournies par les SIA ne peuvent pas facilement être mal interprétées de manière nuisible. ▪ Examiner attentivement les moyens de garantir la mise en œuvre de principes de *design HITL*. **b.** Utilisation SIA ▪ Exercer une réflexion et une implication actives *(active thinking and involvement)* lorsqu'il	*à une personne âgée VS confronté à une personne jeune*). **1.2 Marché de l'emploi** Aider (le plus tôt possible, c.à.d. avant la perte d'emplois) les personnes susceptibles d'être affectées négativement par l'automation à faire une transition dans leur carrière via : formation, requalification, communication d'infos sur les risques/opportunités, mise en place d'incitations à la formation, etc. **2. Accessibilité** Consulter les parties prenantes qui peuvent être affectées par le SIA au cours de son cycle de vie. **3. Discrimination** **3.1 Conformité aux lois** Mettre en place des mesures pour garantir que les décisions prises par des SIA sont conformes aux lois anti-discrimination. **3.2 *Data inputs*** -Au stade du développement des SIA, porter une attention particulière aux ***training data*** pour garantir qu'elles soient exemptes de biais ou de caractéristiques susceptibles d'entraîner des résultats inéquitables par l'algorithme; -Veiller à ce que les ***training data*** incluent un échantillon robuste et inclusif; -Tenir compte des impacts de la discrimination (indirecte) résultant de l'utilisation de **variables fortement corrélées** à d'autres variables discriminatoires;	**Le consentement doit être :** **a.** Volontaire; **b.** Actuel / à jour; →Pour cette exigence, envisager l'incorporation du **droit à l'oubli** dans la législation. **c.** Spécifique; **d.** Transparent. **L'individu doit :** **a.** Être adéquatement informé; **b.** Avoir la capacité de comprendre et de communiquer son consentement. **2.2 *Consumer Data Right*** Envisager un nouveau droit, le *Consumer Data Right*, qui permet aux consommateurs de consentir à partager leurs données avec des destinataires de confiance (*EX : entreprises*). Ce consentement est explicite (pas de consentement implicite après le partage initial) et peut être révoqué. **3. Violations des données (*data breach*)** **3.1** Examiner régulièrement les politiques de collection et d'utilisation des données. **3.2** Signaler aux personnes concernées que leurs données personnelles ont été consultées ou divulguées d'une manière non autorisée qui est	qu'ils continuent de répondre à leur objectif. **3. Gestion des risques** Résoudre tout problème identifié avec une gestion continue des risques.	**1.3** Informer les gens qu'une décision qui les concerne a été prise par un SIA. **1.4** Communiquer les informations qui sont utilisées par l'algorithme dans sa prise de décision. **1.5** Fournir des justifications raisonnables des résultats des SIA dans un format convivial (*user friendly*) (*EX : facteurs-clés utilisés dans la prise de décision)*. *Ces divulgations (1.1 à 1.5) doivent se faire en temps opportun. **2. Mesures proportionnelles** Appliquer les exigences d'une manière proportionnelle à l'impact et aux risques potentiels d'un SIA donné.	personnes et groupes vulnérables. **2. Confiance du public** Communiquer clairement aux gens que la réparation de préjudice est possible – essentiel pour garantir la confiance du public dans l'IA. **3. Contestation efficace** Permettre un accès suffisant aux informations que l'algorithme utilise et aux inférences qu'il en tire. **4. Jugement humain** Pour les décisions qui affectent de manière significative les droits, mettre en place un système de contrôle efficace qui fait un usage approprié du jugement humain.	responsables des résultats. **2. Humains** Tenir compte du principe *HITL* lors de la phase de *design* des systèmes de décision automatisés; S'assurer que des ressources humaines suffisantes sont disponibles pour TX le nombre probable de demandes de renseignements p/r au SIA; Assurer des niveaux appropriés de surveillance humaine du SIA. **3. Identification** S'assurer que l'organisation et les personnes responsables des décisions identifiables. **5. Mécanismes** Mettre en place des mécanismes pour garantir la responsabilité et l'obligation de rendre compte (*responsability & accountability*) des SIA et de leurs résultats pour tout leur cycle de vie (avant/après leur conception, dév., déploiement, fct). **6. Examen externe**

		s'agit de prendre une décision basée sur les conseils de SIA. **3.2 Publicité ciblée** Informer que des techniques d'IA ont été utilisées. **3.3 Nouveaux outils** Développer de nouveaux outils qui permettent d'identifier les fausses informations. *EX : développement d'outils détectant les* vidéos *falsifiées*.	-Rigoureusement **évaluer** et tester toutes les variables utilisées pour développer et entraîner les algorithmes; -Mettre en place des systèmes de formation et d'éducation qui appuient les compétences requises pour lutter contre les biais dans les données d'échantillonnage (*sampling data*). **3.3 Tests et évaluations** -Réaliser des tests ou des simulations des SIA avant de les utiliser sur le public; -Évaluer et surveiller les **modèles** eux-mêmes pour s'assurer que des biais ne s'y infiltrent pas; -Avoir recours à des évaluations externes – les évaluations internes peuvent passer à côté de problèmes-clés si elles se fondent sur les mêmes hypothèses de *fairness* que le *design* original. (*EX : l'évaluation du logiciel COMPAS à l'interne n'avait pas conclu à la présence de résultats discriminatoires*). **3.4 Gouvernance des données** -S'assurer de disposer de « bases solides » en matière de gouvernance des données (*strong data governance foundations*) pour éviter que des SIA soient alimentés de données inappropriées. **4. Prédiction des comportements humains** **4.1 *Designers* d'algorithmes :**	susceptible de leur causer préjudice. (Australie : politique de signalement obligatoire – *Notifiable Data Breaches*). **3.3** Mettre en place des politiques de gouvernance des données plus strictes qui incluent : des correctifs techniques et une législation externe qui crée des répercussions plus importantes en cas de perte de données. **3.4** Mettre en place des mesures appropriées pour la sécurité des données et du SIA, comme l'identification des vulnérabilités de sécurité potentielles. **3.5** S'assurer que les mesures de sécurité tiennent compte des applications non prévues du SIA et des risques d'abus potentiels, et qu'elles prévoient des mesures d'atténuation appropriées. **4. Innovation IA** Gouvernement **4.1** Optimiser l'utilisation et la réutilisation de données publiques (*open data*). **4.2** Partager des données non sensibles *as open by default*. **5. Gouvernance des données** **5.1** Veiller à la gouvernance et à la gestion de toutes les données utilisées et				Pour les SIA ayant un impact significatif sur les droits d'un individu : ces SIA devraient être soumis à un examen externe.

			Porter une attention particulière à la manière dont les SIA parviennent à une prédiction. **4.2 Organes gouvernementaux :** Établir des normes industrielles appuyées par des lignes directrices à jour de manière à soutenir l'utilisation d'algorithmes équitables. **5. Utilisation à long terme d'un SIA** -Vérifier que les données collectées au fil du temps servent toujours le résultat escompté; -Mettre en place des mesures qui garantissent que cette question est régulièrement évaluée. **6. Cycle de vie SIA** Le principe de *Fairness* doit être considéré pour tout le cycle de vie des SIA, c.à.d. pas seulement en ce qui a trait aux *training data* et aux algorithmes.	générées par un SIA au cours de son cycle de vie, par exemple via : -L'anonymisation des données utilisées; -L'évaluation continue de la « solidité du lien » entre les données utilisées et les inférences que le SIA en tire. **5.2** Envisager de nouvelles capacités de l'IA et veiller à ce que le régime de gouvernance des données déjà mis en place reste pertinent eu égard à ces capacités. **5.3** S'assurer de disposer de « bases solides » en matière de gouvernance des données (*strong data governance foundations*) pour éviter la violation des lois sur la confidentialité.				

Meilleures pratiques communes aux 8 principes
(*AI ethical Toolkit* dans *Australia's Ethics Framework: A Discussion Paper*)

Impact Assessments
Pour encourager la responsabilité (*accountability*) et garantir que les principes éthiques sont pris en compte avant la MEO de l'IA.

Review processes
Veiller à ce que les SIA adhèrent aux principes éthiques et aux politiques/lois via :
- L'examen de SIA par des professionnels;
- L'examen de SIA par des SIA - Ex *: outils de Microsoft et de Google qui évaluent la présence de biais*.
  * Procéder à un examen approfondi des SIA utilisés pour évaluer d'autres SIA pour vérifier qu'ils ne présentent pas les mêmes défauts que ceux qu'ils prétendent évaluer.

Risk Assessments
Réaliser des évaluations de risque pour évaluer les facteurs de risque susceptibles de causer des dommages (p. 65 : mesures à prendre selon le niveau de risque identifié).

Education, training and standards
Fournir de la formation pour aider à MEO les normes éthiques dans l'utilisation et le dév. de l'IA;
Formuler des normes pour régir le travail des *data scientists*.

Collaboration

	Promouvoir la collaboration entre les chercheurs/universités et les entreprises pour garantir que l'IA est *ethical by design*. **Mechanisms for monitoring and improvement** Promouvoir une évaluation régulière des SIA et de la manière dont ils sont utilisés pour garantir que les principes éthiques sont pris en compte et pour évaluer si un SIA est encore adapté pour la tâche à accomplir. *Les évaluations initiales (avant le déploiement) sont essentielles mais insuffisantes pour évaluer l'impact de l'IA à travers le temps. **Consultation** Favoriser une consultation régulière public – parties prenantes – experts pour garantir que le dév. et l'utilisation de l'IA correspondent aux attentes du public, pour cerner des risques non identifiés au départ et pour trouver des solutions. ***HITL, SITL*** Adopter des principes de *design* HITL et SITL *HITL* : pour garantir que les humains conservent un rôle de supervision sur les technologies automatisées, que les erreurs sont corrigées et que les humains demeurent responsables. *SITL* : à prendre en compte lorsqu'une technologie a des effets sur la société dans son ensemble – ex : voitures autonomes.							
Entraves aux meilleures pratiques	-	-	**1. Définition du principe** **1.1 Définition de *fairness*** Manque de clarté relativement à ce qu'on entend par *fairness.* **1.2 Définition mathématique de *fairness*** **a.** Plusieurs des définitions mathématiques visant à définir ce que signifie *fairness* dans un algorithme fonctionnent bien lorsqu'elles sont mesurées sous un angle *x*, mais peuvent produire des résultats complètement différents lorsqu'elles sont mesurées sous un autre angle.	**1. Droit à l'oubli (consentement à jour)** Le droit à l'oubli peut être difficile à appliquer et à respecter, surtout si les données ont déjà été intégrées dans un SIA et qu'un modèle a déjà été entraîné. **2. Signalement des violations de données personnelles** Risque lié à cette pratique : perte de confiance des membres du public en la sécurité de leurs informations privées.	-	**1. *Black boxes*** Le fonctionnement interne de certaines technologies d'IA est difficile à expliquer. **2. Obstacle à l'efficacité** Les tentatives d'explication de tous les processus d'un algorithme peuvent ralentir ou nuire à son efficacité. **3. Explications complexes** Les explications peuvent être beaucoup trop	**Notion floue** La notion « d'impact » (dans « permettre la contestation […] d'un SIA qui a un *impact* significatif sur une personne ») est floue. ▪	**Innovation** Le principe de *accountability* pose le risque d'étouffer l'innovation en tenant les développeurs responsables – surtout en ce qui concerne l'attribution de responsabilité pour les conséquences imprévues.

			b. Parfois mathématiquement impossible de respecter toutes les mesures de *Fairness* / nécessité de faire des compromis dans la codification de ce principe dans les algorithmes. **2. Discrimination** **2.1 Ensembles de données non biaisés et résultats discriminatoires** Des ensembles de données non biaisés et précis (dont les informations d'entrée sont raisonnablement exactes) peuvent eux aussi entraîner des résultats discriminatoires : les *input data* proviennent du monde réel, et les données collectées dans le monde réel ne sont pas nécessairement équitables. **2.2 Identification des variables discriminatoires** Il est difficile de savoir si l'utilisation d'une variable en particulier entraînera des résultats discriminatoires. **2.3 Nécessité d'utiliser des variables discriminatoires** L'utilisation de variables discriminatoires peut être nécessaire pour des résultats équitables (*EX : prendre en considération la couleur de la peau pour adéquatement prédire le risque de cancer de la peau*). **2.4 Difficulté d'évaluation des biais** Cette capacité est intrinsèquement liée à la façon dont l'équité (f*airness*) est mesurée (**a**) et au niveau de transparence impliqué (**b**) :	**3. Open Data** Les individus peuvent être identifiés à partir de données non personnelles ou anonymisées en raison de la capacité de l'IA à détecter des modèles et à faire des inférences.		complexes et/ou compréhensibles que par certaines personnes (experts). **4. Secret commercial et « systèmes propriétaires »** Le secret commercial ou les systèmes propriétaires nuisent à l'explicabilité. *EX : les propriétaires d'un logiciel utilisé pour évaluer les performances d'enseignants considéraient les algorithmes utilisés comme des informations confidentielles => impossibilité d'évaluation externe par des humains.* **5. Risques de la transparence** La transparence totale peut être impossible ou indésirable (*EX : si elle entraîne des violations de la vie privée)*. Rendre le fonctionnement interne de l'IA ouvert au public pourrait le rendre susceptible d'être trafiqué (*gamed*).		

			a. Les évaluations du même SIA par des personnes différentes peuvent donner des conclusions différentes en raison de la façon de mesurer l'équité. **b.** Le manque de transparence ou la confidentialité du fct d'un logiciel empêche d'évaluer directement la façon dont un SIA pèse et évalue les *input data* et pose donc problème pour les évaluations externes.					
Stratégies	-	-	**1. Compromis entre les mesures de *fairness* et d'autres objectifs** **1.1** Faire preuve de transparence sur la façon dont certains objectifs sont priorisés pour que le public puisse faire un choix éclairé, et pour que les compagnies elles-mêmes puissent agir conformément aux attentes du public. **1.2** Le gouvernement et la société doivent tenir compte du degré de flexibilité dont les *designers* de SIA devraient bénéficier lorsqu'il s'agit de faire des compromis entre les mesures de *fairness* et d'autres priorités (ex : profit), **c.à.d**. : se demander si les avantages nets de l'algorithme justifient son existence et s'il est justifié dans la manière dont il traite différents groupes.	**1. Droit à l'oubli (consentement à jour)** Observer comment le droit à l'oubli est mis en œuvre et appliqué ailleurs (ex : UE). **2. Signalement des violations des données personnelles** Cette meilleure pratique devrait être soutenue par une éducation et une formation sur la protection des données, ainsi qu'une évaluation régulière des pratiques en matière de données pour que le public ait confiance en la sécurité de leurs informations privées. **3. Open data** **3.1** Utiliser des processus rigoureux de gestion des risques avec une documentation claire des	-	**Black boxes** **Autres possibilités pour atteindre un certain degré de transparence** : -Expliquer et définir les *input data*; -S'assurer que les résultats ne soient pas totalement inattendus ou inexplicables; -En cas de résultats inattendus, les revoir périodiquement jusqu'à ce qu'ils soient compris; -Tester le SIA et procéder à des examens et suivis externes. **Règlementation** Exiger la communication des priorités-clés et des mesures d'équité utilisées par un SIA.	-	-

			1.3 Un décideur humain doit être responsable lorsque les enjeux sont importants. **2. Résultats du SIA** Accorder une attention particulière au type de résultats qui constituent de la discrimination (pour être en mesure de déterminer quelles variables peuvent être incluses dans un SIA).	processus décisionnels qui guident la « publication ouverte » (*open publication*) des données anonymisées. **3.2** Ériger en infraction la réidentification délibérée d'informations gouvernementales anonymisées et rendues publiques.		**Recours** La mise en place de mécanismes de recours est importante dans les cas d'algorithmes de type *black box* dont les résultats impactent des individus. **Certains pays ont mis en œuvre le droit de demander un examen humain des décisions automatisées*. **Multidisciplinarité** Favoriser la multidisciplinarité. ***Intérêt public** Les *black boxes* sont inacceptables lorsque l'intérêt public est en jeu.		

4. CANADA

	La Déclaration de Montréal pour un développement responsable de l'IA, Université de Montréal (2018)									
	10 principes éthiques									
	Bien-être	**Respect de l'autonomie**	**Protection de l'intimité et de la vie privée**	**Solidarité**	**Participation démocratique**	**Équité**	**Inclusion de la diversité**	**Prudence**	**Responsabilité**	**Développement soutenable**
Enjeux, risques ou dilemmes	**Risques** <u>Utilisation de SIA contribuant à</u> : Augmentation du stress, de l'anxiété ou du sentiment de harcèlement liés à l'environnement numérique. **Enjeu** <u>Influence des SIA sur</u> : Conditions de vie et de travail des individus; Santé des individus; Exercice par les individus de leurs préférences; Exercice par les individus de leurs capacités physiques et intellectuelles.	**Risques** <u>Coercition des individus ou des groupes dans leurs choix</u> Influence des SIA sur la réalisation des objectifs moraux et sur la conception de la vie digne d'être vécue des individus : *- Utilisation de SIA par les institutions publiques pour promouvoir ou défavoriser une conception de la vie bonne;* *- Dév. ou utilisation de SIA pour prescrire aux individus un mode de vie particulier via la MEO de mécanismes contraignants.* Propagation d'infos peu fiables, de mensonges ou de propagande; Création de dépendances.	**Risques** Usage préjudiciable de SIA ou de SAAD pour l'intimité de la pensée et des émotions; Absence de droit à la déconnexion numérique des personnes dans leur vie privée; Contrôle restreint des personnes sur les infos relatives à leurs préférences et sur leurs données personnelles; Influence des SIA sur le comportement des personnes via la construction de profils de préférences individuelles; Utilisation de SIA conditionnée à l'abandon de la propriété des données personnelles; Compromission de l'identité personnelle, de la confidentialité des données et de	**Risques** SIA nuisant au maintien de relations humaines affectives et morales épanouissantes ou ne favorisant pas ces relations; Dév. de SIA ne favorisant par la collaboration (humains – SIA, humains-humains); MEO de SIA pour remplacer des personnes pour des tâches requérant une relation humaine de qualité ou MEO de SIA ne facilitant pas cette relation; Dév. de SIA stimulant des comportements cruels avec des robots qui prennent l'apparence d'êtres humains/animaux. **Enjeux** Solidarité intergénérationnelle;	**Risques** <u>Inintelligibilité</u> par les concepteurs du fct des SIA prenant des décisions affectant la vie, la qualité de vie ou la réputation des personnes; <u>Justification</u> inadéquate des décisions des SIA affectant la vie, la qualité de vie ou la réputation des personnes; <u>Inaccessibilité du code</u> des algorithmes nuisant à : → Vérification et contrôle par les autorités publiques compétentes; → Transparence des décisions publiques. <u>Confusion</u> sur la nature de l'interaction ou de la décision (SIA ou personne);	**Risques** Inégalités économiques et sociales; Création, renforcement ou reproduction de discriminations fondées sur les différences sociales, sexuelles, ethniques, culturelles et religieuses; Maintien de relations de domination entre les personnes et les groupes fondées sur la différence de pouvoir, de richesses ou de connaissance; Mauvaises conditions de travail.	**Risques** Uniformisation de la société par la normalisation des comportements et des opinions; Limitation de la liberté d'exprimer des idées et de communiquer des opinions; Constitution de monopoles de fait nuisant aux libertés individuelles; Enfermement des utilisateurs dans un profil d'utilisateur ou une bulle filtrante; Fixation des identités personnelles par le TX des données des activités passés des individus; Réduction des options de dév. personnel des individus.	**Risques** Usages néfastes des SIA ou des données d'utilisateurs; Utilisation détournée d'un SIA posant une probabilité élevée de danger sérieux pour la sécurité ou la santé publique; Compromission de l'intégrité ou de la confidentialité des données personnelles; Tests des SIA mettant en danger la vie des personnes, nuisant à leur qualité de vie ou portant atteinte à leur réputation ou intégrité psychologique.	**Risques** Dév. et utilisation de SIA contribuant à la déresponsabilisation des êtres humains lorsqu'une décision doit être prise; Pouvoir décisionnel indument transféré à un SIA.	**Risques** Inefficacité énergétique; Émissions de gaz à effet de serre; Déchets électriques et électroniques; Impacts sur les écosystèmes, le climat et la biodiversité; Gaspillage de ressources naturelles; Gaspillage de biens produits.

			l'anonymisation des profils personnels; Atteinte à la réputation d'une personne ou manipulation de personnes via l'utilisation de SIA pour imiter ou modifier des caractéristiques individuelles.	Efficacité de la gestion des risques par les SIA; Efficacité de la mutualisation des risques individuels et collectifs par les SIA.	<u>Non-divulgation</u> des dysfonctionnements ou d'informations liés aux SIA. **Enjeu** Influence des SIA sur la vie politique.					
Meilleures pratiques	Les SIA ne peuvent constituer une source de mal-être que si ce dernier permet d'engendrer un bien-être supérieur que l'on ne peut atteindre autrement.	**Encapaciter** Encapaciter les citoyens face aux technologies du numériques en assurant l'accès à différents types de savoir pertinents, le dév. de compétences structurantes (littératie numérique et médiatique) et la formation de la pensée critique. **Conception** Concevoir les SIA dans le but de réduire la propagation d'informations peu fiables, de mensonges et de propagande.	**Espaces d'intimité** Protéger de l'intrusion de SIA ou de SAAD les espaces d'intimité dans lesquels les personnes ne sont pas soumises à une surveillance ou à une évaluation numérique. **Déconnexion numérique** Explicitement offrir aux personnes le choix de la déconnexion à intervalle régulier, sans incitation à rester connecté. **SAAD** Garantir la confidentialité des données et l'anonymisation des profils personnels. **Contrôle** Permettre aux individus d'avoir un contrôle étendu sur les infos relatives à leurs préférences et sur leurs données personnelles (collecte, usage, dissémination);		**Justification** Justifier les décisions des SIA affectant la vie, la qualité de la vie ou la réputation des personnes dans un langage compréhensible et selon les mêmes exigences que celles applicables à un humain prenant le même type de décision. Exposer les facteurs et les paramètres les plus importants de la décision. **Signalement** Signaler la découverte d'erreurs de fct des SIA, d'effets imprévus ou indésirables, de failles de sécurité et de fuites de données aux autorités publiques compétentes, aux parties prenantes concernées et aux personnes affectées par la situation. **Vérification** S'assurer les SIA font ce pour quoi ils ont été programmés	**Reconnaissance** Reconnaître l'activité numérique des utilisateurs de SIA et de services numériques comme un travail qui contribue au fct des algorithmes et qui crée de la valeur. **Accès** Garantir l'accès aux ressources, aux savoirs et aux outils numériques fondamentaux à tous. **Communs algorithmiques, données ouvertes** Soutenir le dév. de communs algorithmiques et des données ouvertes pour les entraîner et les faire fonctionner.	**Considérer** Dès la conception des algorithmes, prendre en considération les multiples expressions des diversités sociales et culturelles pour le dév. et le déploiement des SIA. **Offre SIA** Diversifier l'offre de SIA pour chaque catégorie de service. **Milieux** S'assurer que les milieux de dév. de l'IA (recherche, industrie) soient inclusifs et reflètent la diversité des individus et des groupes de la société.	**Mécanismes** Développer des mécanismes qui tiennent compte du potentiel de double-usage de la recherche en IA et du dév. des SIA. **Restriction** Restreindre la diffusion publique ou l'accès libre à l'algorithme d'un SIA dont l'utilisation détournée peut représenter un danger sérieux pour la sécurité ou la santé publique. **Tests** Avant la mise en marché de SIA, s'assurer qu'ils satisfont à des critères rigoureux de fiabilité, de sécurité et d'intégrité; Tester adéquatement les SIA avant	**Pouvoir décisionnel** La décision finale doit revenir à un être humain dans tous les domaines où une décision qui affecte la vie, la qualité de la vie ou la réputation d'une personne doit être prise. **Humains responsables** Seuls des êtres humains peuvent être tenus responsables : - Des décisions issues de recommandations faites par des SIA et des actions qui en découlent; - De la décision de tuer; - Du crime ou du délit commis par un SIA autorisé par des personnes ou rendu possible par leur négligence.	<u>Équipements de SIA, leurs infrastructures numériques et objets connectés sur lesquels ils s'appuient</u> : Prévoir des filières de maintenance, de réparation et de recyclage dans une logique d'économie circulaire. **Soutien par les acteurs publics et privés** Soutenir le développement de SIA écologiquement responsables.

			Obtenir le consentement libre et éclairé des personnes pour la construction de profils de préférences individuelles.		et ce pour quoi ils sont utilisés. **Information** Informer l'utilisateur d'un service qu'une décision le concernant ou l'affectant a été prise par un SIA; S'assurer que l'utilisateur puisse identifier facilement qu'il interagit avec un SIA ou une personne. **Code** Rendre accessible (aux fins de vérification et de contrôle) le code des algorithmes (publics ou privés) aux autorités publiques compétentes et aux parties prenantes concernées; Rendre accessible le code des algorithmes de décision utilisés par les pouvoirs publics à tous (à l'exception des algorithmes présentant une probabilité élevée de danger sérieux en cas d'usage détourné.) **Recherche** S'assurer que la recherche dans le domaine de l'IA soit ouverte et accessible à tous. **Débat** Permettre aux citoyens de délibérer sur les paramètres sociaux,			leur mise en marché; Ouvrir ces tests aux autorités publiques compétentes et aux parties prenantes concernées. **Anticipation par toutes les personnes impliquées** Anticiper autant que possible les conséquences néfastes de l'utilisation des SIA; Prendre les mesures appropriées pour éviter ces conséquences. **Partage** Institutions publiques et entreprises dans les secteurs qui présentent un danger important pour l'intégrité personnelle et l'organisation sociale : Partager publiquement, à l'échelle mondiale, les erreurs et les failles découvertes dans les SIA et SAAD.		

					des objectifs et des limites d'utilisation des SIA publics ayant un impact important sur leur vie.					
Entraves	-	-	-	-	-	-	-	-	**Attribution de la responsabilité si SIA fiable + usage normal** Il est déraisonnable d'imputer la faute aux personnes impliquées dans le dév. ou l'utilisation d'un SIA lorsqu'un tort a été infligé par un SIA fiable ayant fait l'objet d'un usage normal.	-
Stratégies	-	-	-	-	-	-	-	-	-	-

5. ÉTATS-UNIS

	5.1 Ethical machines: The Human-centric use of artificial intelligence, Lepri et al. (2021)		
	3 domaines d'une IA centrée sur l'humain		
	1. Privacy and data ownership	**2. Accountability and transparency**	**3. Fairness**
Enjeux, risques ou dilemmes	**Enjeu** **Utilisation de la valeur des données comportementales VS préservation du droit à la vie privée.** **Risques** **1. Atteintes à la vie privée** Consentement - Extraction et utilisation de données personnelles sans consentement. Entraînement - Entraînement d'algorithmes à partir de *training data sets* contenant des informations sensibles relatives aux caractéristiques et aux comportements des personnes. Déductions et influence sur le comportement - Déductions algorithmiques sur des informations privées à partir de nouvelles sources de données comportementales (**incluant** les données pseudo-anonymisées) : *réseaux sociaux, données de tél. mobile, transactions de cartes de crédit, etc.* *EX : « Mentions J'aime » sur Facebook pour inférer orientation sexuelle, origine, religion, etc*.	**Risques** **1. Opacité des algorithmes** **1.1** ***Intentional opacity*** **1.2** ***Illeterate opacity*** (Manque de compétences techniques de la population pour comprendre comment les algorithmes fonctionnent et de quelle façon ils construisent des modèles à partir d'*input data*.) **1.3 Intrinsic opacity** (Algorithmes qui sont par nature difficiles à interpréter.) **2. Difficulté de développer des explications compréhensibles pour l'humain** **3. Conséquences découlant de 1 et 2** - Entrave à l'identification, l'évaluation et la réparation des préjudices (*EX : atteintes à la vie privée*) et des effets discriminatoires générés par des algorithmes; - Entrave à la validation de la valeur des outils de prise de décision basés sur l'IA pour l'intérêt public; - Entrave à la responsabilisation (*accountability*);	**Avantages perçus de l'IA** • Utilisation d'algorithmes pour une prise de décision plus objective – absence de préjugés humains, de conflits d'intérêts, de fatigue, etc. • Utilisation d'algorithmes pour résoudre des problèmes sociaux. **Risques** **1. Discrimination et préjugés à l'endroit de groupes défavorisés pouvant résulter de** : • Pondération dans les *input data*; *EX : algorithmes de police prédictive accordant une trop grande importance à l'attribut prédictif du code postal => association entre zones de haute criminalité et certains quartiers;* • Utilisation de *training* data biaisées; • Prise en compte implicite d'attributs sensibles (*sexe, origine, âge, niveau de revenu*) même si exclus de la tâche d'apprentissage; • Mauvaise utilisation de SIA dans des contextes spécifiques (*misuse*); • Décision d'utiliser un algorithme;

	- Influence sur le comportement au moyen d'interventions psychologiques ciblées rendues possibles par la prédiction des profils psychologiques des individus. (*Psychological targeting approcach*). <u>Gouvernements et entreprises</u> - Hausse de l'utilisation des technologies de vision par ordinateur pour détecter traits et attitudes des gens. <u>Gestion des données</u> - Difficulté pour les individus d'avoir un contrôle sur leurs propres données personnelles. <u>Vulnérabilités de confidentialité des SIA</u> - Absence de considération des attaques potentielles dans la conception d'un SIA, menant à des fuites d'infos confidentielles sensibles sur le modèle d'apprentissage ou sur les *training data* du modèle. <u>Vulnérabilités de robustesse et de fiabilité des SIA</u> **2. Asymétrie de pouvoir et d'information** - Inégalité d'accès aux ressources (données); - Difficulté pour les individus d'avoir un contrôle sur leurs propres données personnelles.	▪ Difficulté de savoir qui assume la responsabilité des décisions prises par des algorithmes ou prises à l'aide d'algorithmes; ▪ Difficulté pour les gens de comprendre les processus décisionnels basés sur l'IA et de savoir quand ils devraient s'opposer à une décision.	*=> Choix qui exclut inévitablement la prise en compte de certaines variables contextuelles.* **2. Déni d'opportunités ou de ressources envers des individus en raison de** : • La conduite <u>d'autres personnes</u> avec lesquelles ils partagent des caractéristiques (niveau de revenu, genre, origine ethnique, quartier, traits de personnalité, etc.)
Meilleures pratiques	**Méthodes de confidentialité** **a. Obscurcissement des données (*data obfuscation*)** Masquer les infos personnellement identifiables + autres données sensibles. **b. Anonymisation des données (*data anonymization*)** Supprimer les infos personnellement identifiables + autres données sensibles des ensembles de données. ***c. Entraînement antagoniste (adversial training)***	**A. Atténuer les différents types d'opacité** **1. *Intentional opacity*** Mettre en œuvre des interventions législatives en faveur de l'utilisation d'IA *open source*. **2. *Illeterate opacity*** Établir des programmes éducatifs visant par exemple : les décideurs politiques, les journalistes et les militants en *computational thinking* et en IA. Aider les personnes concernées par les décisions d'apprentissage automatique à recourir aux conseils d'experts techniques indépendants.	**Méthodes** **<u>Pour discrimination directe</u>** : ***Blindess approach*** Empêcher l'utilisation d'attributs sensibles dans la tâche d'apprentissage. **<u>Pour discrimination indirecte</u>** : **Formalisations mathématiques de l'équité <u>individuelle</u>** *EX : Joseph et al. (2016)* et *Hardnt et al. (2016) ont proposé des formalisations mathématiques du concept d'égalité des chances*. *****La **parité statistique** (approche d'équité de <u>groupe</u>) ne semble pas être présentée comme une meilleure pratique en

	Masquer les fonctionnalités pour éviter qu'un « attaquant » puisse déduire des infos sensibles à partir de ces fonctionnalités. **d. Ensembles de données synthétiques (*synthetic datasets*)** **e. Confidentialité différentielle (*differential privacy*)** **f. PPML (*Privacy-Preserving Machine Learning*)** Développer des algorithmes d'apprentissage automatique sécurisés et préservant la confidentialité (PPML) de manière à protéger la confidentialité des input data et/ou des modèles utilisés dans une tâche d'apprentissage. EX de PPML : **1) *Federated learning*** : Approche d'apprentissage-machine qui permet à des organisations de collaborer pour former un modèle en conservant les *training data* décentralisées dans des « nœuds » locaux. Les échantillons de données brutes de chaque entité sont stockés localement et jamais échangés. Seuls les paramètres de l'algo d'apprentissage sont échangés pour générer un modèle global. **2) *Encrypted computation*** Permet de protéger le modèle d'apprentissage en lui permettant de s'entraîner sur des données cryptées. Ainsi, l'organisation qui forme le modèle ne peut pas voir/divulguer les données sous une forme non chiffrée. **Contrôle par les individus de leurs données → coopératives de données** Partage volontaire et collaboratif des individus de leurs données personnelles. Permet de responsabiliser les individus, de leur permettre de mieux contrôler le cycle de	**3. *Intrinsic opacity*** Utiliser d'autres modèles d'apprentissage automatique plus faciles à interpréter par l'humain afin de caractériser les décisions prises par l'algorithme. **B. Fournir des explications** **Mise en œuvre d'un droit à l'explication** *EX : GDPR* 2019 **Différentes stratégies d'explications** **1. Explication simple** Concevoir des algorithmes d'apprentissage automatique qui sont intrinsèquement faciles à interpréter et à expliquer. **2. Explication globale** Fournir une explication globale du modèle d'apprentissage automatique. Ou **Explication locale** Fournir une explication pour une décision spécifique *EX : identifier quelles parties des* training data *sont responsables de la prédiction du modèle*. **3. Explication indépendante du modèle** Développer des explications applicables à tout type de modèle d'apprentissage automatique. (4 méthodes techniques d'explications indépendantes du modèle : *visualizations*, *influence methods*, *example-based explanations*, *knowledge extraction*.) *L'approche d'explication spécifique à un modèle donné (par opposition à indépendante) limite le choix des modèles souvent au détriment de modèles plus prédictifs et précis. **Explications compréhensibles pour l'humain** **Multidisciplinarité**	raison d'un manque de précision. Cette méthode exige qu'une proportion égale de chaque groupe visé par un attribut protégé reçoive chaque résultat possible. EX : allocation de prêt VS pas d'allocation de prêt → la précision de l'algorithme est compromise pcq contraint à prédire une proportion égale de retour sur investissement alors que les 2 groupes comportent des proportions différentes d'individus capables de rembourser leurs prêts.

	vie de leurs données et de favoriser un accès égalitaire aux ressources (données). *Fonctionnement :* **1.** *Le membre d'une coop de données est légalement propriétaire de ses données;* **2.** *Ces données peuvent être collectées dans son* "Personal Data Store" **3.** *Le membre peut ajouter et supprimer des données de son PDS ou en suspendre l'accès;* **4.** *Si le /les PDS sont hébergés dans la coopérative (et non dans des serveurs privés), la protection des données et la conservation des données sont assurées par la coop au profit de ses membres;* **5.** *La coop a une obligation fiduciaire légale envers ses membres (la coop est détenue et contrôlée par les membres).*	Encourager la recherche multidisciplinaire => modèles d'explications développés par les communautés *machine-learning* et HIC (*Human-computer-interaction*) et par les sciences cognitives et sociales. ***Exemple de stratégie issue de la recherche multidisciplinaire** ***T. Miller, 2019*** *(chercheur en IA, a analysé des recherches menées sur les processus d'explication humaine en sciences cognitives, en psychologie cognitive et sociale et en philosophie.)* 4 aspects à considérer pour construire des méthodes d'IA explicables qui peuvent être compréhensibles et utiles pour les humains : **1. Explications « contrastées » (contrastive**) – Les gens ne se demandent pas pourquoi un événement s'est produit, mais plutôt pourquoi cet événement s'est produit au lieu d'un autre. **2. Explications sélectives** – les explications doivent se concentrer seulement sur une ou quelques causes possibles. **3. Explications = « conversation sociale pour le transfert des connaissances »** et donc "the AI-driven explainer should be able to leverage the mental model of the human explainee during the explanation process". **4. Référence aux causes –** La référence aux associations statistiques dans les explications est moins efficace que la référence aux causes.	
	Multidisciplinarité Équipes multidisciplinaires de chercheurs de différents domaines*****, de praticiens, de décideurs et de citoyens pour co-développer et évaluer dans le monde réel les processus de prise de décision algorithmique conçus pour la transparence, l'équité et la responsabilité (*accountability*) qui respectent la vie privée. *******EX : apprentissage-automatique, IHC, sciences cognitives, psychologie sociale et cognitive, éthique, droit, philosophie.*		

Entraves	***Federated learning* – limites** L'approche de ***federated learning*** n'offre pas une garantie complète de la confidentialité des données sensibles pcq certaines caractéristiques des données brutes pourraient être mémorisées au cours de l'apprentissage de l'algorithme et extraites par la suite.	**Opacité intentionnelle requise** Les interventions législatives peuvent se heurter à la nécessité d'utiliser l'opacité intentionnelle pour : • Protéger la propriété intellectuelle; • Éviter la manipulation (*gaming*) du système (fraudes, spams, scams); • Protéger un intérêt commercial légitime (*EX : si la décision algorithmique règlementée par des interventions législatives en est une commerciale).* **Modèles simples = moins précis** Les modèles d'apprentissage automatique plus faciles à interpréter ont tendance à être moins précis et ne permettent pas d'expliquer parfaitement les performances d'un algorithme de type *black box*. **Explications globales difficiles à obtenir** Les explications globales sont difficiles à obtenir, en particulier pour les modèles caractérisés par un grand nombre de paramètres.	**Identification de moyens** La prévention de la discrimination nécessite l'identification de moyens pour détecter ces effets. Cette identification peut être difficile lorsque les choix sont faits par des humains. **Définition d'équité** Difficulté de définir l'équité – varie selon le contexte, les perceptions, les visions du monde, etc. **Contraintes mathématiques** Il est impossible de satisfaire simultanément les contraintes mathématiques des multiples formalisations de l'équité => impossibilité d'une seule définition universellement acceptée de l'équité algorithmique.
Stratégies	**Confidentialité différentielle** La **confidentialité différentielle** peut compléter le *federated learning* en offrant des garanties de garder privée la contribution d'organisations individuelles (« nœuds »).	**Transparence pas tjrs nécessaire** La transparence n'est pas toujours nécessaire pour la responsabilisation (*accountability*). ⇨ Certaines méthodes de calcul sont capables de « rendre des comptes » même lorsque certaines informations "fairness-sensitive" sont cachées. Aussi, pratiques de *PPML*, *federated learning* et de *encrypted computation* (voir meilleures pratiques pour "Privacy an data ownership") suggèrent des solutions à la responsabilité sans divulguer des données sensibles ou des algorithmes. ***Explication post-hoc*** Construire des modèles complexes et très précis (blackbox) et ensuite utiliser différentes techniques permettant de fournir les explications requises (sans connaître le fonctionnement interne du modèle d'apprentissage automatique d'origine).	**Prévention des effets discriminatoires** Mettre en œuvre des lois et règlements pour l'utilisation d'algorithmes d'apprentissage comme **outils de détection** et de prévention de la discrimination. **Définition d'équité** Encourager la **contribution multidisciplinaire** de chercheurs de différents domaines (droit, philosophie morale et politique, ML) pour concevoir, évaluer et valider dans le monde réel des mesures d'équité pour différentes tâches.

		Explications locales ou combinaison locales/globales • Opter pour la stratégie des explications locales; • Combiner les explications globales et locales ("A recent promising line of work is trying to combine the benefits of global and local explanations".)	

Légende : Recommandations applicables aux gouvernements Recommandations applicables aux organisations qui développent, déploient ou utilisent des SIA	**5.2 Responsible AI Global Policy Framework, ITechLaw (2019)[26]**							
	Ethical Purpose and Societal Benefits	**Accountability**	**Transparency and Explainabilty**	**Fairness and Non-Discrimination**	**Safety and Reliability**	**Open Data and Fair Competition**	**Privacy**	**AI and Intellectual Property**
Enjeux, risques ou dilemmes	**Risques** Dév., déploiement ou utilisation de l'IA incompatible avec le respect de l'action humaine (*human agency*) et des droits fondamentaux; Conséquences (intentionnelles ou non) de SIA incompatibles avec les principes de *beneficience*, de *non-maleficience* et des autres principes de cette grille; "*Weaponised AI*"; Impacts de l'implantation de l'IA dans les milieux de travail (ex : survenance d'inégalités socio-économiques); Diffusion de contenu (ex : informations fausses ou trompeuses, contenus haineux) pouvant affecter de manière préjudiciable les individus, groupes ou institutions démocratiques; **Enjeux** Confiance du public dans les SIA;	**Risque** "Accountability gaps" des cadres juridiques et règlementaires actuels applicables aux SIA. **Enjeu** Plus le niveau d'autonomie du SIA est élevé et plus la criticité des résultats est élevée, plus le degré de responsabilité est élevé.	**Risques** MEO d'exigences réduites quant à la transparence et l'explicabilité d'une décision lorsque le processus décisionnel est contrôlé par un SIA et non un humain; Explications incompréhensibles pour l'humain; Informations insuffisantes sur le SIA nuisant à : - La responsabilité (*accountability*) des développeurs, des « déployeurs » et des utilisateurs de SIA; - La vérification des décisions (vérifier si justes et impartiales). **Enjeux** Préservation de la confiance du public dans les SIA; Détermination de l'intensité des obligations de transparence et d'explicabilité;	**Risques** SIA biaisés ou erronés (ex : groupes défavorisés incorrectement représentés dans les *training data*) menant à : - Décisions injustes ou discriminatoires; - Perpétuation et exacerbation de biais. Décisions prises selon des normes s'écartant des processus décisionnels menés entièrement par des humains; Utilisation de SIA comme argument justifiant l'exemption ou l'atténuation du principe d'équité. **Enjeux** Les SIA reflètent les objectifs, connaissances et expériences des créateurs et les ensembles de données utilisés; La MEO efficace du principe de *Fairness and Non-Discrimination* dépend de la MEO des principes	**Risques** SIA qui ne fct pas sur la base de principes clairement définis ou qui n'y adhère pas; Absence de limites restreignant les pouvoirs de décision du SIA; SIA formé à partir de données imprécises ou erronées. **Enjeux** Acceptabilité et confiance du public dans les SIA; Principes éthiques et moraux non uniformes (influencés par des considérations géographiques, religieuses, sociales); Cadres juridiques existants (ex : responsabilité du fait des produits) non adaptés aux caractéristiques propres aux SIA.	**Enjeux** **Concurrence** Dév. et déploiement des SIA dans l'intérêt des consommateurs; Accès pour les gouvernements aux SIA innovants susceptibles de présenter un avantage particulier pour la société ou de faire progresser la technologie. **Risques** Partage ou octroi de licence de données non conforme aux obligations ou exigences légales, réglementaires, contractuelles ou autres exigences relatives aux données concernées (confidentialité, sécurité, liberté d'information).	**Risques** Impacts des SIA sur la vie privée; Règlementation des SIA en matière de vie privée + frein à **l'**innovation; **Enjeux** Conflit entre utilisation croissante des SIA pour gérer les données privées **ET** protection règlementaire internationale croissante accordée aux données personnelles et aux autres données privées; Respect des exigences des différents régimes législatifs nationaux.	**Risques** Lois actuelles sur la PI pas suffisamment « équipées » pour faire face à la création d'œuvres par des IA. **Enjeu** Définition de la notion de PI dans un contexte d'IA et protection des œuvres résultant de SIA.

[26] ITechLaw (*International Technology Law Association*) a publié une version mise à jour de cette publication en 2021. La mise à jour comporte aussi un outil d'évaluation de l'impact de l'IA selon des facteurs de risque. Voir : https://www.itechlaw.org/ResponsibleAI2021.

	Détermination du « moment d'intervention » des humains en cas de défaillance d'un SIA; Impacts des SIA sur l'environnement; Identification des contenus interdits dans le respect des droits à la dignité et à l'égalité et du droit à la liberté d'expression. Ajouter : risques, enjeux, dilemmes de la nouvelle section Human Agency and Autonomy p. 24 PDF.		Dépendance d'autres principes (*Accountability*, *Fairness and Non-Discrimination*, *Safety and Reliability*, *Privacy*, *Lawful Use*, *Consent*) au principe de *Transparency and Explainabilty*.	de *Transparency* et de *Accountability*.				
Meilleures pratiques	**Objectifs** Identifier les objectifs du SIA; Exiger que les objectifs de la MEO de SIA soient identifiés. **Surveillance, évaluation** Surveiller la MEO du SIA; Évaluer les implications sociales, politiques et environnementales du SIA via une *Responsible AI Impact Assessment* qui évalue les risques de préjudices et qui propose des stratégies pour les atténuer. **Travail** Permettre la participation des employés concernés dans l'implantation de l'IA dans le milieu de travail; Réaliser une *Responsible AI Impact Assessment* pour déterminer les effets d'une	**Responsabilité** Les humains doivent toujours rester responsables des actes et des omissions des SIA; Désigner une/des personne(s) responsable(s) du respect par l'organisation des 8 principes; Communiquer cette/ces désignation(s) sur demande. **Principes éthiques** MEO des politiques et des pratiques pour donner effet aux 8 principes (ou d'autres principes adoptés) comme : → Mettre en place des procédures pour répondre aux plaintes et aux demandes de renseignements;	**Informations** Lorsque l'IA est utilisée dans les processus de prise de décision, fournir des infos concernant : - Le fait qu'un SIA est utilisé; - Le but du SIA; - La manière dont il sera ou pourra être utilisé; - Les types d'ensembles de données utilisés; - La logique de la prise de décision - en termes humainement compréhensibles. **Intensité des obligations de transparence et d'explicabilité** Augmente si la sensibilité des ensembles de données utilisées augmente;	**Éducation** Sensibiliser et éduquer sur les possibilités et les limites des SIA; Informer les utilisateurs que les SIA reflètent les objectifs, connaissances et expériences des créateurs et les ensembles de données utilisés. **Réparation** Mettre en place un moyen efficace permettant aux utilisateurs de demander réparation lorsqu'une situation discriminatoire ou injuste survient. **Dév. et surveillance** Prioriser l'équité (*fairness*) dans le *design* de l'IA en s'attaquant à la question des biais dans les algorithmes et les données dès le début; Mettre en place des comités d'éthique et des codes de conduite;	**Principes** Définir et valider périodiquement les principes éthiques et moraux qui sous-tendent le SIA et clairement limiter ses pouvoirs de décision. **Ajustement** Envisager de rendre « ajustables » les SIA pour qu'ils puissent répondre aux normes locales (ex : considérations géographiques, religieuses, sociales). **Tests** Tester les SIA lors de leur fct. **Normes** Soutenir et participer à la coordination internationale pour élaborer des normes	**Concurrence, *open data*** Participer à la coordination internationale (ex : via des organismes comme l'OCDE et le *International Competition Network*); Réaliser des examens réguliers pour vérifier que les cadres du droit de la concurrence (*competition law frameworks*) et les outils d'application à la disposition des autorités compétentes sont suffisants et efficaces; Veiller à ce que les données détenues par les organismes du secteur public soient accessibles et ouvertes; Encourager et faciliter les infrastructures nationales nécessaires pour promouvoir le libre accès aux ensembles de données à des fins de recherche et/ou d'utilisation non commerciale – à cet effet, envisager des modèles d'accès « à 2 niveaux » (accès gratuit pour les fins	Déterminer si les processus actuels au sein de l'organisation doivent être mis à jour pour garantir que le respect de la vie privée est une considération centrale; Envisager la MEO de garanties opérationnelles pour protéger la confidentialité (ex : principes de *privacy by design* spécifiquement adaptés aux caractéristiques des SIA); Nommer un *AI ethics officer* ayant pour mission d'examiner la conformité éthique et réglementaire de l'utilisation de l'IA de l'organisation.	Enquêter sur la manière dont les œuvres créées par l'IA peuvent être davantage protégées, sans créer de nouveau droit de la PI; Avoir recours aux lois existantes sur les droits de PI; Prendre les mesures nécessaires pour protéger les droits sur les œuvres d'IA : - Revendication ou obtention de droits d'auteur; - Obtention de brevets; - Dispositions contractuelles.

	MEO de l'IA dans le milieu de travail; Identifier les secteurs de l'économie où les travailleurs humains sont le plus touchés par la MEO de l'IA; Solliciter les commentaires et les données en provenance des différentes parties prenantes (industrie, employés et autres) concernant l'impact des SIA dans les milieux de travail; Développer un « forum ouvert » pour le partage des expériences et meilleures pratiques; Promouvoir des politiques éducatives qui permettent aux enfants d'acquérir les compétences requises par la « nouvelle économie » et qui favorisent l'apprentissage continu (*life-long learning*); Encourager la création d'opportunités permettant aux adultes d'acquérir de nouvelles compétences, en particulier pour ceux « déplacés » par l'automation; Envisager la possibilité de nouveaux systèmes de protection sociale et d'avantages sociaux pour réduire les inégalités socio-économiques causées par l'introduction de l'IA. **Environnement** Évaluer l'impact environnemental global des SIA (consommation de ressources, coûts énergétiques du stockage et du TX des données, efficacité énergétique nette, avantages	→ Former les employés et leur communiquer des informations sur les politiques et les pratiques de l'organisation. **Régime de responsabilité** Promouvoir les 8 principes ou des considérations similaires dans l'élaboration des cadres juridiques et règlementaires; Ne pas accorder de personnalité juridique distincte aux SIA.	Augmente si le résultat décisionnel d'un SIA augmente en importance; Doit tenir compte de la mise en balance des intérêts de la personne soumise à la décision (ses attentes raisonnables) et des intérêts de l'organisation qui prend la décision. **Transparence et explicabilité *by design*** S'assurer le *design* du SIA permet de faciliter la MEO des exigences de transparence et d'explicabilité; Préférer l'option la plus transparente lorsqu'un choix doit être fait entre différentes architectures +/- opaques; **Recours** Mettre en place un moyen efficace permettant aux utilisateurs de demander réparation lorsque les organisations ne sont pas transparentes.	Adopter des normes à l'échelle de l'industrie et des labels de qualité internationalement reconnus (*internationally recognised quality seals*); Avoir recours de façon périodique à des évaluations et des tests indépendants pour tester les SIA qui ont un impact social important; Accorder une attention particulière aux groupes défavorisés lors du dév. (ex : représentation dans les *training* data) et de la surveillance des SIA. **Multidisciplinarité** Favoriser la collaboration étroite d'experts techniques des domaines liés à l'IA avec des statisticiens et des chercheurs en sciences sociales; Favoriser la collaboration gouvernement – organisations qui travaillent avec l'IA – public.	internationales pour le dév. et le déploiement de SIA fiables.	universitaires et de recherche, accès payant pour les fins commerciales); Ouvrir l'accès aux ensembles de données de l'organisation et/ou octroyer des licences pour ces derniers, par ex. via des fiducies de données; Développer et déployer les SIA sur une base de *compliance by design* par rapport au droit de la concurrence + *antitrust*.	**IA comme outil** Avoir recours aux technologies d'IA pour se conformer aux obligations en matière de confidentialité. **Régimes législatifs** Organisations et gouvernements : Prévoir d'autres bases légales pour la collecte et le TX des données personnelles par les SIA.	

	environnementaux potentiels); Promouvoir et MEO les utilisations de SIA qui visent la neutralité carbone globale ou la réduction des émissions de carbone. Encourager la recherche en IA sur la réduction du gaspillage de ressources nuisibles à l'environnement et les inefficacités. Mettre en place des politiques pour garantir que les données pertinentes sont accessibles et utilisables d'une manière cohérente avec le respect des autres principes. **Confiance du public** Envisager la MEO de règles garantissant une enquête complète et transparente sur les résultats nuisibles ou imprévus de SIA s'étant produits en cours d'utilisation; Utiliser ces enquêtes pour ajuster le cadre règlementaire des SIA, en particulier pour déterminer comment et quand les opérateurs humains devraient intervenir en cas de défaillance; **"Weaponised AI"** Respecter les principes et les normes de droit international humanitaire sur l'utilisation des armes et le droit international des droits humains dans l'utilisation de systèmes d'armes létaux autonomes (*lethal autonomous weapons systems – LAWS*);							

	MEO des mécanismes multilatéraux pour contrôler le respect des accords internationaux concernant le dév. éthique, l'utilisation et le commerce des *LAWS*. Informer les employés concernés qu'ils sont affectés à des projets liés aux *LAWS*. Ne pas développer, vendre ou utiliser de *LAWS* capables de sélectionner et *d'engage* des cibles sans contrôle et surveillance humains. **Informations fausses ou trompeuses** P/r aux organisations qui dév., déploient ou utilisent des SIA pour filtrer ou promouvoir le contenu informatif sur Internet : Minimiser la diffusion d'informations fausses ou trompeuses lorsqu'il y a un risque important que ces informations puissent entraîner un préjudice important pour des individus, des groupes ou des institutions démocratiques; Permettre aux utilisateurs de signaler un contenu potentiellement dangereux en temps opportun; Permettre aux fournisseurs de contenu (*content providers*) de contester la suppression de leur contenu en temps opportun; Fournir des directives claires aux organisations pour les aider à identifier les contenus interdits d'une manière qui respecte les droits à la dignité et à							

	l'égalité et le droit à la liberté d'expression. **Recherche** Encourager la recherche sur les moyens pour l'IA d'identifier/supprimer de manière proactive les contenus illégaux d'une manière compatible avec la liberté d'expression; ***Exigence** Les tribunaux doivent rester les « arbitres ultimes » en ce qui concerne le contenu licite.							
Entraves	-	-	-	-	-	**Secteur public** *Open data* + possibilité de conflit avec le mandat du secteur public de récupérer l'investissement des contribuables dans la collecte et la conservation de ces données. **Secteur privé** - Pas tjrs possible de rendre les ensembles de données ouverts; - Des organisations peuvent décider de ne pas publier leur SIA (logiciel *open source* OU octroi d'une licence commerciale). **Méconnaissance** Méconnaissance des avantages à tirer des données *open access*, des structures par lesquelles les ensembles de données peuvent être partagés, et des processus par lesquels les données peuvent être « adaptées » à un *open access* (ex : *API standardisation*, pseudonymisation, agrégation).	-	-
Stratégies	-	-	-	-	-	Mettre en place des programmes d'incitation ciblés pour les organisations;	-	-

						Soutenir les initiatives en matière de *open data* dans le secteur public ou privé en fournissant des conseils et en menant des recherches (notamment pour partager les avantages à tirer de l'accès ouvert aux données); Encourager les organisations qui décident de ne pas partager leurs SIA comme logiciels *open source* à accorder une licence commerciale; Encourager les organisations qui ne publient pas *l'intégralité* de leurs SIA (*open source* OU licence) à accorder une licence pour le plus grand nombre possible des composantes qui pourraient être réutilisées dans d'autres cas d'utilisation de l'IA.		

6. ANGLETERRE

	6.1 An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations, Floridi et al. (2018) *Inclut : A Unified Framework of Five Principles for AI in Society Floridi & Cowls, 2019				
	5 principes éthiques				
	1. Beneficience	**2. Non-maleficience**	**3. Autonomy**	**4. Justice**	**5. Explicability** Comprend : **intelligibility** et **accountability**
Enjeux, risques ou dilemmes	**Enjeux** : • Promotion du bien-être des personnes; • Promotion du bien-être de la planète (durabilité environnementale); • Préservation de la dignité humaine.	**Risque** : Survenance de préjudices accidentels ou délibérés résultant respectivement d'un usage excessif *(overuse)* ou d'un mauvais usage *(misuse)* de l'IA. *EX : atteinte à la vie privée.* **Enjeu** : Sécurité entourant les capacités actuelles et futures de l'IA.	**Opportunités IA VS risques corrélatifs** **Opportunités IA** : • Développement de nouvelles capacités et compétences; • Amélioration du libre-arbitre. **Risques corrélatifs** : Délégation trop importante du pouvoir décisionnel humain à un agent autonome entraînant : • Absence de responsabilité; • Réduction de la capacité humaine à surveiller la performance des SIA, à prévenir et à corriger les erreurs ou les dommages; • Déqualification dans des domaines qui requièrent des compétences spécifiques (*EX : aviation*) créant des vulnérabilités	**Opportunités IA VS risques corrélatifs** **Opportunités IA** : • Réalisation de soi (automatisation de certaines tâches => libère du temps pour d'autres activités); • Développement de nouvelles capacités et compétences; • Élimination de la discrimination; • Création et partage d'avantages; • Prévention de nouveaux préjudices. **Risques corrélatifs** : • Dévalorisation trop rapide de certaines compétences entraînant une transition inéquitable caractérisée par diverses conséquences individuelles (I) et sociétales (S) comme : I : atteinte à l'identité personnelle ou à l'estime de soi, préjudice économique (par ex. en cas de licenciement). S : distribution inéquitable des coûts et des bénéfices liés à l'IA, fragilisation des structures sociales existantes et discrimination (via par ex. les biais dans les ensembles de données d'entraînement). **Enjeux** • Attribution de la responsabilité et mécanismes de recours en cas d'effets indésirables; • Législation et règlementation ne suivant pas le rythme de développement des technologies d'IA; • Capacité des institutions existantes (ex : tribunaux) à réparer les erreurs et les préjudices résultant de l'utilisation de l'IA;	**Enjeu** : Confiance du public. **Risques** : Absence d'explication ou incompréhension du fonctionnement de l'IA pour la majorité des gens => entrave aux 4 autres principes : *Beneficience, Non-maleficience* Requiert de comprendre le bien/mal que l'IA fait réellement à la société et de quelle façon; *Autonomy* Requiert de comprendre comment l'IA réagirait à la place d'un humain pour faire le choix de déléguer ou non le pouvoir décisionnel; *Justice* Requiert de comprendre les raisons pour lesquelles un résultat nuisible s'est produit afin de tenir les personnes et organisations responsables.

			dangereuses en cas d'attaque ou de dysfonctionnement.	• Confiance du public.	
Meilleures pratiques[27]	**DÉVELOPPER** **(8)** Développer un **"trust comparison index"** à partir de paramètres évaluant la fiabilité des produits et services d'IA pour permettre aux consommateurs de les comparer et pour créer une compétitivité autour du développement d'une IA sûre et socialement bénéfique. À plus long terme, envisager un système de certification plus large. **ENCOURAGER FINANCIÈREMENT** : **(12)** Le dév. et l'utilisation de technologies d'IA socialement préférables (et non simplement acceptables) **et** favorables à l'environnement (et non simplement durables). => Une « **approche par défi »** pourrait permettre d'encourager la créativité et de promouvoir la concurrence dans le dév. de solutions d'IA éthiques qui favorisent le bien commun. **(13)** Un effort de recherche adapté aux spécificités de l'IA et ayant pour **mission** de faire progresser l'IA pour le bien social. **(15)** L'inclusion de considérations éthiques, juridiques et sociales dans les projets de recherche en IA et en //, encourager l'examen régulier de la législation pour tester dans quelle mesure elle favorise l'innovation socialement positive.	**DÉVELOPPER** **(9)** Développer une nouvelle **agence de surveillance** chargée de la protection du bien-être public via l'évaluation scientifique et la supervision des produits, logiciels, SIA, services d'IA. **(9)** Développer un **système de surveillance** "**post-release"** pour l'IA. Ce système devrait comporter des obligations en matière de déclaration (*reporting-duties*) et des mécanismes de signalement (*reporting mechanisms*). **ENCOURAGER FINANCIÈREMENT** : **(16)** Le développement et l'utilisation de **zones spéciales déréglementées** pour les tests empiriques et le développement de SIA – *protection by design*. *EX : « laboratoire vivant », « autoroutes d'essai »*	**DÉLÉGATION** ***"Decide-to-delegate" model*** Les humains doivent toujours avoir le choix (prendre la décision ou déléguer). Toute délégation doit être réversible. **ÉVALUER** **(2)** Via des mécanismes participatifs, évaluer quelles tâches et fonctionnalités de prise de décision ne devraient <u>pas</u> être déléguées à des SIA. L'évaluation devrait prendre en compte la législation existante et être soutenue par un dialogue permanent entre toutes les parties prenantes – gouv, industrie, société.	**NOUVELLES SOLUTIONS** Envisager de nouvelles solutions pour une transition juste. *EX : envisager une forme de revenu de base universel.* **ÉVALUER** **(1)** Évaluer la capacité des institutions existantes (*EX* : *tribunaux*) à réparer les erreurs et préjudices résultant de l'utilisation de SIA. **ÉVALUER** **(3)** Évaluer si la **réglementation** en vigueur est suffisamment fondée sur l'éthique pour fournir un cadre législatif capable de suivre le rythme des développements technologiques. **DÉVELOPPER** **(7)** Développer un processus ou mécanisme de **recours** accessible et fiable qui attribue clairement la responsabilité aux personnes et/ou organisations concernées en cas d'effets indésirables d'un SIA. **Peut être utile de s'inspirer de l'industrie aérospatiale, qui dispose d'un système pour traiter les conséquences indésirables de façon approfondie.* **(7)** Envisager des **solutions institutionnelles supplémentaires**, nationales ou supranationales, en cas d'identification d'un manque de capacité des institutions existantes à réparer les préjudices (*recommandation #1 ci-dessus*) <u>EX de solutions institutionnelles nationales ou supranationales</u> : -Ombudsman; -Processus pour enregistrer une plainte (semblable à une demande d'accès à l'information); -Assurance responsabilité civile obligatoire pour certaines catégories d'IA. **DÉVELOPPER** **(6)** Développer des **mécanismes d'audit** pour les SIA afin d'identifier les conséquences indésirables comme les préjugés injustes.	**DÉVELOPPER** **(4)** Développer, en multidisciplinarité (associations professionnelles, experts en sciences, en affaires, en droit et en éthique) des **cadres spécifiques à différents secteurs pour améliorer l'explicabilité** des SIA qui prennent des décisions socialement importantes de manière à permettre aux individus d'obtenir une explication claire du processus de prise de décision, en particulier en cas de conséquences indésirables. **DÉVELOPPER** **(8)** Développer un **"trust comparison index**"à partir de paramètres évaluant la fiabilité des produits et services d'IA pour améliorer la compréhension des consommateurs. **(8)** Charger les parties qui exploitent ou se servent de SIA et qui en tirent profit de financer ou d'aider à développer des **"AI literacy programs**" pour les consommateurs. **ENCOURAGER FINANCIÈREMENT** : **(16)** Le développement et l'utilisation de **zones spéciales**

[27] Les 20 recommandations sont classées dans l'article en 4 catégories : *Assessment*, *Development*, *Incentivisation* et *Support*. Ces recommandations ont été classées dans la grille en fonction du ou des principes éthiques qui leur correspondaient le plus.
Les pratiques qui ne sont pas numérotées **(#)** se retrouvaient dans la 1re partie de l'article et non dans la section des 20 recommandations.

		SOUTENIR **(18)** Soutenir l'élaboration de codes de conduite d'autorèglementation comportant des obligations éthiques spécifiques pour les professions liées aux données ou à l'IA. Mènerait à une certification « d'IA éthique » par le biais de *trust labels* pour s'assurer que les gens comprennent les mérites de l'IA éthique et qu'ils l'exigent des fournisseurs.		**(6)** Pour faire face aux risques graves dans les secteurs à forte intensité d'IA : - Développer un « mécanisme de solidarité » (par ex. en collaboration avec le secteur des assurances); - Développer d'autres mécanismes multipartites en amont pour atténuer ces risques. **(11)** Développer des instruments juridiques et des modèles de contrats pour parvenir à une collaboration homme-machine fluide dans les milieux de travail. **(11)** Envisager la création d'un **fonds** d'ajustement de l'IA (dans article Floridi "*European AI Adjustement Fund*") pour favoriser une innovation inclusive et faciliter la transition vers de nouveaux types d'emplois. **SOUTENIR** **(18)** Soutenir l'élaboration de **codes de conduite** d'autorèglementation comportant des obligations éthiques spécifiques pour les professions liées aux données ou à l'IA. Mènerait à une certification « d'IA éthique » par le biais de *trust labels* pour s'assurer que les gens comprennent les mérites de l'IA éthique et qu'ils l'exigent des fournisseurs. **(19)** Soutenir la **capacité des CA** d'entreprises à assumer la responsabilité des implications éthiques des technologies d'IA, via, par ex : • Meilleure formation des CA; • Comités d'examen éthique.	**déréglementées** pour les tests empiriques et le développement de SIA pour contribuer à une éducation pratique et à la promotion de la responsabilité (*accountability*).
				DÉVELOPPER **(5)** Développer des procédures juridiques appropriées et améliorer l'infrastructure informatique du système judiciaire pour permettre **l'examen des décisions algorithmiques devant les tribunaux**. À cet effet, créer un cadre spécifique au système judiciaire pour l'explicabilité de l'IA (*lien avec recommandation #4 dans la section explicabilité*).	
	Recommandations qui semblent applicables aux 5 principes **Observatoire (recommandation #10)** - Développer un observatoire pour l'IA chargé notamment d'émettre des lignes directrices et des recommandations.				

	Approche multipartite (recommandation #14) - Encourager financièrement la coopération et le débat multidisciplinaire en ce qui a trait aux intersections entre la technologie, les questions sociales, le droit et l'éthique. Avoir recours à un groupe diversifié (genre, classe, ethnicité). **Participation publique (recommandation #17)** - Encourager financièrement la recherche sur la perception et la compréhension du public de l'IA et la MEO de mécanismes de consultation publique en vue de mesurer l'opinion du public et de co-créer des politiques, des normes, des règles et des bonnes pratiques pour l'IA. *EX de mécanismes : sondages d'opinion, groupes de discussion, proposition d'exemples simulés de dilemmes éthiques propres à l'IA, expériences dans des labos de sciences sociales, etc.* **Éducation (recommandation #20)** Soutenir la création de programmes éducatifs et d'activités de sensibilisation du public par rapport à l'impact sociétal, juridique et éthique de l'IA. *EX :* — Programmes scolaires faisant de l'informatique une discipline de base; — Programmes de qualification dans les entreprises pour éduquer les employés sur l'impact sociétal, juridique et éthique du fait de travailler avec l'IA; — Inclure l'éthique et les droits humains dans les programmes de formation scientifique et d'ingénierie qui travaillent avec des systèmes informatiques et des SIA; — Programmes similaires pour le grand public et pour les personnes impliquées à chaque étape de la gestion de la technologie d'IA (notamment : fonctionnaires, politiciens et journalistes.)				
Entraves	-	-	-	-	-
Stratégies	-	-	-	-	-

	6.2 How to Design AI for Social Good: Seven Essential Factors, Floridi et al. (2020)[28]						
	1. Falsifiability and incremental deployment	***2. Safeguards against the manipulation of predictors***	***3. Receiver-contextualised intervention***	***4. Receiver-contextualised explanation and transparent purposes***	***5. Privacy protection and data subject consent***	***6. Situational fairness***	**7. Human-friendly semanticisation**
	Principes éthiques correspondants						
	Beneficience[29]						
	Nonmaleficience	***Nonmalificience***	***Autonomy***	***Explicability***	***Nonmaleficience, Autonomy)***	***Justice***	***Autonomy***
Enjeux, risques ou dilemmes	**<u>Enjeu</u>** : **Fiabilité du SIA**	**<u>Risques</u>** **1. Altération du pouvoir prédictif de l'IA** La manipulation des *input data* par des personnes empêche le déploiement d'interventions individuelles socialement bonnes par le SIA, et nuit ainsi à son pouvoir prédictif. **Particulièrement si c'est la personne de qui on veut prédire le comportement qui modifie les données d'entrée/variables prédictives de son comportement*. **2. Traitement inefficace d'un problème par un recours excessif aux indicateurs non-causaux** Le recours excessif aux indicateurs corrélés, mais non-causaux, risque de faire en sorte que le SIA s'attaque aux *symptômes* d'un problème plutôt qu'à ses causes.	**<u>Risque :</u>** **Inefficacité des interventions d'un SIA par une trop grande (a) ou trop modeste (b) atteinte à l'autonomie d'un utilisateur** : **a. Si atteinte trop grande** : risque de rejet de la technologie par l'utilisateur => interventions futures nécessaires rendues impossibles. **b. Si absence d'atteinte ou atteinte trop minime** : risque d'interventions futures incorrectement contextualisées en raison d'un manque de données pertinentes *(c.à.d. les réponses de l'utilisateur sur lesquelles se base le SIA pour ses interventions futures.)*	**<u>Risque</u>** : **Explication inefficace** Une explication inefficace mine la confiance dans le SIA, ne favorise pas l'adoption de solutions et ne protège pas l'autonomie du destinataire. **<u>Dilemme</u> :** **Avantages VS inconvénients de la transparence des objectifs d'un SIA et des motivations des développeurs (a et b) :** **a. Avantages** : favorise l'autonomie de l'utilisateur, révèle les véritables intentions du SIA, diminue les malentendus et les risques de préjudices. **b. Inconvénients :** - Révéler les objectifs endogènes du SIA peut mener au rejet de la technologie par les utilisateurs. - Révéler un objectif peut le rendre moins susceptible d'être atteint. - Certaines circonstances sont incompatibles avec le dévoilement des objectifs d'un SIA (*ex : contexte expérimental où des étudiants devaient interagir avec un assistant de cours sans savoir*	**<u>Risques</u>** : **Sécurité compromise** (*Ex : acteur ou État malveillant + contrôle sur des individus via des atteintes à leur vie privée.)* **Atteinte à la dignité humaine** **Atteinte à la cohésion sociale** – *la vie privée sous-tend la cohésion sociale* **Utilisation de données personnelles pour le dév. de logiciels (malveillants ou non) <u>sans</u> consentement / ou avec consentement <u>vicié</u>** *Ex :* → *Consentement en ligne de type « à prendre ou à laisser » / manque de choix* → *Utilisation de <u>seconde main</u> de données personnelles partagées en ligne à des fins auxquelles l'utilisateur n'a probablement pas consenti ou qui n'ont pas été portées à sa connaissance.* **<u>Enjeu :</u>** Seuil/type de consentement requis en fonction du contexte – *consentement présumé, éclairé, général (ex : consentement à*	**<u>Risque</u>** : **Prise de décision algorithmique injuste** en raison de biais socialement significatifs dans les données d'un SIA. *Décisions injustes si basées sur :* *1. Des « facteurs d'importance éthique » <u>non pertinents</u> à la prise de décision;* *2. Des facteurs <u>pertinents</u>, mais légalement protégés comme facteurs non discriminatoires.* 3. *Des facteurs qui ne sont pas d'une importance éthique évidente, mais qui entrainent quand même une prise de décision injustement biaisée.* **<u>Risque</u>** : **Reproduction de biais (cercle vicieux)** : données initialement biaisées => décisions et interventions du SIA biaisées => le SIA collecte et utilise ces données biaisées (apprentissage) => les données sont de plus en plus biaisées. ***<u>Risque exacerbé – IA</u>*** Le risque des préjugés historiques affectant la prise de décision => pas nouveau, mais dans le cas de l'IA, ces préjugés risquent d'être <u>intégrés, renforcés et perpétués</u> via des	**<u>Risques</u>** : **Perte du « capital sémantique »** des individus – *pouvoir de donner un sens à qqc et de le comprendre* – en raison de la capacité technique de l'IA d'automatiser le sens et la création de sens. *Ex : via l'attribution de sens par l'IA d'une manière qui diverge de nos choix (par tirage au sort, de façon arbitraire, etc.)* ou *via la définition de mots par leurs usages préexistants qui empêche la redéfinition de sens pour l'avenir.* **<u>Sémantisation impossible</u>** La sémantisation est subjective => impossible pour un SIA, dans un contexte social, de définir tous les sens possibles – *ex : IA peut <u>détecter</u> que quelqu'un semble triste mais ne peut pas changer/définir le sens de la tristesse*.

[28] Prémisse : 7 facteurs essentiels (et d'importance égale) à des initiatives **<u>d'AI4SG</u>** (*AI for Social Good*).
[29] Le principe de *Beneficience* → condition préalable à l'AI4SG et applicable aux 7 facteurs.

				qu'il s'agissait en réalité d'un robot.)	*l'usage des données pour « tout usage médical » sans fins spécifiques), dynamique (surveillance et ajustement des préférences de confidentialité par les utilisateurs)…*	mécanismes d'apprentissage par renforcement.	
Meilleures pratiques ***S'adressent aux concepteurs***	**1. SIA falsifiables** : Identifier les exigences critiques d'un SIA, s'assurer qu'elles sont falsifiables et les tester progressivement du « laboratoire au monde extérieur ». ***Falsifiabilité – démarche*** **1.1** Identifier les exigences critiques du SIA (ex : sécurité); **1.2** S'assurer que les exigences critiques sont falsifiables; **Exigences critiques non falsifiables = SIA non fiable.* **1.3** Tester empiriquement ces exigences selon un déploiement progressif pour améliorer progressivement les niveaux de fiabilité; **1.3.1** Entreprendre des tests de falsification (tester les hypothèses initiales) des exigences les plus critiques dans un contexte contrôlé; **Si ces hypothèses ne sont pas réfutées dans un petit nombre de contextes contrôlés :* **1.3.2** Effectuer des tests dans des contextes de moins en moins contrôlés et/ou tester un plus grand nombre d'exigences moins critiques;	**Garanties (*safeguards*) comme facteurs de *design* des projets d'AI4SG** **1. Manipulation des données** Adopter des garanties qui limitent la connaissance de la manière dont les *input data* affectent les résultats. **2. Indicateurs non-causaux** Adopter des garanties à l'effet que les indicateurs non causaux ne faussent pas les interventions du SIA de manière inappropriée. *(EX : limiter la sélection des indicateurs à utiliser, limiter la mesure dans laquelle ils façonnent les interventions, etc.)*	**5 dimensions d'une intervention correctement contextualisée dans la conception et le déploiement d'un SIA** : **1.** Consulter les utilisateurs qui interagissent avec le SIA et qui sont impactés par eux. *Tenir compte des :* **2.** Caractéristiques des utilisateurs; **3.** Méthodes de coordination entre l'utilisateur et le système; **4.** Buts et effets d'une intervention du SIA. *Et :* **5.** Respecter le droit d'option de l'utilisateur *(c.à.d. le droit d'ignorer une intervention tout en ayant la possibilité d'accepter une intervention ultérieure, ou encore le droit de modifier une intervention ou de demander une intervention plus appropriée.)*	**1. Explication adéquatement conceptualisée et contextualisée** **1.1** Choisir un niveau d'abstraction/cadre conceptuel (*LoA*) qui remplit le but explicatif souhaité et qui convient au système et au destinataire de l'explication. **1.2** Déployer des arguments rationnellement persuasifs pour le public cible de manière à ce que ce dernier soit en mesure de fournir l'explication à son tour – *explication compréhensible et intuitive pour le public cible.* **2. Évaluation du niveau de transparence requis** Au stade de la conception, évaluer le niveau de transparence requis en fonction de l'objectif général du SIA et du contexte de sa mise en œuvre. **2.1 Si transparence possible/requise** S'assurer que le but pour lequel un SIA est développé et déployé est connu par défaut des destinataires. **2.2 Choix entre opacité VS transparence dans certaines**	**1. Consentement préalable** Obtenir le consentement relativement à l'utilisation des données personnelles avant qu'elles ne soient utilisées. **2. Seuil** Respecter le seuil de consentement établi pour le traitement des ensembles de données personnelles.	**Données** **1. «** Nettoyer » les ensembles de données utilisés pour entrainer l'IA, mais conserver les données représentatives des nuances contextuelles qui aident la prise de décision éthique et qui favorisent l'inclusion. *(Ex : un logiciel de TX de texte, pour favoriser l'inclusion, devrait fonctionner différemment si un non-voyant l'utilise)* **2.** Éliminer les facteurs d'importance éthique (et leurs *proxys*) qui n'ont pas rapport avec le résultat, mais conserver ces mêmes facteurs lorsqu'ils sont nécessaires pour des raisons de sécurité, d'inclusivité ou pour d'autres considérations éthiques. **Considérer les groupes cibles (identification) et la manière dont ils interagissent** Pour l'utilisation de l'IA dans les interactions humains-ordis (*ex : chatbots),* tenir compte des groupes auxquels appartient un utilisateur et des caractéristiques qu'ils incarnent lorsqu'ils interagissent (3.2) avec le logiciel.	**Ne pas empêcher les gens de sémantiser** : - Faire la distinction entre les tâches qui devraient être déléguées à un SIA et celles qui ne devraient pas l'être. **-** Déployer l'IA pour faciliter une sémantisation "*human-friendly*", et non pour fournir la sémantisation elle-même.

	1.3.3 Tester le SIA dans le monde réel seulement s'il est sécuritaire de le faire; **Si les hypothèses sont réfutées ou si de nouveaux dilemmes du monde réel le justifient :* **1.3.4** Rectifier les hypothèses initiales. ***Falsiabilité – Détails*** ***Monde réel** : Nécessité de tester éventuellement le SIA dans le monde réel. *Les méthodes de falsification qui ne testent pas les SIA dans le monde réel comportent des limites. EX : limites des méthodes de vérifications formelles ou de simulations → ce qui semble être fiable avec ces méthodes peut se révéler invalide dans le monde réel.* ***Rectification des hypothèses initiales** : Nécessité de rectifier les hypothèses initiales, et non pas d'essayer de comprendre des résultats à travers des hypothèses initiales erronées (*retrodictive approach*) ou de laisser un SIA constamment adapter sa « compréhension » sur la base d'hypothèses initiales erronées (*on-the-fly approach* / ex : chatbot de Twitter et langage offensant). **2. Possibilité d'interruption ou de modification du déploiement du SIA** L'interruption ou la modification du déploiement du SIA doit être possible à tout moment si des effets indésirables ou dangereux surviennent.			**situations *(ex : contexte expérimental)* :** Se rapporter à des notions préexistantes de consentement éclairé pour des expériences sur sujets humains (*ex : Code de Nuremberg, Déclaration d'Helsinki et Rapport Belmont.*)			

Entraves	**1. Tests impossibles** Impossibilité de tester le SIA dans *tous* les contextes, certitude hors de portée. **2. Fiabilité intrinsèque** Le processus de falsifiabilité permet de savoir *quand* le SIA n'est pas fiable (c.à.d. lorsqu'une exigence critique donnée n'est pas rencontrée), mais pas de de savoir si le SIA est fiable en soi. **3. Monde réel** Impossibilité de tester dans le monde réel.	-	-	**Explicabilité** Situations où l'explication est impossible ou indésirable. *(Ex : utilisateur ayant une déficience cognitive => efficacité limitée de l'explication contextualisée)*	**« Situations de vie ou de mort » (urgences nationales, pandémies, etc.)** (*Ex : Ebola : analyse des enregistrements de données d'appels d'utilisateurs de cellulaires de la région pour permettre aux épidémiologistes de suivre la propagation de la maladie.)*	**Associations injustes** Les ensembles de données de formation *(training datasets)* peuvent contenir un langage naturel porteur d'associations injustes *(ex : associations injustes entre le genre et les mots*).	-
Stratégies	**Zones dérèglementées** Création de zones dérèglementées *(Ex : Allemagne et voitures autonomes)*	-	-	**Transparence des objectifs** La transparence des objectifs du SIA prend davantage d'importance dans les cas où l'explication des résultats est impossible ou indésirable.	-	-	-

***

BIBLIOGRAPHIE

SECTION A

Modèles :

Modèles gouvernementaux :

Artificial Intelligence Ethics Framework for the Intelligence Community
Office of the Director of National Intelligence. "Artificial Intelligence Ethics Framework for the Intelligence Community" (2020), En ligne : <https://www.dni.gov/files/ODNI/documents/AI_Ethics_Framework_for_the_Intelligence_Community_10.pdf>

Data Ethics Framework
Gouvernement du Royaume-Uni. "Data Ethics Framework" (2020), En ligne : <https://www.gov.uk/government/publications/data-ethics-framework>

The Assessment List for Trustworthy Artificial Intelligence (ALTAI)
Groupe d'experts indépendants de haut niveau sur l'IA. "ALTAI - The Assessment List on Trustworthy Artificial Intelligence" (2020), En ligne : <https://futurium.ec.europa.eu/en/european-ai-alliance/pages/altai-assessment-list-trustworthy-artificial-intelligence>

Modèles d'entreprises

A Framework for Systematically Applying Humanistic Ethics when Using AI as a Design Material
Dent, Kyle, Richelle Dumond et Mike Kuniavsky. *A Framework for Systematically Applying Humanistic Ethics when Using AI as a Design Material*, SSRN, (Rochester, NY: Social Science Research Network, 2019), En ligne : <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3455518>

Everyday Ethics for Artificial Intelligence
IBM. "Everyday Ethics for Artificial Intelligence" (2019), En ligne : <https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf>

Guidelines for Human-AI Interaction
Amershi, Saleema et al. "Guidelines for Human-AI Interaction" (2019), En ligne : <https://www.microsoft.com/en-us/research/publication/guidelines-for-human-ai-interaction/>

Responsible AI by Design in Practice
Benjamins, Richard, Alberto Barbado et Daniel Sierra. "Responsible AI by Design in Practice" (2019), En ligne : <http://arxiv.org/abs/1909.12838>

Responsible bots: 10 guidelines for developers of conversational AI

Microsoft. "Responsible bots: 10 guidelines for developers of conversational AI" (2018), En ligne : <https://www.microsoft.com/en-us/research/publication/responsible-bots/>

The Box: Dynamics of AI principles (*AI Ethics Lab,* 2020)
AI Ethics Lab. "Tool: The Box", (2020), En ligne : *Toolbox: Dynamics of AI Principles* <http://aiethicslab.com/the-box/>

Modèles d'associations et d'OSBL

AI Ethics Framework
Digital Catapult. "AI Ethics Framework", (2018), En ligne : *Machine Intelligence Garage* <https://www.migarage.ai/ethics/ethics-framework/>

AI Ethics Label[30]
AIEI Group. "AI Ethics Impact Group: From Principles to Practice – VDE", (2020), En ligne : <https://www.ai-ethics-impact.org/en>

Data Ethics Canvas[31]
Open Data Institute. "The Data Ethics Canvas" (2019), En ligne : <https://theodi.org/article/data-ethics-canvas/>

Éthique et numérique : un référentiel pratique pour les acteurs du numérique
Cigref et Syntec Numérique. « Éthique & numérique : un référentiel pratique pour les acteurs du numérique » (2018), En ligne : *Cigref* <https://www.cigref.fr/ethique-numerique-un-referentiel-pratique-pour-les-acteurs-du-numerique>

Responsible AI Design Assistant[32]
AI Global. "Responsible AI Design Assistant" (2020), En ligne : <https://oproma.github.io/rai-trustindex/>

Modèles académiques

[30] Note : les auteurs proposent aussi une **matrice de risques** qui complète le modèle et qui n'a pas été détaillée dans le cadre du présent rapport. Voir ***Annexe J*** et consulter plus spécifiquement les pages 35-40 de la publication *From Principles to Practice*.

[31] **Une version mise à jour du modèle a été publiée en 2021**. Voir : https://theodi.org/article/the-data-ethics-canvas-2021/#1563365825519-a247d445-ab2d.

[32] **AI Global a changé de nom en 2021 pour *Responsible Artificial Intelligence Institute* (RAII)**.
Des modifications ont été apportées à l'outil en 2021. Désormais, **6 dimensions éthiques** d'une IA fiable sont examinées : *Organization Maturity, Accountability, Data, Fairness, Interpretability, Robustness*.
Le lien vers la nouvelle version de l'outil est le suivant : https://designassistant.responsible.ai/.

An Ethical Framework for Good AI Society: Opportunities, Risks, Principles, and Recommendations
Floridi, Luciano et al. “AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations” (2018) 28:4 Minds & Machines 689–707

Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications
Ryan, Mark et Bernd Carsten Stahl. "Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications" (2020), Journal of Information, Communication and Ethics in Society, En ligne : <https://doi.org/10.1108/JICES-12-2019-0138>

DEDA : Data Ethics Decision Aid
Utrecht Data School et Université d’Utrecht. “Data Ethics Decision Aid (DEDA)” (2020), En ligne : *Utrecht Data School* <https://dataschool.nl/en/deda/>

FactSheets: Increasing Trust in AI Services through Supplier’s Declarations of Conformity
Arnold, Matthew et al. "FactSheets: Increasing Trust in AI Services through Supplier’s Declarations of Conformity" (2019), En ligne : <http://arxiv.org/abs/1808.07261>

How to Design AI for Social Good: Seven Essential Factors
Floridi, Luciano et al. "How to Design AI for Social Good: Seven Essential Factors" (2020) 26:3 Sci Eng Ethics 1771–1796, En ligne : <https://link.springer.com/article/10.1007/s11948-020-00213-5>

Autres sources

Fjeld, Jessica et al. "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI", Berkman Klein Center for Internet & Society (2020), En ligne : <https://dash.harvard.edu/handle/1/42160420>

Jobin, Anna, Marcello Ienca et Effy Vayena. "The global landscape of AI ethics guidelines" (2019) 1:9 Nat Mach Intell 389-399

Hagendorff, Thilo. "The Ethics of AI Ethics: An Evaluation of Guidelines" (2020) 30:1 Minds Mach 99-120

Hagendorff, Thilo. "The Ethics of AI Ethics: An Evaluation of Guidelines" (2020) 30:3 Minds Mach 457-461

SECTION B

AI Ethics Label
AIEI Group. "AI Ethics Impact Group: From Principles to Practice – VDE" (2020), En ligne : <https://www.ai-ethics-impact.org/en>

An Ethical Framework for Good AI Society: Opportunities, Risks, Principles, and Recommendations
Floridi, Luciano et al. “AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations” (2018) 28:4 Minds Mach 689–707, En ligne : <https://link.springer.com/article/10.1007/s11023-018-9482-5>

Australia’s Artificial Intelligence Ethics Framework (2020)
Gouvernement de l’Australie. "Australia’s Artificial Intelligence Ethics Framework" (2020), En ligne : <https://www.industry.gov.au/data-and-publications/australias-artificial-intelligence-ethics-framework>

Australia’s Ethics Framework: A Discussion Paper (2019)
Gouvernement de l’Australie, "Australia’s Ethics Framework: A Discussion Paper" (2019), En ligne : <https://consult.industry.gov.au/australias-ai-ethics-framework>

Ethical machines: The Human-centric use of artificial intelligence
Lepri et al. "Ethical machines: The Human-centric use of artificial intelligence" (2021), En ligne: <https://www.sciencedirect.com/science/article/pii/S2589004221002170>

How to Design AI for Social Good: Seven Essential Factors
Floridi, Luciano et al. "How to Design AI for Social Good: Seven Essential Factors" (2020) 26:3 Sci Eng Ethics 1771–1796, En ligne : <https://link.springer.com/article/10.1007/s11948-020-00213-5>

La Déclaration de Montréal (2018)
Université de Montréal. « La Déclaration de Montréal pour un développement responsable de l’intelligence artificielle » (2018), En ligne : <https://www.declarationmontreal-iaresponsable.com/la-declaration>

Lignes directrices en matière d’éthique pour une IA digne de confiance (2019)
Groupe d’experts indépendants de haut niveau sur l’IA. « Lignes directrices en matière d’éthique pour une IA digne de confiance » (2019), En ligne : *Shaping Europe’s digital future - Commission européenne* <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>

Recommandation du Conseil sur l’intelligence artificielle (2019)
Organisation de coopération et de développement économiques. « Recommandation du Conseil sur l’intelligence artificielle » (2019), En ligne : <https://legalinstruments.oecd.org/fr/instruments/OECD-LEGAL-0449>

Responsible AI: A Global Policy Framework (2019)[33]

[33] ITechLaw (*International Technology Law Association*) a publié une version mise à jour de cette publication en 2021. La mise à jour comporte aussi un outil d’évaluation de l’impact de l’IA selon des facteurs de risque.

International Technology Law Association. "Responsible AI: A Global Policy Framework" (2019), En ligne : *ITechLaw* <https://www.itechlaw.org/ResponsibleAI>

The Assessment List for Trustworthy Artificial Intelligence (ALTAI)
Groupe d'experts indépendants de haut niveau sur l'IA. "ALTAI - The Assessment List on Trustworthy Artificial Intelligence" (2020), En ligne : <https://futurium.ec.europa.eu/en/european-ai-alliance/pages/altai-assessment-list-trustworthy-artificial-intelligence>

Trustworthy use of artificial intelligence: priorities from a philosophical, ethical, legal, and technological viewpoint as a basis for certification of AI
Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS. "Trustworthy use of artificial intelligence: priorities from a philosophical, ethical, legal, and technological viewpoint as a basis for certification of AI" (2020), En ligne : <https://www.iais.fraunhofer.de/en/press/press-release-190702.html>

Voir : https://www.itechlaw.org/ResponsibleAI2021.

DOCUMENTS DE RÉFÉRENCE

Chartes et principes éthiques

La Déclaration de Montréal (2018)

Université de Montréal. « La Déclaration de Montréal pour un développement responsable de l'intelligence artificielle » (2018), En ligne : <https://www.declarationmontreal-iaresponsable.com/la-declaration>

- Cadre éthique composé de 10 principes guidant l'évolution de l'IA pour des résultats moralement et socialement acceptables.
- Principes : bien-être, autonomie, protection de l'intimité et de la vie privée, solidarité, participation démocratique, équité, diversité, prudence, responsabilité et développement soutenable.

Lignes directrices en matière d'éthique pour une IA digne de confiance (2019)

Groupe d'experts indépendants de haut niveau sur l'IA. « Lignes directrices en matière d'éthique pour une IA digne de confiance » (2019), En ligne : *Shaping Europe's digital future - Commission européenne* <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>.

- Proposition de 7 principes éthiques pour une IA digne de confiance.
- Principes : action humaine et contrôle humain (1), robustesse technique et sécurité (2), respect de la vie privée et gouvernance des données (3), transparence (4), diversité, non-discrimination et équité (5), bien-être sociétal et environnemental (6) et responsabilité (7).

Ethically Aligned Design (2019)

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, 2e édition. IEEE. (2019), En ligne: <https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf>

- Proposition de 8 principes éthiques et de recommandations pratiques propres à chacun d'eux.
- Principes : Human rights, Well-being, Data agency, Effectiveness, Transparency, Accountability, Awareness of Misuse et Competence.

Responsible AI: A Global Policy Framework (2019)[34]

International Technology Law Association. "Responsible AI: A Global Policy Framework" (2019), En ligne : *ITechLaw* <https://www.itechlaw.org/ResponsibleAI>.

- Proposition de 8 principes éthiques pour le développement et l'utilisation responsable de l'IA.
- Principes : Ethical purpose and social benefit (1), Accountability (2), Transparency and Explainability (3), Fairness and non-discrimination (4), Safety and Reliability (5), Open Data and Fair Competition (6), Privacy (7) et AI and Intellectual Property (8).

Normes juridiques

[34] ITechLaw (*International Technology Law Association*) a publié une version mise à jour de cette publication en 2021. La mise à jour comporte aussi un outil d'évaluation de l'impact de l'IA selon des facteurs de risque. Voir : https://www.itechlaw.org/ResponsibleAI2021.

Charte canadienne du numérique (2019)

Gouvernement du Canada. « Charte canadienne du numérique : La confiance dans un monde numérique - Innover pour un meilleur Canada » (2019), En ligne : <https://www.ic.gc.ca/eic/site/062.nsf/fra/h_00108.html>.

- 10 principes guidant la transformation numérique responsable par le gouvernement du Canada.

Charte éthique européenne d'utilisation de l'intelligence artificielle dans les systèmes judiciaires (2018)

Commission européenne pour l'efficacité de la justice. « Charte éthique européenne d'utilisation de l'intelligence artificielle dans les systèmes judiciaires » (2018), En ligne : <https://www.coe.int/fr/web/cepej/cepej-european-ethical-charter-on-the-use-of-artificial-intelligence-ai-in-judicial-systems-and-their-environment>

- 5 principes éthiques pour l'utilisation de l'IA dans les systèmes judiciaires.
- Principes : respect des droits fondamentaux (1), non-discrimination (2), qualité et sécurité (3), transparence, neutralité et intégrité intellectuelle (4) et maîtrise par l'utilisateur (5).
- Exemples de technologies visées par la charte : moteurs de recherche de jurisprudence, "chatbots" facilitant l'accès aux sources d'information pertinentes pour les justiciables, catégorisation de contrats, etc.

Sites et publications recensant les outils et lignes directrices pour le développement éthique de l'IA[35]			
Liens URL	**Mise à jour**	**Auteur**	**Notes**
AI Ethicist	2021	Merve Hickok	**Sections** : AI Principles AI Governance National Strategies AI Organizations Ethics Cases Research on AI Biais, Fairness
AI Ethics Guidelines Global Inventory	2021-04	AlgorithmWatch (Allemagne, Berlin)	Cartographie des cadres de l'IA opérationnalisant des principes éthiques.
AI Principles Map	2021	AI Ethics Lab (É-U, Cambridge)	Cartographie des cadres de l'IA opérationnalisant des principes éthiques.
AI Ethics Typology (2019)	- *(Article)*	Morley, Floridi et al. (R-U, Oxford)	Typologie discutée dans From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principle into Practices (2019)
AI Global Bubble	2018	N/D	Cartographie des cadres de l'IA opérationnalisant des principes éthiques. Version interactive Données
Dote Everyone	2021	DoteEveryone *Site désormais géré par le Ada Lovelace Institute (R-U, Londres)	"This directory focuses on initiatives which help make more tech responsible/ethical, rather than individually ethical products or services. So: standards, training courses, advocacy, guidelines, campaigns, regulations, networks, tools, etc."
Ethical frameworks, tool kits, principles, and oaths	2020-10	Kathy Baxter (É-U, San Francisco)	Frameworks, Tools & Toolkits, Checklists, Principles, Oaths, Manifestoes, and Codes of Conducts, Policy Papers, White Papers, Statements, Reports, Other Resources

[35] Mise à jour du tableau : 2021-06-15.

Ethics in AI research papers and articles	2021-05		Peer-reviewed, Governement, NGO and expert publications, Popular Press
Ethics Standards repository	2021	Open Community for Ethics in Autonomous and Intelligent Systems (OCEANIS) (R-U)	"The Global AI Standards Repository is the world's first centralized, transparent notification system that captures AI and Autonomous and Intelligent Systems standards and standards in progress."
Linking-ai-principles	2020	Yi Zeng – Institute of Automation, Chinese Academy of Sciences (Chine, Beijing)	"Linking Artificial Intelligence Principles (LAIP) is an initiative and platform for integrating, synthesizing, analyzing, and promoting global Artificial Intelligence Principles and their social and technical practices World Wide, from different research institutes, non-profit organizations, non-governmental organizations, companies, etc. The efforts aim at understanding in which degree do these different AI Principles proposals share common values, differ and complete each other."
An Updated Round Up of Ethical Principles of Robotics and AI	2019	Alan Winfield	"This blogpost is an updated round up of the various sets of ethical principles of robotics and AI that have been proposed to date, ordered by date of first publication."
Ethics in Context	2021-06	Centre for Ethics (Canada, Université de Toronto)	The Oxford Handbook of Ethics of AI: Online Supplement The Oxford Handbook of Ethics of AI: An Annotated Bibliography
Stanford HAI: AI Index Report 2021	N/A *(Rapport)*	Human-Centered Artificial Intelligence (É-U, Université de Stanford)	Rapport Données
Implementing AI Principles: Frameworks, Processes, and Tools (2021)	N/A *(Article)*	Boza et Evgeniou (Hongrie, Budapest)	Examen des principes, cadres et outils éthiques actuels et identification des lacunes potentielles.

ANNEXES

ANNEXES DE LA SECTION A[36]

ANNEXE A – ODNI, 2020

Extrait de: ODNI. Artificial Intelligence Ethics Framework for the Intelligence Community (2020).

Les 10 objectifs du guide et les questions correspondantes	
Objectifs	**Questions**
1. Purpose: Understanding Goals and Risks	• What is the goal you are trying to achieve by creating this AI, including components used in AI development? Is there a need to use AI to achieve this goal? Can you use other non-AI related methods to achieve this goal with lower risk? Is AI likely to be effective in achieving this goal? • Are there specific AI system methods suitable and preferred for this use case? Does the efficiency and reliability of the AI in this particular use case justify its use for this purpose? • What benefits and risks, including risks to civil liberties and privacy, might exist when this AI is in use? Who will benefit? Who or what will be at risk? What is the scale of each and likelihood of the risks? How can those risks be minimized and the remaining risks adequately mitigated? Do the likely negative impacts outweigh likely positive impacts? • What performance metrics best suit the AI, such as accuracy, precision, and recall, based on risks determined by mission managers, analysts, and consumers given the potential risks; and how will the accuracy of the information be provided to each of those stakeholders? What impacts could false positive and false negative rates have on system performance, mission goals, and affected targets of the analysis? • Have you engaged with the AI system developers, users, consumers, and other key stakeholders to ensure a common understanding of the goal of the AI and related risks of utilizing AI to achieve this goal? • How are you documenting the goals and risks?
2. Legal Obligations and Policy Considerations Governing the AI and the Data	• What authorities, agreements, or contracts govern the collection or acquisition of all sources of the data related to the model (training, testing, and operational data)? Who can clarify limitations from the agreements or contracts? • What legal or policy restrictions exist on the use of data under this authority/agreement/ contract? (For example, data subject to the Privacy Act should be used for a purpose that is compatible with that for which the data was collected). • How must data be stored, shared, retrieved, accessed, used, retained, disseminated, and dispositioned under the authority/agreement/contract, as well as relevant constitutional, statutory, and regulatory provisions? • What authorities or agreements apply to the AI itself, including the use, modification, storage, retrieval, access, retention, and disposition of the AI? Are there any proposed downstream applications of the AI that are legally restricted from using the underlying data? • Does combining data with other inputs from the AI create new legal, records management, or classification risks relating to how the information is maintained and protected?

[36] Le contenu de ces annexes est directement extrait des publications originales.

3. Human Judgment and Accountability	• Given the purpose of the AI and potential consequences of its use, at what points, if any, are a human required as part of the decision process? If the AI could result in significant consequences such as an action with the potential to deprive individuals of constitutional rights or the potential to interfere with their free exercise of civil liberties, how will you ensure individual human involvement and accountability in decisions that are assisted through the use of AI? • Where and when should the human be engaged? Before the results are used in analysis? Before the outputs are provided for follow-on uses? • Who should be the accountable human(s)? Do they know that they are designated as the accountable human(s)? What qualifications are required to serve in that role? How is accountability transferred to another human? • What are the access controls and training requirements for those operating at different stages in the AI lifecycle? • What does the accountable human need to know about the AI to judge its reliability and accuracy? • How may introducing an accountable human produce cognitive biases and/or confirmation bias? • Who should be engaged for unresolved issues and disputes regarding the AI or its outputs?
4. Mitigating Undesired Bias and Ensuring Objectivity	• How complete are the data on which the AI will rely? Are they representative of the intended domain? How relevant is the training and evaluation data to the operational data and context? How does the AI avoid perpetuating historical biases and discrimination? • What are the correct metrics to assess the AI's output? Is the margin of error one that would be deemed tolerable by those who use the AI? What is the impact of using inaccurate outputs and how well are these errors communicated to the users? • What are the potential tradeoffs between reducing undesired bias and accuracy? To what extent can potential undesired bias be mitigated while maintaining sufficient accuracy? • Do you know or can you learn what types of bias exist in the training data (statistical, contextual, historical, or other)? How can undesired bias be mitigated? What would happen if it is not mitigated? Is the selected testing data appropriately representative of the training data? Based on the purpose of the AI, how much and what kind of bias, if any, are you willing to accept in the data, model, and output? Is the team diverse enough in disciplinary, professional, and other perspectives to minimize any human bias? • How will undesired bias and potential impacts of the bias, if any, be communicated to anyone who interacts with the AI and output data?
5. Testing Your AI	• Based on the purpose of the AI and potential risks, what level of objective performance for the desired performance metric, (e.g., precision, recall, accuracy, etc.) do you require? • Has the AI been evaluated for potential biased outcomes or if outcomes cause an inappropriate feedback loop? Have you considered applicable methods to make the AI more robust to adversarial attacks? • How and where will you document the test methodology, results, and changes made based on the test results? • If a third party created the AI, what additional risks may be associated with that third party's assumptions, motives, and methodologies? What limitations might arise from that third party claiming its methodology is proprietary? What information should you require as part of the acquisition of the analytic? What is the minimum amount of information you must have to approve an AI for use? • Was the AI tested for potential security threats in any/all levels of its stack (e.g. software level, AI framework level, model level, etc.)? Were resulting risks mitigated?
6. Accounting for Builds,	• As you refine the AI, how does the data you have used, the parameters and weights you have chosen, and the outputs ensure that this version or evolution is designed to achieve the authorized purpose?

Versions, and Evolutions of an AI	• Have you accounted for natural data drift within the operational environment compared to training data? • Have you documented provenance of data, outputs of the iteration, and test results (accuracy) in a way that will provide for repeatability, auditing, and oversight? If the AI is continuously modified, are all critical aspects, dependencies, and artifacts version controlled and documented? Will it be clear to anyone auditing the AI or consumers of the AI's outputs which version was in use at any given moment in time? Will it be clear which iteration of a model drew on which data and produced what outputs? • Where will you save documentation on versions of AI and relevant training and test data? Have you made that information available to users and consumers of the AI? How will this documentation be retained and made discoverable to ensure compliance with your Agency's records management responsibilities? • Have you accounted for changes in demographics of your customer for your AI capability (e.g., changing user experience needs) or the changing needs of the mission?
7. Documentation of Purpose, Parameters, Limitations, and Design Outcomes	• How can you store the documentation in a way that is available to all potential consumers of this AI? • Have you documented where the data came from and its downstream uses and sharability? The downstream uses and sharability of the AI? • Have you documented what rules apply to the data as a whole? What rules apply to subsets? • Have you documented the potential risks of using the AI and its output data, and the steps taken to minimize these risks? • Have you documented use cases for which the AI was and was not specifically designed? • Have you documented the process for discovering undesired bias and the conclusions? • Have you documented how to verify and validate the model as well as the frequency with which these checks should be performed?
8. Transparency: Explainability and Interpretability	• Given the purpose of the AI, what level of explainability or interpretability is required for how the AI made its determination? If a third party created the AI, how will you ensure a level of explainability or interpretability? Does this conform with Intelligence Community Directive 203: Analytic Standards? • How are outputs marked to clearly show that they came from an AI? • How might you respond to an intelligence consumer asking "How do you know this?" How will you describe the dataset(s) and tools used to make the output? How was the accuracy or appropriate performance metrics assessed? How were the results independently verified? Have you documented and explained that machine errors may differ from human errors?
9. Periodic Review	• How will user and peer engagement be integrated into the model development process and periodic performance review once deployed? • Given the purpose of this AI, what is an appropriate interval for checking whether it is still accurate, unbiased, explainable, etc.? What are the checks for this model?

	• As time passes and conditions change, is the training data still representative of the operational environment? • How will the appropriate performance metrics, such as accuracy, of the AI be monitored after the AI is deployed? How much distributional shift or model drift from baseline performance is acceptable? • Who is responsible for checking the AI at these intervals? • How will the accountable human(s) address changes in accuracy and precision due to either an adversary's attempts to disrupt the AI or unrelated changes in the operational/business environment, which may impact the accuracy of the AI?
10. Stewardship and Accountability: Training Data, Algorithms, Models, Outputs of the Models, Documentation	• Who will be responsible for maintaining, re-verifying, monitoring, and updating this AI once deployed? • Who is ultimately responsible for the decisions of the AI and is this person aware of the intended uses and limitations of the analytic? • Who is accountable for the ethical considerations during all stages of the AI lifecycle? • If anyone believed that the AI no longer meets this ethical framework, who will be responsible for receiving the concern and as appropriate investigating and remediating the issue? Do they have authority to modify, limit, or stop the use of the AI?

ANNEXE B – AI HLEG, 2020

Extrait de : Groupe d'experts indépendants de haut niveau sur l'IA. "ALTAI - The Assessment List on Trustworthy Artificial Intelligence" (2020).

Les 7 principes éthiques et les questions correspondantes	
Principes	**Questions**
#1 Human Agency and Oversight	**Human Agency and Autonomy** Is the AI system designed to interact, guide or take decisions by human end-users that affect humans or society? - Could the AI system generate confusion for some or all end-users or subjects on whether a decision, content, advice or outcome is the result of an algorithmic decision? - Are end-users or other subjects adequately made aware that a decision, content, advice or outcome is the result of an algorithmic decision? Could the AI system generate confusion for some or all end-users or subjects on whether they are interacting with a human or AI system? - Are end-users or subjects informed that they are interacting with an AI system? Could the AI system affect human autonomy by generating over-reliance by end-users? - Did you put in place procedures to avoid that end-users over-rely on the AI system? Could the AI system affect human autonomy by interfering with the end-user's decision-making process in any other unintended and undesirable way? - Did you put in place any procedure to avoid that the AI system inadvertently affects human autonomy? Does the AI system simulate social interaction with or between end-users or subjects? Does the AI system risk creating human attachment, stimulating addictive behaviour, or manipulating user behaviour? Depending on which risks are possible or likely, please answer the questions below: - Did you take measures to deal with possible negative consequences for end-users or subjects in case they develop a disproportionate attachment to the AI System? - Did you take measures to minimise the risk of addiction? o Did you take measures to mitigate the risk of manipulation?
	Human Oversight Please determine whether the AI system (choose as many as appropriate): - Is a self-learning or autonomous system; - Is overseen by a Human-in-the-Loop; - Is overseen by a Human-on-the-Loop - Is overseen by a Human-in-Command Have the humans (human-in-the-loop, human-on-the-loop, human-in-command) been given specific training on how to exercise oversight? Did you establish any detection and response mechanisms for undesirable adverse effects of the AI system for the end-user or subject? Did you ensure a 'stop button' or procedure to safely abort an operation when needed? Did you take any specific oversight and control measures to reflect the self-learning or autonomous nature of the AI system?
	Resilience to Attack and Security

#2 Technical Robustness and Safety	Could the AI system have adversarial, critical or damaging effects (e.g. to human or societal safety) in case of risks or threats such as design or technical faults, defects, outages, attacks, misuse, inappropriate or malicious use? Is the AI system certified for cybersecurity (e.g. the certification scheme created by the Cybersecurity Act in Europe) or is it compliant with specific security standards? How exposed is the AI system to cyber-attacks? - Did you assess potential forms of attacks to which the AI system could be vulnerable? - Did you consider different types of vulnerabilities and potential entry points for attacks such as: — Data poisoning (i.e. manipulation of training data); — Model evasion (i.e. classifying the data according to the attacker's will); — Model inversion (i.e. infer the model parameters) Did you put measures in place to ensure the integrity, robustness and overall security of the AI system against potential attacks over its lifecycle? Did you red-team/pentest the system? Did you inform end-users of the duration of security coverage and updates? - What length is the expected timeframe within which you provide security updates for the AI system?
	General Safety Did you define risks, risk metrics and risk levels of the AI system in each specific use case? - Did you put in place a process to continuously measure and assess risks? - Did you inform end-users and subjects of existing or potential risks? Did you identify the possible threats to the AI system (design faults, technical faults, environmental threats) and the possible consequences? - Did you assess the risk of possible malicious use, misuse or inappropriate use of the AI system? - Did you define safety criticality levels (e.g. related to human integrity) of the possible consequences of faults or misuse of the AI system? Did you assess the dependency of a critical AI system's decisions on its stable and reliable behaviour? - Did you align the reliability/testing requirements to the appropriate levels of stability and reliability? Did you plan fault tolerance via, e.g. a duplicated system or another parallel system (AI-based or 'conventional')? Did you develop a mechanism to evaluate when the AI system has been changed to merit a new review of its technical robustness and safety?
	Accuracy Could a low level of accuracy of the AI system result in critical, adversarial or damaging consequences?

	Did you put in place measures to ensure that the data (including training data) used to develop the AI system is up-to-date, of high quality, complete and representative of the environment the system will be deployed in? Did you put in place a series of steps to monitor, and document the AI system's accuracy? Did you consider whether the AI system's operation can invalidate the data or assumptions it was trained on, and how this might lead to adversarial effects? Did you put processes in place to ensure that the level of accuracy of the AI system to be expected by end-users and/or subjects is properly communicated?
	Reliability, Fall-back plans and Reproducibility Could the AI system cause critical, adversarial, or damaging consequences (e.g. pertaining to human safety) in case of low reliability and/or reproducibility? - Did you put in place a well-defined process to monitor if the AI system is meeting the intended goals? - Did you test whether specific contexts or conditions need to be taken into account to ensure reproducibility? Did you put in place verification and validation methods and documentation (e.g. logging) to evaluate and ensure different aspects of the AI system's reliability and reproducibility? - Did you clearly document and operationalise processes for the testing and verification of the reliability and reproducibility of the AI system? Did you define tested failsafe fallback plans to address AI system errors of whatever origin and put governance procedures in place to trigger them? Did you put in place a proper procedure for handling the cases where the AI system yields results with a low confidence score? Is your AI system using (online) continual learning? - Did you consider potential negative consequences from the AI system learning novel or unusual methods to score well on its objective function?
#3 Privacy and Data Governance	**Privacy** Did you consider the impact of the AI system on the right to privacy, the right to physical, mental and/or moral integrity and the right to data protection?

	Depending on the use case, did you establish mechanisms that allow flagging issues related to privacy concerning the AI system?
	Data Governance Is your AI system being trained, or was it developed, by using or processing personal data (including special categories of personal data)? Did you put in place any of the following measures some of which are mandatory under the General Data Protection Regulation (GDPR), or a non-European equivalent? — Data Protection Impact Assessment (DPIA) — Designate a Data Protection Officer (DPO) and include them at an early state in the development, procurement or use phase of the AI system; — Oversight mechanisms for data processing (including limiting access to qualified personnel, mechanisms for logging data access and making modifications); — Measures to achieve privacy-by-design and default (e.g. encryption, pseudonymisation, aggregation, anonymisation); — Data minimisation, in particular personal data (including special categories of data); - Did you implement the right to withdraw consent, the right to object and the right to be forgotten into the development of the AI system? - Did you consider the privacy and data protection implications of data collected, generated or processed over the course of the AI system's life cycle? Did you consider the privacy and data protection implications of the AI system's non-personal training-data or other processed non-personal data? Did you align the AI system with relevant standards (e.g. ISO25, IEEE26) or widely adopted protocols for (daily) data management and governance?
#4 Transparency	**Traceability** Did you put in place measures that address the traceability of the AI system during its entire lifecycle? - Did you put in place measures to continuously assess the quality of the input data to the AI system? - Can you trace back which data was used by the AI system to make a certain decision(s) or recommendation(s)? - Can you trace back which AI model or rules led to the decision(s) or recommendation(s) of the AI system? - Did you put in place measures to continuously assess the quality of the output(s) of the AI system? - Did you put adequate logging practices in place to record the decision(s) or recommendation(s) of the AI system?
	Explainability Did you explain the decision(s) of the AI system to the users? Do you continuously survey the users if they understand the decision(s) of the AI system?
	Communication In cases of interactive AI systems (e.g., chatbots, robo-lawyers), do you communicate to users that they are interacting with an AI system instead of a human? Did you establish mechanisms to inform users about the purpose, criteria and limitations of the decision(s) generated by the AI system?

	- Did you communicate the benefits of the AI system to users? - Did you communicate the technical limitations and potential risks of the AI system to users, such as its level of accuracy and/ or error rates? - Did you provide appropriate training material and disclaimers to users on how to adequately use the AI system?
#5 Diversity, Non-discrimination and Fairness	**Avoidance of Unfair Bias** Did you establish a strategy or a set of procedures to avoid creating or reinforcing unfair bias in the AI system, both regarding the use of input data as well as for the algorithm design? Did you consider diversity and representativeness of end-users and/or subjects in the data? - Did you test for specific target groups or problematic use cases? - Did you research and use publicly available technical tools, that are state-ofthe-art, to improve your understanding of the data, model and performance? - Did you assess and put in place processes to test and monitor for potential biases during the entire lifecycle of the AI system (e.g. biases due to possible limitations stemming from the composition of the used data sets (lack of diversity, non-representativeness)? - Where relevant, did you consider diversity and representativeness of end-users and or subjects in the data? Did you put in place educational and awareness initiatives to help AI designers and AI developers be more aware of the possible bias they can inject in designing and developing the AI system? Did you ensure a mechanism that allows for the flagging of issues related to bias, discrimination or poor performance of the AI system? - Did you establish clear steps and ways of communicating on how and to whom such issues can be raised? - Did you identify the subjects that could potentially be (in)directly affected by the AI system, in addition to the (end-)users and/or subjects? Is your definition of fairness commonly used and implemented in any phase of the process of setting up the AI system? - Did you consider other definitions of fairness before choosing this one? - Did you consult with the impacted communities about the correct definition of fairness, i.e. representatives of elderly persons or persons with disabilities? - Did you ensure a quantitative analysis or metrics to measure and test the applied definition of fairness? - Did you establish mechanisms to ensure fairness in your AI system?
	Accessibility and Universal Design Did you ensure that the AI system corresponds to the variety of preferences and abilities in society? Did you assess whether the AI system's user interface is usable by those with special needs or disabilities or those at risk of exclusion? - Did you ensure that information about, and the AI system's user interface of, the AI system is accessible and usable also to users of assistive technologies (such as screen readers)? - Did you involve or consult with end-users or subjects in need for assistive technology during the planning and development phase of the AI system?

	Did you ensure that Universal Design principles are taken into account during every step of the planning and development process, if applicable? Did you take the impact of the AI system on the potential end-users and/or subjects into account? - Did you assess whether the team involved in building the AI system engaged with the possible target end-users and/or subjects? - Did you assess whether there could be groups who might be disproportionately affected by the outcomes of the AI system? - Did you assess the risk of the possible unfairness of the system onto the end-user's or subject's communities?
	Stakeholder Participation Did you consider a mechanism to include the participation of the widest range of possible stakeholders in the AI system's design and development?
#6 Societal and Environmental Well-being	**Environmental Well-being** Are there potential negative impacts of the AI system on the environment? - Which potential impact(s) do you identify? Where possible, did you establish mechanisms to evaluate the environmental impact of the AI system's development, deployment and/or use (for example, the amount of energy used and carbon emissions)? - Did you define measures to reduce the environmental impact of the AI system throughout its lifecycle?
	Impact on Work and Skills Does the AI system impact human work and work arrangements? Did you pave the way for the introduction of the AI system in your organisation by informing and consulting with impacted workers and their representatives (trade unions, (European) work councils) in advance? Did you adopt measures to ensure that the impacts of the AI system on human work are well understood? - Did you ensure that workers understand how the AI system operates, which capabilities it has and which it does not have? Could the AI system create the risk of de-skilling of the workforce? - Did you take measures to counteract de-skilling risks? Does the system promote or require new (digital) skills? - Did you provide training opportunities and materials for re- and up-skilling?
	Impact on Society at large or Democracy Could the AI system have a negative impact on society at large or democracy? - Did you assess the societal impact of the AI system's use beyond the (end-)user and subject, such as potentially indirectly affected stakeholders or society at large? - Did you take action to minimize potential societal harm of the AI system? - Did you take measures that ensure that the AI system does not negatively impact democracy?
#7 Accountability	**Auditability** Did you establish mechanisms that facilitate the AI system's auditability (e.g. traceability of the development process, the sourcing of training data and the logging of the AI system's processes, outcomes, positive and negative impact)?

	Did you ensure that the AI system can be audited by independent third parties?
	Risk Management Did you foresee any kind of external guidance or third-party auditing processes to oversee ethical concerns and accountability measures? - Does the involvement of these third parties go beyond the development phase? Did you organise risk training and, if so, does this also inform about the potential legal framework applicable to the AI system? Did you consider establishing an AI ethics review board or a similar mechanism to discuss the overall accountability and ethics practices, including potential unclear grey areas? Did you establish a process to discuss and continuously monitor and assess the AI system's adherence to this Assessment List for Trustworthy AI (ALTAI)? - Does this process include identification and documentation of conflicts between the 6 aforementioned requirements or between different ethical principles and explanation of the 'trade-off' decisions made? - Did you provide appropriate training to those involved in such a process and does this also cover the legal framework applicable to the AI system? Did you establish a process for third parties (e.g. suppliers, end-users, subjects, distributors/vendors or workers) to report potential vulnerabilities, risks or biases in the AI system? - Does this process foster revision of the risk management process? For applications that can adversely affect individuals, have redress by design mechanisms been put in place?

ANNEXE C – DENT et al (PARC), 2020

Extrait de : Dent et al. "A Framework for Systematically Applying Humanistic Ethics when Using AI as a Design Material", (2020).

Liste de vérifications préliminaires (*Preliminary Ckecklist)*

When the AI learns from existing data, ask yourself:

- Does your project data represent individuals or populations of people?
- Does the data contain individual personal attributes (especially protected attributes such as race, sex, gender identity, ability, status, socio-economic status, education level, religion, country of origin)?
- Is the goal to make predictions about people's behavior?
- Is the goal to classify people or otherwise make predictions about them?
- Is the goal to make decisions about people or populations of people that could have a significant impact on their lives (e.g. job performance, judicial sentencing, fraudulent activity, etc.)

When AI participates in physical space as with cameras, microphones, sensors, or anything that records human likenesses or activity:

- Is it in a public place?
- Is it hidden from anyone who might be recorded? In other words, could subjects be recorded without knowing they are being recorded?
- Is it possible that any members of vulnerable populations (this could be any disadvantaged sub-segment of an overall population, e.g. children, prisoners, refugees, people facing discrimination) might be recorded?
- Is there heavy equipment or is it operating at high speed?

Answering 'yes' to any of the above questions indicates additional scrutiny is most likely needed.

Directives détaillées (*Comprehensive Guidelines*)

Human data	**Have you taken appropriate data protection steps given the sensitivity of the data?** - Any personally identifiable information is potentially sensitive and, in many contexts, such as medical information, requires different levels of handling when managing data about people. **If the project is for a public entity, can you disclose the sources of your data?** - Where the data came from, who collected it, and for what purposes can affect not only the quality of the data but may also inform other qualities including possible biases. **Have you confirmed that your data accurately reflects the real-world situation for the problem you are trying to solve? Can you also consider alternative data sources?** - Getting the right data can be expensive and sometimes even impossible. Working with data because it's what's available is a surprisingly common practice. It has also led to the unfair treatment of individuals due to mistaken assumptions and wrong conclusions. **Have you checked the internal consistency of the data (through random sampling, for example)? Should you report any issues to stakeholders? How will you assess the data for both explicit and implicit biases?** - Consider specific subpopulations too. Perform appropriate checks for statistical biases and think through whether or not the data matches real-life phenomena. It is important to communicate all limitations and assumptions to all stakeholders. **Will the benefits of your design extend to the entire population or could there be subgroups who are inadvertently excluded?** - In instances where there is a known risk of discrimination, include details of the application and its functions along with samples of the data and details about the data source. Is there a process to take effective action to mitigate discrimination? **For projects that have a significant risk of causing human rights abuse, is it possible to submit them for independent third-party audits?**

	- Third-party involvement may not always be appropriate since many of us do work for private companies; however, if possible, an independent review of the system and its outcomes when it's deployed is a good check on your work especially if the system is making decisions that could have a significant impact on people's lives. **Can stakeholders audit the system?** - At a minimum, you should not be delivering black-box systems that are not inspectable and subject to verification by those involved. **Is there a plan to communicate the decisions about trade-offs, project assumptions, shortcomings, error rates, etc. to stakeholders?** - Throughout the design and development process, you have had to make decisions about assumptions and trade-offs. If those are not captured and communicated, they will never be known beyond the development team. **Is there a process for people to correct data or contest erroneous decisions?** - Systems are fallible and never operate with 100% accuracy. People affected by them must have a way to question the results. **If appropriate, is there a way for data to be removed when necessary (e.g., the right to be forgotten as practiced in the EU and Argentina)?**
Social impact	**AI decisions that could bias or alter societal norms** - Are there any subgroups who benefit more or less from the system? Give special consideration to the effects of errors in the system. Are any groups likely to be harmed more significantly than others? - Think about what services, products, and industries – and perhaps jobs – might be replaced by the deployment of the algorithm/system. How will this impact society? You may not be able to do anything about it, and it might still be appropriate to develop the innovation for the long-term, but from an insider's point-of-view, you may be able to offer insights or suggestions that could mitigate the negative effects your system will have on communities or other systems and groups. - When using human data, do the benefits outweigh the risks to those involved? Consider the benefits of a project and the potential harm to those affected by it. - Are the decisions produced by an algorithmic system explainable to the people affected by those decisions? Explanations help with the validation of results and build trust in the system. - Useful exercise: Imagine yourself subject to all possible outcomes of a system, are they all equally fair and just from all points of view? (Rawls 1999). As you design a system, consider yourself in a position where you are subject to its decisions and predictions. We cannot eliminate all of our inherent biases, but imagine yourself with different natural abilities and a different position in society with different identities related to gender, race, and nationality.
Physical interaction	**Section 2.4 de l'article**
Miseuse or malicious intent	**Are there guardrails in place to prevent off-label usage (the intentional or accidental misuse of the design)?** **Has the system been adequately secured to prevent manipulation from outside?**
Environmental Impact	**Consider the energy impact of the computational resources necessary for the project. Is there a way to minimize that impact? Can the computation be handled differently to use less energy?** - Do the resources provisioned for the project match the requirements, or are they too excessive? **Does the algorithm/system/design encourage unsustainable behavior?**

	- The innovation of coffee pods, for example, has made single-cup, on-demand cups of coffee tremendously convenient, but the pods are generally not recyclable and end up in landfills.
Post deployment	**Do individuals and groups have access to meaningful remedy and redress?** - This may include, for example, creating clear processes for redress following adverse decisions or effects. **What will the reporting process and process for recourse be?** **Is there a plan for what to do if the project has unintended consequences?** - This may be part of a maintenance plan and should involve post-launch monitoring plans.

ANNEXE D – IBM, 2019

Extrait de : IBM. "Everyday Ethics for Artificial Intelligence", (2019).

5 domaines d'intérêt éthique	Recommandations pratiques	Actions à considérer	Questions
1. Accountability	• Make company policies clear and accessible to design and development teams from day one so that no one is confused about issues of responsibility or accountability. As an AI designer or developer, it is your responsibility to know. • Understand where the responsibility of the company/software ends. You may not have control over how data or a tool will be used by a user, client, or other external source. • Keep detailed records of your design processes and decision making. Determine a strategy for keeping records during the design and development process to encourage best practices and iteration. • Adhere to your company's business conduct guidelines. Also, understand national and international laws, regulations, and guidelines8 that your AI may have to work within. You can find other related resources in the IEEE Ethically Aligned Design document.	• Understand the workings of your AI even if you're not personally developing and monitoring its algorithms. • Refer to secondary research by sociologists, linguists, behaviorists, and other professionals to understand ethical issues in a holistic context.	• How does accountability change according to the levels of user influence over an AI system? • Is the AI to be embedded in a human decision-making process, is it making decisions on its own, or is it a hybrid? • How will our team keep records of our process? • How do we keep track of ethical design choices and considerations after the launch of the AI? Will others new to our effort be able to understand our records?
2. Value Alignment	• Consider the culture that establishes the value systems you're designing within. Whenever possible, bring in policymakers and academics that can help your team articulate relevant perspectives. • Work with design researchers to understand and reflect your users' values. You can find out more about this process here. • Consider mapping out your understanding of your users' values and aligning the AI's actions accordingly with an Ethics Canvas.12 Values will be specific to certain use cases and affected communities. Alignment will allow users to better understand your AI's actions and intents.	• If you need somewhere to start, consider IBM's Standards of Corporate Responsibility or use your company's standards documentation. • Values are subjective and differ globally. Global companies must take into account language barriers and cultural differences. • Well-meaning values can create unintended consequences. e.g. a tailored political newsfeed provides users with news that aligns with their beliefs but does not holistically represent the gestalt.	• Which group's values are expressed by our AI and why? • How do we agree on which values to consider as a team? (For more reading on moral alignment, check here. • How do we change or adjust the values reflected by our AI as our values evolve over time?
3. Explainability	• Allow for questions. A user should be able to ask why an AI is doing what it's doing on an ongoing	• Explainability is needed to build	• How do we build explainability into

	basis. This should be clear and up front in the user interface at all times. • Decision-making processes must be reviewable, especially if the AI is working with highly sensitive personal information data like personally identifiable information, protected health information, and/or biometric data. • When an AI is assisting users with making any highly sensitive decisions, the AI must be able to provide them with a sufficient explanation of recommendations, the data used, and the reasoning behind the recommendations. • Teams should have and maintain access to a record of an AI's decision processes and be amenable to verification of those decision processes.	public confidence in disruptive technology, to promote safer practices, and to facilitate broader societal adoption. • There are situations where users may not have access to the full decision process that an AI might go through, e.g., financial investment algorithms. • Ensure an AI system's level of transparency is clear. Users should stay generally informed on the AI's intent even when they can't access a breakdown of the AI's process.	our experience without detracting from user experience or distracting from the task at hand? • Do certain processes or pieces of information need to be hidden from users for security or IP reasons? How is this explained to users? • Which segments of our AI decision processes can be articulated for users in an easily digestible and explainable fashion?
4. Fairness	• Real-time analysis of AI brings to light both intentional and unintentional biases. When bias in data becomes apparent, the team must investigate and understand where it originated and how it can be mitigated. • Design and develop without intentional biases and schedule team reviews to avoid unintentional biases. Unintentional biases can include stereotyping, confirmation bias, and sunk cost bias. • Instill a feedback mechanism or open dialogue with users to raise awareness of user-identified biases or issues. e.g., Woebot asks "Let me know what you think," after suggesting a link.	• Diverse teams help to represent a wider variation of experiences to minimize bias. Embrace team members of different ages, ethnicities, genders, educational disciplines, and cultural perspectives. • Your AI may be susceptible to different types of bias based on the type of data it ingests. Monitor training and results in order to quickly respond to issues. Test early and often.	• How can we identify and audit unintentional biases that we run into during the design and development of our AI? • The status quo changes over time. How do we instill methods to reflect that change in our ongoing data collection? • How do we best collect feedback from users in order to correct unintentional bias in design or decision-making?
5. User Data Rights	• Users should always maintain control over what data is being used and in what context. They can deny access to personal data that they may	• Employ security practices including encryption, access control	• What types of sensitive personal data does the AI utilize and how will

	find compromising or unfit for an AI to know or use. • Users' data should be protected from theft, misuse, or data corruption. • Provide full disclosure on how the personal information is being used or shared. • Allow users to deny service or data by having the AI ask for permission before an interaction or providing the option during an interaction. Privacy settings and permissions should be clear, findable, and adjustable. • Forbid use of another company's data without permission when creating a new AI service. • Recognize and adhere to applicable national and international rights laws when designing for an AI's acceptable user data access permissions.	methodologies, and proprietary consent management modules to restrict access to authorized users and to de-identify data in accordance with user preferences. • It is your responsibility to work with your team to address any lack of these practices.	this data be protected? • What contractual agreements are necessary for data usage and what are the local and international laws that are applicable to our AI? • How do we create the best user experience with the minimum amount of required user data?

ANNEXE E – AMERSHI et al (Microsoft), 2019

Extrait de : Amershi et al (Microsoft). "Guidelines for Human-AI Interaction", à la p. 3, (2019).

		AI Design Guidelines	Example Applications of Guidelines
Initially	G1	**Make clear what the system can do.** Help the user understand what the AI system is capable of doing.	[Activity Trackers, Product #1] "Displays all the metrics that it tracks and explains how. Metrics include movement metrics such as steps, distance traveled, length of time exercised, and all-day calorie burn, for a day."
	G2	**Make clear how well the system can do what it can do.** Help the user understand how often the AI system may make mistakes.	[Music Recommenders, Product #1] "A little bit of hedging language: 'we think you'll like'."
During interaction	G3	**Time services based on context.** Time when to act or interrupt based on the user's current task and environment.	[Navigation, Product #1] "In my experience using the app, it seems to provide timely route guidance. Because the map updates regularly with your actual location, the guidance is timely."
	G4	**Show contextually relevant information.** Display information relevant to the user's current task and environment.	[Web Search, Product #2] "Searching a movie title returns show times in near my location for today's date"
	G5	**Match relevant social norms.** Ensure the experience is delivered in a way that users would expect, given their social and cultural context.	[Voice Assistants, Product #1] "[The assistant] uses a semi-formal voice to talk to you - spells out "okay" and asks further questions."
	G6	**Mitigate social biases.** Ensure the AI system's language and behaviors do not reinforce undesirable and unfair stereotypes and biases.	[Autocomplete, Product #2] "The autocomplete feature clearly suggests both genders [him, her] without any bias while suggesting the text to complete."
When wrong	G7	**Support efficient invocation.** Make it easy to invoke or request the AI system's services when needed.	[Voice Assistants, Product #1] "I can say [wake command] to initiate."
	G8	**Support efficient dismissal.** Make it easy to dismiss or ignore undesired AI system services.	[E-commerce, Product #2] "Feature is unobtrusive, below the fold, and easy to scroll past...Easy to ignore."
	G9	**Support efficient correction.** Make it easy to edit, refine, or recover when the AI system is wrong.	[Voice Assistants, Product #2] "Once my request for a reminder was processed I saw the ability to edit my reminder in the UI that was displayed. Small text underneath stated 'Tap to Edit' with a chevron indicating something would happen if I selected this text."
	G10	**Scope services when in doubt.** Engage in disambiguation or gracefully degrade the AI system's services when uncertain about a user's goals.	[Autocomplete, Product #1] "It usually provides 3-4 suggestions instead of directly auto completing it for you"
	G11	**Make clear why the system did what it did.** Enable the user to access an explanation of why the AI system behaved as it did.	[Navigation, Product #2] "The route chosen by the app was made based on the Fastest Route, which is shown in the subtext."
Over time	G12	**Remember recent interactions.** Maintain short term memory and allow the user to make efficient references to that memory.	[Web Search, Product #1] "[The search engine] remembers the context of certain queries, with certain phrasing, so that it can continue the thread of the search (e.g., 'who is he married to' after a search that surfaces Benjamin Bratt)"
	G13	**Learn from user behavior.** Personalize the user's experience by learning from their actions over time.	[Music Recommenders, Product #2] "I think this is applied because every action to add a song to the list triggers new recommendations."
	G14	**Update and adapt cautiously.** Limit disruptive changes when updating and adapting the AI system's behaviors.	[Music Recommenders, Product #2] "Once we select a song they update the immediate song list below but keeps the above one constant."
	G15	**Encourage granular feedback.** Enable the user to provide feedback indicating their preferences during regular interaction with the AI system.	[Email, Product #1] "The user can directly mark something as important, when the AI hadn't marked it as that previously."
	G16	**Convey the consequences of user actions.** Immediately update or convey how user actions will impact future behaviors of the AI system.	[Social Networks, Product #2] "[The product] communicates that hiding an Ad will adjust the relevance of future ads."
	G17	**Provide global controls.** Allow the user to globally customize what the AI system monitors and how it behaves.	[Photo Organizers, Product #1] "[The product] allows users to turn on your location history so the AI can group photos by where you have been."
	G18	**Notify users about changes.** Inform the user when the AI system adds or updates its capabilities.	[Navigation, Product #2] "[The product] does provide small in-app teaching callouts for important new features. New features that require my explicit attention are pop-ups."

Table 1: Our 18 human-AI interaction design guidelines, roughly categorized by when they likely are to be applied during interaction with users, along with illustrative applications (rated as "clearly applied" by participants) across products tested by participants in our user study.

ANNEXE F – BENJAMINS et al (Telefonica), 2019

Extrait de : Benjamins et al (Telefonica). : "Responsible AI by Design in Practice", tableaux des pp 3-4, (2019).

Voir Table 1, Table 2* et *Table 3.

initiative involving different departments such as Engineering, Security, Legal, Business, IT, Human Resources, Procurement, as well as an endorsement of top management.

Table 1 below illustrates how the Principles are implemented in the organization. For each principle, several questions are defined that must be answered by the respective responsible persons. Several of the questions require certain understanding of AI and Machine Learning, and therefore specific tools and training are required. The next sections will indicate the details and proposals for each of the ingredients mentioned before.

Table 1: Questions and their corresponding proposals.

Principle	Question to be asked	Implemented through
Fair AI	Does your data set contain sensitive variables?	Training
	Does any of the variables strongly correlate with sensitive variables?	Technical tool
	Is/are your training data set(s) biased with respect to the target groups in case those include "protected groups"?	Technical tool
	Is there an important impact in the specific domain of false positives (FP) and/or false negatives (FN)?	Training
	Are FP and FN unequally distributed across different (protected) groups	Technical tool
Transparent & Explainable AI	Could the user think that s/he interacts with a person rather than with your system?	Training
	Is the AI system's outcome significantly affecting people's lives?	Training
	Do you lack sufficient understanding of how the AI-generated decisions are constructed for the domain at hand?	Training
	Could the user request an explanation for the AI-generated conclusion?	Training
	Is it difficult to be explicit about whether the data used is personal or non-personal, and about the purpose the AI system uses the data for?	Training
	Is it possible to understand how the algorithm has reached its conclusions? For example, what variables have influenced the result of the algorithm and how much?	Technical tool
Human-centric AI	Is there a possibility that your P&S has a negative impact on Human Rights?	Training
	Does your P&S negatively impact the UN's SDGs?	Training
Privacy & Security by Design	Does your AI system use personal data?	Training
	Has your Privacy Impact Assessment revealed any important concerns?	Training
	In case your P&S uses anonymized data, is there an unreasonable risk of re-indentification?	Technical tool
	Has your Security Assessment revealed any important concerns?	Training
	Is the system robust against attacks that seek to exploit weaknesses in it and manipulate the outputs?	Technical tool
Third parties	Do you need more information from your supplier to understand whether the AI module is consistent with the Principles?	Training

Table 2: Key content of the training components for each of the AI principles and questions.

Principle	Question to be asked	Training content
Fair AI	Does your data set contain sensitive variables?	Explain what is sensitive personal data as defined by law
	Is there an important impact of false positives (FP) and/or false negatives (FN)?	Show examples of FP and FN and their concrete impact on people
Transparent & Explainable AI	Could the user think that s/he interacts with a person rather than with your system?	Show examples of people and machines interacting with different degrees of machine sophistication
	Is the AI system's outcome significantly affecting people's lives?	Show examples of different degrees of decisions' impact on peoples' lives
	Do you lack sufficient understanding of how the AI-generated decisions are constructed for the domain at hand?	Explain what types of algorithms are explainable (e.g. decision trees) and what not (e.g. deep learning). Show different ways of how AI decisions can be explained (algorithm, outcome)
	Could the user request an explanation for the AI-generated conclusion?	This may require building in certain features during design
	Is it difficult to be explicit about whether the data used is personal or non-personal, and about the purpose the AI system uses the data for?	Explain what is personal data, psydonymized data and anonymous data. Give examples of how certain types of data are used for different purposes in AI systems
Human-centric AI	Is there a possibility that your P&S has a negative impact on Human Rights?	General training on Human Rights and examples of P&S that impact them in both positive and negaitive ways. See Ethics Canvas (ADAPT 2017)
	Does your P&S negatively impact the UN's SDGs?	Explain the 17 SDGs of the UN and give examples of P&S that impact them in both positive and negative ways
Privacy & Security by Design	Does your AI system use personal data?	Explain what is personal data, psydonymized data and anonymous data.
	Has your Privacy Impact Assessment revealed any important concerns?	Explain Privacy by Design, and what a PIA is, using for examples templates from ICO (ICO)
	Has your Security Assessment revealed any important concerns?	Explain Security by Design
Third parties	Do you need more information from your supplier to understand the AI module is consistent with the Principles?	Explain what kind of questions to ask a provider and how to interpret the answers (Hind et al 2018)

Technical tools required

Many of the tools required to support "Responsible AI by Design" are still in early stage, though they are being developed quickly and its expected that the companies include them soon in their processes. The technical tools proposal that appears within this methodology spans from the usage of stand-alone tools to the reference of more academic research solutions in order to offer a complete portfolio that could be useful to all the profiles within the company.

Table 3 Summarizes the key functionalities of the required tools.

Table 3: Types of technical tools to help visualize and detect potential problems.

Principle	Question to be asked	Type of tool
Fair AI	Does any of the variables strongly correlate with sensitive variables?	Check correlations between all variables in data set and visualize result
	Is/are your training data set(s) biased with respect to the target groups in case those include "protected groups"?	Check for disparate impact, i.e. whether different subgroups are treated differently
	Are FP and FN unequally distributed across different (protected) groups	Check for predictive parity, i.e. what is the false positive rate for different subgroups as well as for the overall population
Transparent & Explainable AI	Is it possible to understand how the algorithm has reached its conclusions? For example, what variables have influenced the result of the algorithm and how much?	Define what transparency level is required according to the profile of the users of the system and use an available solution to generate local or global explanations
Privacy & Security by Design	In case your P&S uses anonymized data, is there an unreasonable risk of re-indentification?	Notify when anonymization algorithm is not sufficient to render data sets anonymous
	Is the system robust against attacks that seek to exploit weaknesses in it and manipulate the outputs?	Before passing a model to production, it is recommended to assess its vulnerabilities to these attacks with the tools available

ANNEXE G – MICROSOFT, 2018

Extrait de : Microsoft. "Responsible bots : 10 guidelines for developers of conversational AI', (2018).

Responsible bots: 10 guidelines for developers of conversational AI

More people are using bots in their everyday lives, whether it's to get a quick answer to a customer service problem or to help people out with things like managing their calendars, checking the weather or ordering pizza. Bots, or more generally, conversational AI, have the ability to help people achieve more, and we are only starting to see their potential to augment what we can do.

In order for people and society to realize the full potential of bots, they need to be designed in such a way that they earn the trust of others. These guidelines are aimed at helping you to design a bot that builds trust in the company and service that the bot represents. These guidelines are not intended as legal advice and you should separately ensure that your bot complies with the fast-paced developments in the law in this area. Also, in designing your bot, you should consider a broad set of responsibilities you have when developing any data-centric AI system, including ethics, privacy, security, safety, inclusion, transparency and accountability. See, for example, Microsoft's six principles for the responsible development of AI published in the January 2018 book, The Future Computed.

These guidelines are just that — guidelines, and for the most part not hard-and-fast rules. They are most relevant to bots that may affect people in *consequential ways* — such as helping people to navigate information relating to employment, finances, health or the like. You should use your best judgment when applying these guidelines, always with a view toward the ultimate question of whether your design will deliver an experience that end users appreciate, in a manner that builds their trust in your company and services. These are v.1.0 guidelines, so we fully expect that they will be revised over time in response to your feedback and our own experiences.

Guidelines

1. Articulate the purpose of your bot and take special care if your bot will support consequential use cases.

The purpose of your bot is central to ethical design, and ethical design is particularly important when it is anticipated that a consequential use will be served by the bot you are developing. Consequential use cases include access to services such as healthcare, education, employment, financing or other services that, if denied, would have meaningful and significant impact on an individual's daily life.

- **Before beginning design work, carefully specify how your bot will benefit the user or the entity deploying the bot.** If the bot is likely to affect the well-being of the user, such as providing access to a consequential service like healthcare, attention to these guidelines will be especially important. Be sure to pause to research, learn and deliberate on the impact of the bot on people's lives. When in doubt, seek guidance.
- **Assess whether the bot's intended purpose can be performed responsibly.** Some purposes may inherently require human judgment, empathy and expertise or a very high degree of reliability and accuracy, e.g., healthcare diagnosis or financial planning. Be sure to consider the nature and type of errors in the performance of the bot and their cost to users. Consider if you have access to relevant expertise in the domain in which your bot would operate.
- **Develop metrics to assess user satisfaction.** Metrics for your bot should cover not only whether the user feels that the bot served its intended purpose, but also the user's sense of well-being and comfort while using the bot.

2. Be transparent about the fact that you use bots as part of your product or service.

Users are more likely to trust a company that is transparent and forthcoming about its use of bot technology, and a bot is more likely to be trusted if users understand that the bot is working to serve their needs and is clear about its limitations.

- **It should be apparent to the user that they are not having an interaction with another person.** Since designers might endow their bots with "personality" and natural language capabilities, it is important to convey to users that they are not interacting with another person and some aspects of their interaction are being performed by a bot. There are variety of design choices that can be made to accomplish this that do not degrade the user experience.
- **Establish how the bot can help and the limitations associated with its use.** Users are more likely to find a bot to be trustworthy if the bot sets reasonable expectations for what it can do and what it does not do well. Users should be able to easily find information about the limitations of the bot, including the possibility of errors and the consequences that can flow from such errors. For users who wish to "learn more," you should offer a more detailed explanation of the purpose and operation of the bot.

3. Ensure a seamless hand-off to a human where the human-bot exchange leads to interactions that exceed the bot's competence.

If your bot will engage people in interactions that may require human judgment, provide a means or ready access to a human moderator.

- **Respect individual engagement preferences, particularly if your bot deals in consequential matters.** Bots designed for use in consequential matters should incorporate the ability to transfer an engagement to a human moderator as soon as the user asks, otherwise indicates, or if the bot recognizes (e.g., through sentiment analysis) that the user is dissatisfied. If users feel trapped or alienated by a bot, they will quickly lose trust in the technology and in the company that has deployed it.

4. Design your bot so that it respects relevant cultural norms and guards against misuse.

Since bots may have human-like personas, it is especially important that they interact respectfully and safely with users and have built-in safeguards and protocols to handle misuse and abuse.

- **Limit the surface area for norms violations where possible.** Every bot should be designed to follow a specific set of values and cultural norms. To reduce the possibility of conflicting with those values and cultural norms, limit the surface area for norms violations. For example, if your bot is designed to take pizza orders, limit it to that purpose only, so that it does not engage on topics such as race, gender, religion, politics and the like.
- **Where appropriate, point to a relevant "code of conduct" for users.** Consider whether your bot should be subject to a user code of conduct (from your organization or the entity deploying the bot) that, for example, includes prohibitions on hate speech, bullying and threatening others, and provides appropriate notice to the user of any code of conduct.
- **Apply machine learning techniques and keyword filtering mechanisms to enable your bot to detect and — critically — respond appropriately to sensitive or offensive input from users.** Deploy a two-way filtering mechanism with a customizable threshold of tolerance for what your bot takes in from users, as well as what your bot says in response. In most cases, we

recommend the bots simply steer clear of controversial subjects (especially hate speech). Open domain conversations are considered high-risk because they require significant investment in both content operations and social media monitoring capabilities and must be maintained 24/7 with bugfix service level agreements. You should leverage products that include offensive text classifiers, such as the Microsoft Bot Framework, , to protect your bot from abuse if it engages in open domain conversations. Sensitive categories include adult content, extremism, drugs, alcohol and tobacco, profanity, vulgarity, harassment, bullying, violence and gore, and hate speech (relating, for example, to ethnicity or race, gender identity or sexuality, religion, or people with disabilities). Public-facing bot APIs should also be reviewed to assess whether they could be used by people outside your organization to create a bot that would engage in hate speech or otherwise reflect poorly on your organization.

5. Ensure your bot is reliable.

Ensure that your bot is sufficiently reliable for the function it aims to perform, and always take into account that since AI systems are probabilistic they will not always provide the correct answer.

- **Establish reliability metrics and review them periodically.** Consider what questions your bot needs to answer and rigorously test its performance and ongoing effectiveness. Because the performance of AI-based bot systems may vary from development to implementation, and over time as the bot is rolled out to new users and in new contexts, it is important to continually monitor reliability. Reliability signals can be developed to help drive decisions about when to pass the baton to a human, or when a bot should announce that it cannot perform the requested function reliably. If an AI-based bot system can determine that it has made a mistake, that fact should be communicated to the user.
- **Be transparent about bot reliability.** Particularly for bots operating in sensitive domains, you should make available information concerning the reliability of the bot, such as summaries of general statistical performance, performance under particular circumstances, or in the context of specific examples.
- **Build traceability capabilities into your bot.** When something goes wrong with your bot during a high-value interaction, it is important to have traceability (monitoring and auditing), for example through Microsoft Azure Application Insights, in order to troubleshoot the issue. For more information on Application Insights, refer to: https://azure.microsoft.com/en-us/services/application-insights/.
- **Provide a feedback mechanism.** Users will feel more comfortable with bots if they can provide feedback on their operation (and feedback is essential in any event, as with all product development work). Bots should actively ask for feedback. Set expectations as to whether the user will get any response to feedback provided.
- **For sensitive uses, obtain domain expertise.** If you are building a bot to deliver services in areas such as health, employment, finance or law enforcement, ensure that you obtain and take account of input from experts in these areas as you design and deploy your bot.

6. Ensure your bot treats people fairly.

The possibility that AI-based systems will perpetuate existing societal biases, or introduce new biases, is one of the top concerns identified by the AI community relating to the rapid deployment of AI. Development teams must be committed to ensuring that their bots treat all people fairly.

- **Systematically assess the data used for training your bot.** Systematically assess the data used for training your bot to ensure that it has appropriate representativeness and quality, and take steps to understand the lineage and relevant attributes of the training data. As bias detection tools become more broadly available, adopt them as an additional means to ensure the fairness of your bot and make such tools available for customer use and adoption.
- **Strive for diversity amongst your development team.** Employing a diverse team focused on the design, development and testing of bot technology will help ensure that your bot operates fairly.

7. Ensure your bot respects user privacy.

Privacy considerations are especially important for bots. While the Microsoft Bot Framework does not store session state, you may be designing and deploying authenticated bots in personal settings (like hospitals) where bots will learn a great deal about users. People may also share more information about themselves than they would if they thought they were interacting with a person. And, of course, bots can remember everything. All of this (plus legal requirements) makes it especially important that you design bots from the ground up with a view toward respecting user privacy. This includes giving users sufficient transparency into bots' data collection and use, including how the bot functions, and what types of controls the bot offers users over their personal data.

- **Inform users up front about the data that is collected and how it is used and obtain their consent beforehand.** Provide easy access to a valid privacy statement and applicable service agreement and include a "profile page" for users to obtain information about the bot with links to relevant privacy and legal information. Making this information available and easily accessible in the "first run" experience is highly recommended.
- **Collect no more personal data than you need, limit access to it and store it for no longer than needed.** Collect only the personal data that is essential for your bot to operate effectively. If your bot will share data (such as with another bot), be sure only to share the minimum amount of user data necessary in order to compete the requested function on behalf of the user. If you enable access by other agents to your bot's user data, do so only for the minimum time necessary in order to compete the requested function. Always give users the opportunity to choose which agents your bot will share data with and what data is suitable for sharing. Consider whether you can purge stored user data from time to time while still enabling your bot to learn. Shorter retention periods minimize security risks for users and will help to position your bot as privacy-friendly.
- **Provide privacy-protecting user controls.** For bots that store personal information, such as authenticated bots, consider providing an easy-to-find "Show me all you know about me" button, and similar buttons to "Forget my last interaction," "Delete all you know about me," and so forth. In some cases, such buttons may be legally required.
- **Obtain legal and privacy review.** The privacy aspects of bot design are subject to important and increasingly stringent legal requirements. Be sure to obtain both a legal and a privacy review of your bot's privacy practices through the appropriate channels in your organization.

8. Ensure your bot handles data securely.

Users have every right to expect that their data will be handled securely. Follow security best practices that are appropriate for the type of data your bot will be handling.

- **Establish secure development and secure operations foundations.** Traditional secure software foundations are critical. As with any AI system, your bot should ensure proper authentication, separation of duty, input validation and mitigations for denial-of-service attacks.
- **Your bot should be resilient.** Design your bot to identify abnormal behaviors and prevent manipulation. Pinpoint "users" (who could in fact be malicious bots) who deviate from norms established by large clusters of other users — such as users who seem to respond too fast, don't sleep, or trigger parts of your bot code paths that other users do not.
- **Ensure the integrity of your training data.** All AI systems must be able to distinguish between maliciously introduced data (which must be purged) and data that is merely rare, yet valid and potentially important. This is particularly critical for bots which employ automatic or supervised learning techniques.
- **Obtain security review.** If available, work with the appropriate security team within your organization to conduct a security review on your bot and supporting services. Given the close relationship of security and privacy in this space, a joint security/privacy review is recommended to ensure the best depth and breadth of coverage.

9. Ensure your bot is accessible.

Bots can benefit everyone, but only if they are designed to be inclusive and accessible to people of all abilities. Microsoft's mission to empower every person to achieve more includes ensuring that new technology interfaces can be used by people with disabilities, including users of assistive technology.

- **If you are developing a bot, consider how your bot complies with commonly used accessibility standards, such as WCAG 2.0 AA, and U.S. Section 508 and EN 301 549 standards.** Customers with disabilities should be able to use your bot as effectively as those without disabilities. Complying with the international web accessibility standard WCAG 2.0 AA (codified as ISO 40500:2012) and U.S. and European procurement standards will help enable users who rely on screen readers, navigate UI using only keyboard, are hard of hearing, require color contrast or cannot distinguish between colors, to use your bot. Many of these requirements carry dependencies on the conversational canvas.
- **Have people with disabilities test your bots.** In addition to complying with accessibility standards, getting feedback from users with disabilities on your bot prior to launch will help determine whether the bot can be used as intended by the broadest possible audience.
- **Design bots to respect the full range of human abilities.** Use Microsoft's Inclusive Design toolkit to design bots which recognize exclusion, learn from diversity and solve for ability constraints.

10. Accept responsibility.

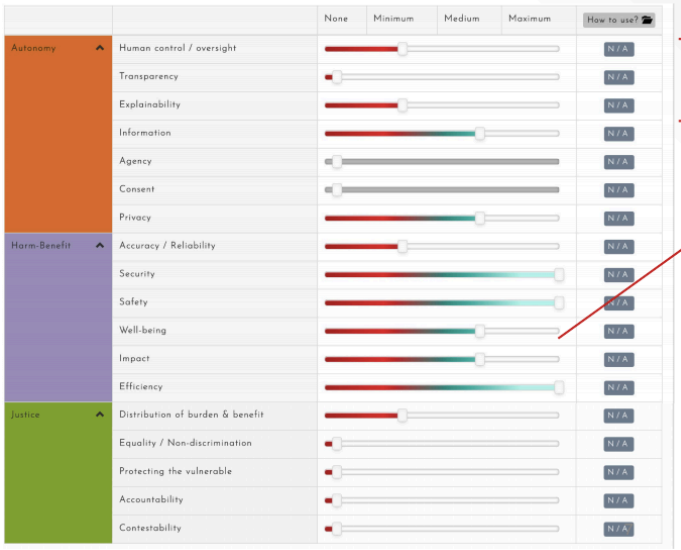
We are a long way away from bots that can truly act autonomously, if that day will ever come. Humans are accountable for the operation of bots.

- **Developers are accountable for the bots they deploy.** If you are developing a bot that your organization will deploy, you should recognize that you are fully responsible for its operation and how it affects people. If you are designing a bot to be deployed by a third party, come to a shared understanding with them of who is ultimately responsible for the bot and document that understanding.

ANNEXE H – AI ETHICS LAB, 2020

Extrait de : AI Ethics Lab. "Tool: The Box", (2020).

Autonomy	human control / oversight	*Does the system allow for human oversight and control?* *Does the system provide the necessary information for meaningful oversight and control?*
	transparency	*Are the system's abilities and limitations clear to the users and those who are subject to the system?* *Is the system transparent in how it affects individual decision-making?*
	explainability	*Is it explainable how the system reaches its decisions and outcomes?* *Can humans understand how the system produces its results?*
	information	*Does the system provide individuals with accurate and relevant information regarding the system?* *Does the system enable individuals' access to accurate and relevant information for decision-making?*
	agency	*Does the system enable individuals pursue their goals or help their pursuit?*
	consent	*Is the system designed to ensure rational, voluntary, and informed consent of individuals when they use its functions?*
	privacy	*Does the system allow individuals control their privacy? Does the system protect individual privacy?*
Harm-Benefit	accuracy / reliability	*Is the system accurate and reliable? Is the system designed to be measured for accuracy and reliability?*
	security	*Is the system secure and designed against security attacks?*
	safety	*Is the system safe for users and those who are subject to the system?* *Is the system designed to minimize risks of harm from the system or any other risks of harm?*
	well-being	*Does the system promote and protect individual and societal well-being?* *Does the system maximize well-being and benefits in society?*
	impact	*Does the system have a significant beneficial impact? Is the system designed to be measured for impact?*
	efficiency	*Is the system efficient in achieving the set goals? Is the system designed to be measured for efficiency?*
Justice	distribution of burden & benefit	*Is the system designed and developed to ensure (1) not putting disproportionate burden on a group and (2) not concentrating benefits on another group*
	equality / non-discrimination	*Is the system developed with an understanding of social equality and aim to ensure it?* *Is the system designed to reduce or eliminate unfair bias?*
	protecting the vulnerable	*Does the system protect vulnerable individuals and the worst-off from harm and aim to benefit them?*
	accountability	*Is the system designed for accountability and auditability? Is the system designed for traceability?*
	contestability	*Does the system have a mechanism for appeals?*



Low scores on transparency, explainability, & human control:

The system is a "black box" and its function and limitations are unclear to the applicants and to the HR.

Average scores on well-being & impact:

The system might be overall beneficial and impactful.

Low scores on justice:

The system uses historical data that is most likely to be biased against female applicants. Masking the gender is unlikely to be enough to solve this problem.

The system also does not have mechanisms for appeal and accountability.

AI ETHICS LAB

ANNEXE I – DIGITAL CATAPULT, 2018

Extrait de : Digital Catapult. "AI Ethics Framework", (2018).

Les 7 concepts éthiques et les questions correspondantes	
Concepts	**Questions**
1. Clear benefits	**What are the goals, purposes and intended applications of the product or service?** • Who or what might benefit from the product/service? Consider all potential groups of beneficiaries, whether individual users, groups or society and environment as a whole. • Are those benefits common to the application type, or specific to the technology or implementation choices? • How can the products or services be monitored and tested to ensure they meet these goals, purposes and intended applications? • How likely are the benefits and how significant? How can it be assessed what the benefits are? • How are these benefits obtained by the various stakeholders? • Can the benefits of the product/service be demonstrated? • Might these benefits change over time? • What is the company's position on making (parts of) the products/services available on a non-commercial basis, or on sharing AI knowledge which would enable more people to develop useful AI applications?
2. Know and manage the risks	• Have the risks of other foreseeable uses of the technology, including accidental or malicious misuse of it, been considered? • Have all potential groups at risk, whether individual users, groups or society and environment as a whole, been considered? • Is there currently a process to classify and assess potential risks associated with the use of the product or service? • Who or what might be at risk from the intended and non-intended applications of the product or service? • Consider all potential groups at risk, whether individual users, groups, society as a whole or the environment. • Are those risks common for application area or technology, or specific to the technology or implementation choices? • How likely are the risks, and how significant? Is there a plan to mitigate and manage the risks? • How can the potential risks or perceived risks to users, potentially affected parties, purchasers or commissioners be communicated? • How do third-parties or employees report potential vulnerabilities, risks or biases, and what processes are in place to handle these issues and reports? • How can it be known whether a bias has been created or reinforced with the system? • As a result of assessing potential risks, are there customers or use cases that would be chosen not to work with?

	• How are these decisions made and documented?
3. Use data responsibly	• How was the data obtained and was consent obtained (if required)? Is the data current? • Is the training data appropriate for the intended use? • Is the data pseudo-anonymised or de-identified? If not, why not? • Is the data usage proportionate to the problem being addressed? • Is there sufficient data coverage for all intended use-cases? • What are the qualities of the data (for example, is the data coming from a system prone to human error?) • Have potential biases in the data been examined, well-understood and documented and is there a plan to mitigate against them? • Is there a process for discovering and dealing with inconsistencies or errors in the data? • What is the quality of the data analysis? How much uncertainty / error is there? • What are the consequences which might arise from errors in analysis and how can they be mitigated? • Can it be clearly communicated how the data is being used and how decisions are being made? • What systems are in place to ensure data security and integrity? • Are there adequate methods in place for timely and auditable data deletion, once data is no longer needed? • Can individuals remove themselves from the dataset? Can they also remove themselves from any resulting models? • Is there a publicly available privacy policy in place, and to what extent are individuals able to control the use of data about them, even when they are not users of the service or product? • Are there adequate mechanisms for data curation in place to ensure external auditing and replicability of results, and, if a risk has manifested itself, attribution of responsibility? • Can individuals access data about themselves? Is data being made available for research processes?
4. Be worthy of trust	• Within the company, are there sufficient processes and tools built-in to ensure meaningful transparency, auditability, reliability and suitability of the product output? • Have the limitations of the company's experience on the system being built been acknowledged and how can these reflect on the system in place? What steps are being taken to address these limitations? • Is the nature of the product or technology communicated in a way that the intended users, third parties and the general public can access and understand? • Are (potential) errors communicated and their impact explained? • Does the company actively engage with its employees, purchasers/commissioners, suppliers, users and affected third-parties so that ethical (including safety, privacy and security) concerns can be voiced, discussed, and addressed? • Does the company work with researchers where appropriate to explore or question areas of the technology? • Is there a process to review and assure the integrity of the AI system over time and take remedial action if it is not operating as intended? • If human labour has been involved in data preparation (eg image labelling by Mechanical Turk workers) have the workers involved been fairly compensated?

	• If data comes from another source, have the data owner's rights been preserved (eg copyright, attribution) and has permission been obtained? • Who is accountable if things go wrong? Are they the right people? Are they equipped with the skills and knowledge they need to take on this responsibility? • What is/are the quality or standards to which the product / technology must conform (e.g. academic; peer-review, technical), what are the reasons for choosing the particular standards; and what does the company propose to do to maintain such standards? • In order to engender trust, are there customers, suppliers or use cases that may not be worked with? How are these decisions made and documented? • Does the company have a clear and easy to use system for third party/user or stakeholder concerns to be raised and handled? • Is adequate training in the safe and secure use of your product or service provided to all of your operators, customers and/or users? • Has it been considered how to embed ethics within the organisation? • Has it been considered how to embed integrity and fair dealing in the culture? • How would a person raise a concern with the company? • To inform your processes and culture, could mentors or innovation hubs be consulted?
5. Diversity, equality and inclusion	• Are there processes in place to establish whether the product or service might have a negative impact on the rights and liberties of individuals or groups? Please consider: – varied social backgrounds and education levels – different ages – different gender and or sexual orientation – different nationalities or ethnicity – different political, religious and cultural backgrounds – physical or hidden disabilities. • What actions can be taken if negative impacts are identified? • Social impact can be difficult to demonstrate: have processes been considered that can demonstrate the positive impact the product or service brings? • Has putting in place a diversity and inclusiveness policy in relation to recruitment and retention of staff been considered? • Has how to balance the specific responsibilities of a startup against other factors such as cost and freedom of choice for users been considered? • Are potential biases in the data and processes are examined, well-understood and documented and is there a plan to mitigate against them? • Where do hiring practices and building culture fit in? For instance, are ethical questions raised at interviews? • Are any principles/risk considerations communicated to new hires? • Does the company have a diversity and inclusiveness policy in relation to recruitment and retention of staff?

6. Transparent communication	• Does the company communicate clearly, honestly and directly about any potential risks of the product or service being provided? • What does it communicate and when? Does the company communicate clearly, honestly and directly about the processes in place to avoid, minimise or mitigate potential risks? • Does the company have a clear and easy to use system for third party/user or stakeholder concerns to be raised and handled? • Are the company's policies relating to ethical principles available publicly and to employees? • Are the processes to implement and update the policies open and transparent? • Does the company disclose issues other than the product e.g. projects, studies and other activities funded by the company or which the company may work in conjunction or otherwise be involved with; the major sources of data and expertise that inform the insights of AI solutions and the methods used to train those systems and solutions? • Has a communication strategy and process if something goes wrong been considered?
7. Business model	• What kind of corporate structure best meets the company's needs? As well as the traditional company limited by shares, there are a variety of 'social enterprise' alternatives, including community interest company, co-operative, B-Corp and company limited by guarantee. Are any of these of interest? • Data exchange: are free services being provided in exchange for user data? Are there any ethical implications for this? • Do users have a clear idea of how the data will be used, including any future linking/sale of the data? • What happens if the company is acquired? For example, what happens to its data and software? • Pricing: have differential prices been considered ? Are there any ethical considerations regarding the pricing strategy? • Are there any vulnerable groups that may be offered lower prices? • Data philanthropy: is there data that others (e.g. charities, researchers) could use for public purpose benefits? • Is integrity and fair dealing embedded in the organisational culture? • Has the environmental impact of development / deployment of the technology been considered? • Is environmental impact considered when choosing suppliers? Have less energy-intensive options been considered?

ANNEXE J – AIEI GROUP, 2020

Extrait de : AIEI GROUP. "AI Ethics Impact Group: From Principles to Practice – VDE", (2020).

Exemple d'attribution d'une note à la valeur *Transparency* à l'aide de l'approche VCIO.

*Les cases de la *Figure 2* se transposent sur les cases correspondantes de la *Figure 1*.

Dans cet exemple, la valeur *Transparency* obtient la note de C.

- Les notes des éléments observables deviennent les notes des indicateurs correspondants.
  - Les indicateurs 1.1.1 et 1.1.2 obtiennent donc les notes de A et B respectivement.
  - Les indicateurs 1.2.1 et 1.2.2 obtiennent donc les notes de A et C respectivement.
- Le critère 1.1 obtient la note de B, soit la note la plus faible des indicateurs 1.1.1 et 1.1.2.
- Le critère 1.2 obtient la note de C, soit la note la plus faible des indicateurs 1.2.1 et 1.2.2.
- La valeur *Transparency* obtient la note de C, soit la note la plus faible des deux critères.
- La note de C est transposée sur le AI Ethics Label (*Figure 3*).

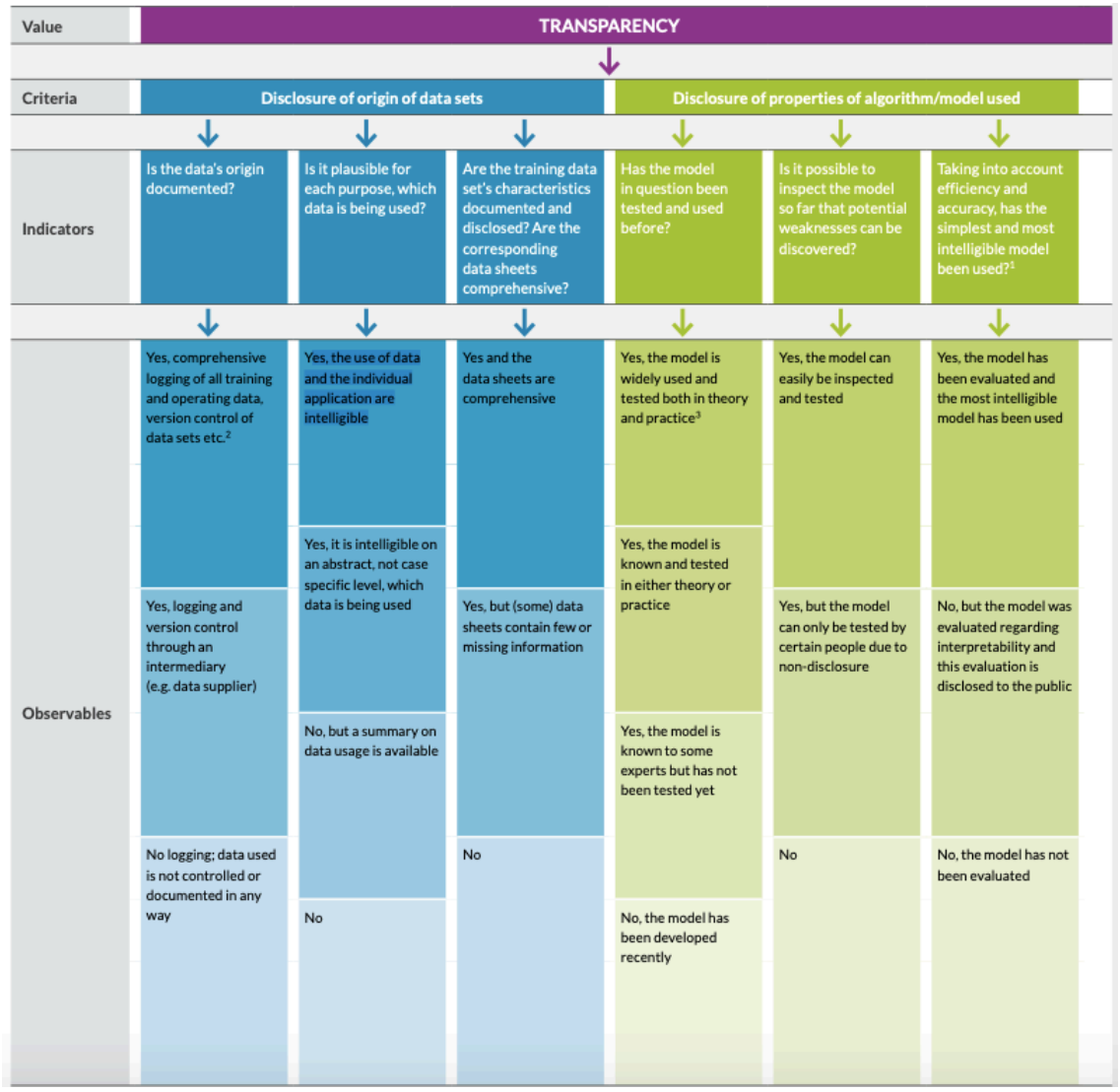


Figure 1

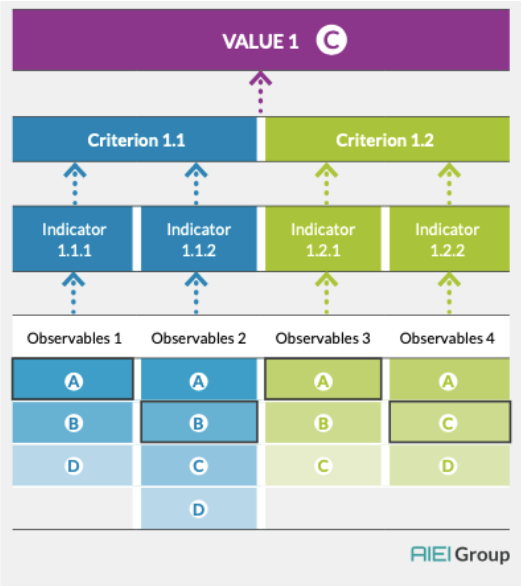


Figure 2

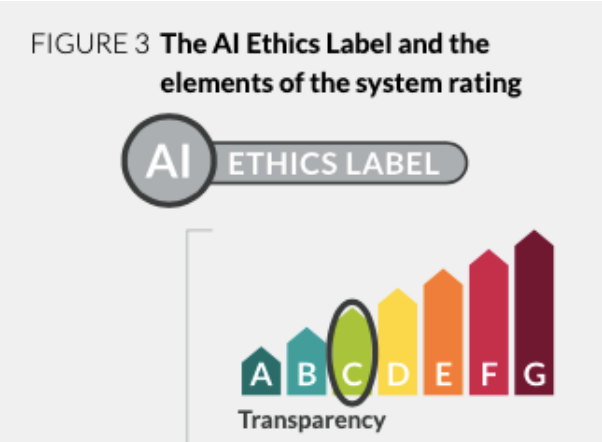
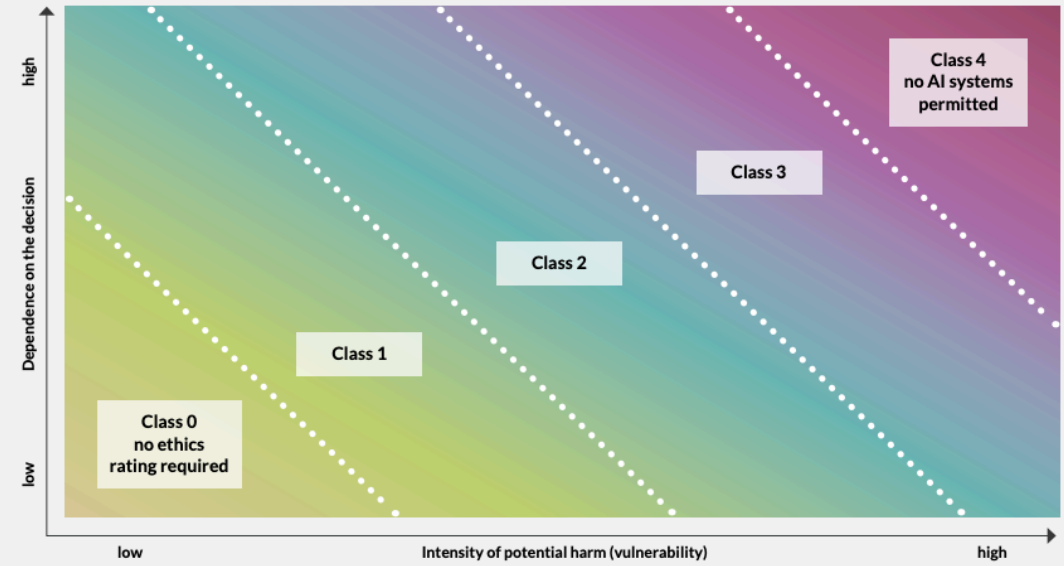


Figure 3

Matrice de risques et interaction avec les autres sections du modèle :

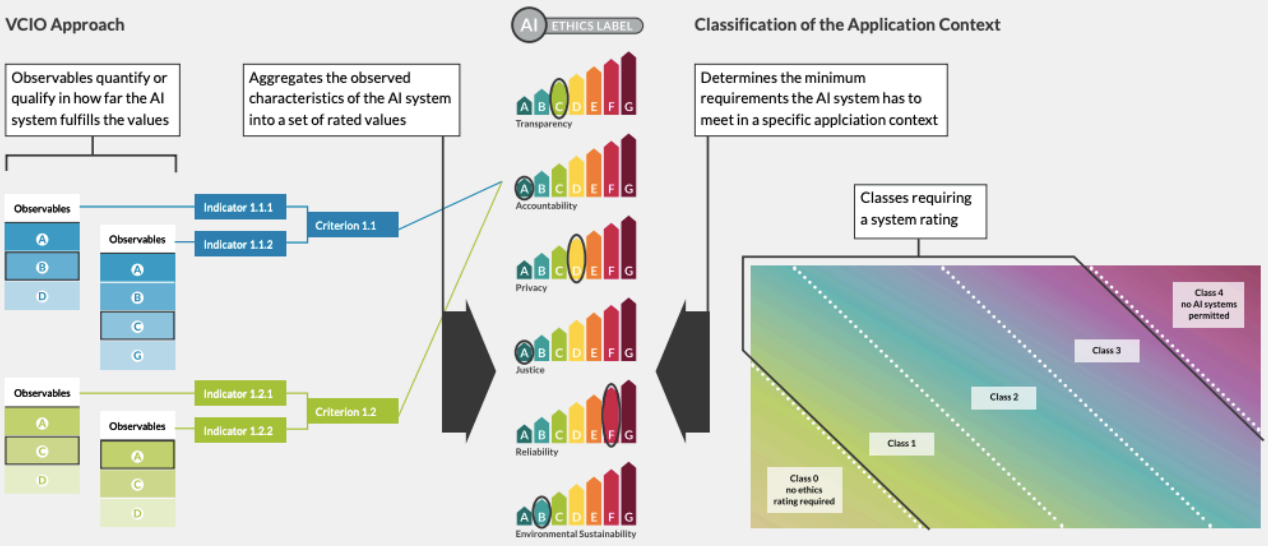
FIGURE 7 **Risk matrix with 5 classes of application areas with risk potential ranging from 'no ethics rating required' in class 0 to the prohibition of AI systems in class 4**



Source: Krafft and Zweig 2019

AIEI Group

FIGURE 8 **Illustration of the composition of the whole system rating using minimum requirements**



AIEI Group

*Consulter les pages 35 à 40 de *AI Ethics Impact Group: From Principles to Practice – VDE* (2020) pour plus de détails.

ANNEXE K – OPEN DATA INSTITUTE, 2019

Extrait de : Open Data Institute. "The Data Ethics Canvas", (2019).[37]

[37] **Une version mise à jour du modèle a été publiée en 2021**.
Voir : https://theodi.org/article/the-data-ethics-canvas-2021/#1563365825519-a247d445-ab2d.

ANNEXE L – CIGREF et SYNTEC NUMÉRIQUE, 2018

Extrait de : Cigref et Syntec numérique. Éthique & numérique : un référentiel pratique pour les acteurs du numérique, aux pp 12-14 (2018).

› RÉFÉRENTIEL ÉTHIQUE ET NUMÉRIQUE

3 questions préliminaires

- ***L'éthique du numérique fait-elle l'objet d'un traitement spécifique dans l'entreprise (y a t-il des comités dédiés, des programmes de sensibilisation portant exclusivement sur l'éthique du numérique, de l'IA etc.) ?***
- ***L'éthique du numérique fait-elle partie des enjeux de gouvernance globale de la transformation numérique ?***
- ***Le sujet de l'éthique du numérique est-il défini clairement et différencié des sujets de conformité / du travail des directions juridiques ?***

ETHIQUE BY DESIGN

Déontologie des développeurs et concepteurs (éditeurs, intégrateurs, designers…) de solutions et services numériques	*La DSI a-t-elle des cycles de formation au sujet de l'éthique dans la conception des outils numériques ?*	Instaurer des ateliers de formation et/ou des stages de mise à niveau au sein de la DSI
	Les concepteurs de solutions sont-ils représentatifs de la diversité sociale et de la mixité ?	Etablir une politique de RH assurant la diversité sociale et la mixité au sein des collaborateurs
Protection de la vie privée et des données personnelles	*Les nouveaux projets sont-ils soumis à des évaluations en matière d'impact sur la vie privée ?*	Mettre en place un comité éthique pour valider les projets sensibles
	Les données personnelles sont-elles protégées dès la conception des outils et solutions ? *Le droit à l'oubli est-il pris en compte dans la chaîne de conception ?*	Adopter une approche ' privacy by design ', conformément aux exigences du RGPD : Il s'agit d'intégrer la protection des données à caractère personnel non seulement dès la conception des produits et services mais également, « par défaut » (notamment avec le principe de minimisation introduit par le RGPD) ; l'enjeu est aussi culturel car il faut intégrer cette notion en amont des projets.
	La corrélation de données issues de diverses sources induit-elle la production d'informations personnelles (dans le cadre de projets big data et d'IA notamment) ?	Mettre en place un dispositif qui mesure la personnalisation des données à la sortie du traitement
Ethique algorithmique et intelligence artificielle		Former les concepteurs aux risques de biais liés aux jeux de données utilisés pour l'apprentissage machine
	Les risques de biais liés aux jeux de données utilisés sont-ils identifiés et traités ?	Mettre en place un DIA (discrimination impact assessment) tel que proposé dans le rapport Villani (p.148), s'inspirant du PIA (privacy impact assessment) présent dans le RGPD, afin d'analyser les possibles impacts discrimants des algorithmes dès leur conception
		Mettre en place des procédures de vérification à chaque étape du développement pour s'assurer de l'absence de biais dans les résultats
	La logique de fonctionnement des algorithmes déployés en intelligence artificielle peut-elle être expliquée ?	Avoir une politique d'explicabilité des systèmes, sur l'ensemble de la chaîne (provenance des données, explication du raisonnement suivi)
		Développer des algorithmes transparents dès la conception, afin de faciliter leur explication et l'analyse de leur raisonnement
		Rentrer dans une démarche de labellisation (via un système de scoring / rating éthique) et d'accompagnement éthique

ETHIQUE D'USAGE (avec les collaborateurs, utilisateurs, partenaires)

Accessibilité des solutions pour les personnes handicapées	*Les outils numériques sont-ils conçus en prenant en compte l'accessibilité pour les personnes en situation de handicap ?*	Concevoir par défaut des solutions accessibles aux personnes en situation de handicap
Accès aux données par les collaborateurs	*Des règles éthiques sur la collecte et le traitement des données sont-elles définies et partagées en interne ?*	Sensibiliser les équipes via des ateliers dédiés à l'éthique des données
	Les droits d'accès internes à des données personnelles et/ou sensibles sont-ils encadrés ?	Définir des modalités d'accès aux données sensibles en fonction des profils et missions des collaborateurs
Ethique managériale	*Les questions liées à l'éthique sont-elles abordées de façon transversale au sein de l'entreprise ?*	Mettre en place un programme de sensibilisation de l'ensemble des collaborateurs (information et exemples de bonnes pratiques) Nommer un Chief Digital Ethics Officer chargé d'assurer la cohérence globale de la politique « éthique et numérique » de l'entreprise
	Les collaborateurs sont-ils informés des conditions de traitement et de conservation de leurs données et de leurs droits afférents ?	Informer les collaborateurs sur les conditions de traitement et de conservation des données les concernant, ainsi que leurs droits afférents (affichage, mise à jour du règlement intérieur de l'entreprise)
	Les conséquences de l'usage de certains outils numériques en interne sont-elles évaluées ?	Réaliser une étude sur l'impact des outils numériques sur le quotidien des collaborateurs au sein de l'entreprise
Éthique avec les utilisateurs	*Des moyens sont-ils proposés aux utilisateurs de services personnalisés pour gérer leurs paramètres ?*	Assurer la clarté et la transparence de l'information fournie aux utilisateurs Donner la possibilité aux utilisateurs de paramétrer facilement la gestion de leurs données personnelles, et d'opérer des choix éclairés
	Les utilisateurs sont-ils informés sur les conditions d'usage d'une solution ou application numérique ?	Définir une charte d'usage numérique précisant les conditions d'usage éthique d'une solution Préciser les cadres d'utilisation prévus d'une solution par voie contractuelle, avec la possibilité pour le concepteur de s'opposer à une utilisation non conforme
Éthique partenariale	*Existe-t-il une politique permettant de vérifier, dans l'assemblage des solutions numériques entre divers partenaires, que le process dans son ensemble est éthique ?*	S'assurer de la loyauté de l'écosystème et donner à chaque partenaire une vision de la finalité de la solution globale Faire appel à des tiers de confiance, certifications et/ou labels, démontrant l'engagement éthique de chaque partie prenante

ETHIQUE SOCIETALE (impact des solutions numériques sur la société)

Empreinte écologique et sociétale des solutions	*Une démarche visant à améliorer l'empreinte environnementale du SI a-t-elle été mise en place ?*	Identifier un responsable et définir un plan d'action incluant la sensibilisation de tous les collaborateurs DSI et utilisateurs, appuyé sur des référentiels reconnus
		Prendre en compte l'impact environnemental lors de toute passation de marché ayant une conséquence sur l'empreinte environnementale du SI
	L'évaluation des impacts environnementaux du SI couvre-t-elle l'énergie primaire, les émissions de gaz à effet de serre, l'eau, l'épuisement des ressources abiotiques, le papier et les DEEE (déchets d'équipements électriques et électroniques) ?	Conduire une évaluation régulière de l'empreinte environnementale du SI (au maximum tous les deux ans), en s'appuyant sur des indicateurs reconnus et auditables (Green IT ou WWF France)
	L'impact sociétal des projets (origine des matériaux, bonnes pratiques des partenaires, etc.) est-il pris en compte ?	Réaliser une étude d'impact sociétal des projets
		Avoir recours à des entreprises de l'Économie Sociale et Solidaire et de l'Économie Adaptée
Impact économique et acceptabilité des innovations	*Une étude est-elle faite sur l'impact des innovations sur l'emploi au sein de l'entreprise, notamment avec l'automatisation ?*	Anticiper, à l'aide d'équipes spécialisées en prospective et stratégie, les impacts des transformations technologiques pour les métiers et activités de l'entreprise
		Inclure les impacts de l'automatisation et plus largement du numérique dans le plan de Gestion Previsionnelle de l'Emploi et des Compétences (GPEC)
Économie de l'attention et bulles informationnelles	*Les phénomènes d'addiction sont-ils pris en compte dans la conception des solutions numériques ?*	Décourager formellement l'utilisation de « dark patterns » (astuces dans le design d'interface destinées à tromper l'utilisateur)
		Se doter de normes de « haute qualité attentionnelle » ou de « labels d'attention responsable » (conférence « Ethics by design » mai 2017)
	Les risques de biais cognitifs humains sont-ils pris en compte dans la conception des solutions numériques ?	S'assurer que les applications et solutions numériques n'ont pas été conçues de manière à manipuler volontairement l'utilisateur par l'exploitation de biais cognitifs

ANNEXE M – AI GLOBAL[38], 2020

Extrait de : AI Global. "Responsible AI Design Assistant", (2020).

Les 5 dimensions d'une IA fiable et des exemples de questions		
Dimensions d'une IA fiable	**Questions**	**Choix de réponses**
1. Accountability	**Has a risk benefit analysis of all aspects of this system including looking at aspects avoidance, mitigation, transference, and acceptance been completed?**	• No • Yes, we have done analysis including, but not limited to, feedback from user surveys, tracking of system performance, short and long-term product health (eg. click-through rate and customer lifetime values), and false positive and false negative rates sliced across different subgroups • Yes, we have ensured that the metrics are appropriate for the context (eg. fire alarm systems should have high recall, even if that means the occasional false alarm)
	To what extent is the review of ethics built in to your organization's practice of implementing responsible programs, processes, and technology?	• Initial, there is limited discussion about different trade-offs of te system within our organization. • Managed, ethical and responsible processes are planned, documented, performed, monitored, and controlled at the project level. Often reactive. • Defined, processes are well characterized and understood. Processes, standards, procedures, tools, etc are defined at the organizational level. Proactive. • Quantitatively managed, processes aer controlled using statistical and other quantitiative techiques. • Optimized, process performance continually improved through incremental and innovative technological improvements.
	Is there a level of specialized knowledge required to operate your system? Are users made aware prior to use?	• No, systems is available to all users with all degrees of knowledge • Yes, users have been clearly notified in advance that there is a degree of knowledge required to operate the system. • Yes, any recipient of the outcomes of this system are made aware that a qualified person is operating the system. • Yes, users have been notified and the appropriate training has been made available to all users. For internal users, training has been delivered or will be delivered prior to deployment • Yes, for internal users, training has been delivered or will be delivered prior to deployment • Yes, users won't use the tool if it's too complicated

[38] **AI Global a changé de nom en 2021 pour *Responsible Artificial Intelligence Institute* (RAII)** et des modifications ont été apportées à l'outil en 2021. Désormais, **6 dimensions éthiques** d'une IA fiable sont examinées : *Organization Maturity, Accountability, Data, Fairness, Interpretability, Robustness*. **Le lien vers la nouvelle version de l'outil est le suivant** : https://designassistant.responsible.ai/.

	At what stage of development is your organization's risk management process?	• Does not exist, risk assessment for processes for technology and business decisions do not currently occur. • Initial, organization considers risks in an ad hoc manner without following defined processes or polices. • Repeatable but intuative, There is an emerging understanding that risks are important and need to be considered. Some approaches to risk assessment exist, but the process is still immature and under development. • Defined process, an organization-wide risk management policy defines when and how to conduct risk assessments. Risk assessment follows a defined process that is documented and available to all staff through training. • Managed and measurable, the assessment of risk is a standard procedure and exceptions to following the procedure would be noticed by management. It is likely that risk management is a defined management function with senior level responsiblity. Senior management have determined the levels of risk that the organiation will tolerate and have standards measures for risk/ return ratios.
	What type of technology are you using? Select all that apply.	• Image and object recognition: analyzing datasests to automate the recognition, classification, reinforcement, and context associated with an image or object. • Text and speech analysis: analyzing datasets to recognize, process, and tag text, speech, voice, and make recommendations based on the tagging. • Risk assessment: Analyzing datasets to identify patterns and and recommend courses of action and in some cases trigger specific actions. • Content generation: Analyzing dataset to categorize, process, triage, personalize, and serve specific content for specific contexts. • Process optimization and workflow automation: Analyzing datasets to identify anomalies, cluster patterns, predict outcomes or ways to optimize and automate specific workflows. • Other
	Is there ongoing monitoring of the system to ensure that the system is operating as intended?	• Yes • No
	What type(s) of unintended outcomes could occur from the use of this system?	Material, Moral, Physical, None
	Were all labour laws or procedures in your jurisdiction followed throughout the development life-cycle of this project?	• Yes • No

	Is there a robust review of the inputs of your system, including but not limited to the data and algorithms used to train and operate the system?	Yes, inputs are reviewed internally as part of a general technology review Yes, inputs are reviewed internally independent of a technology review and include at least one of the following activities: • Analysis of the systems training/testing data • Multiple tests with different training/testing data, for example, gather more training data for certain subsets (e.g., parts/slices of the feature space), or gather test data for categories of interest • Testing of different models • Testing for right degree of transparency to mitigate any potential abuse • System is built so users can set their own degree of sensitivity • The smallest set of inputs necessary were used to make it clear which factors are affeting the model • The simplest model to meet the performance goal was used • Model learns on casual relationships not correlations when and where possible • Model is trained to match true goal • Model has been constrained to produce input-output relationships that reflect domain expert knowledge • Metrics used address the particular benefits and risks of your specific context • The model's sensitivity has been trained for different subsets of examples Yes, inputs are reviewed by a trusted third-party independent of a technology review No third third-party review is completed
	Is the personal data information flow in the system compliant with EU GDPR?	• • Yes • No • Not applicable
	Is the personal data information flow in the system compliant with California Privacy Protection?	
	Is there a log which determines who has generated and had access to the inputs of the system?	• No • Yes, the log identifies the authority or delegated authority responsible • Yes, the log records all the recommendations or decisions made by the system and are easy to identify
	Is there a contingency plan in place if the system is not available?	• No • Yes, there is a contigency plan in place if the system is not available
	Is there a process in place for determining if the automated activity will be flagged for human oversight?	▪ System is only used to assist a decision maker ▪ System is replacing a decision that would otherwise be made by a human and no judgement or discretion is required ▪ System is replacing a decision that would otherwise be made by a human and judgement or discretion is required

	Does your organization have a framework, standard, or set of controls in place to evaluate responsible AI development and deployment in your organization?	• Yes, a framework based on key principles of the organization • Yes, evaluatable standards or controls • No, we leverage a measurable framework from a credible institution • No, we haven't developed this yet
2. Bias and Fairness	**Is your user base comprised of individuals or groups from vulnerable populations?**	• Yes - Most users will be individuals or groups from vulnerable populations • Yes - Some users will be individuals or groups from vulnerable populations • No - There are currently no identified vulnerable populations in the user base
	What is the likelihood of an unintended outcome occuring during the operation of your system?	• Zero Likelihood - All outcomes of the system are determined in advance, there are no learning components to the system • Low Likelihood - It is possible that an unexpected event could occur, but the likliehood is low • High Likelihood - There are many unknown variables within the system, including how the system learns, as a result, there is a high probability that something unlikely could happen during the operation of the system.
	If an unintended outcome(s) were to occur, who would it impact?	• All users equally • A segment of the population • Your organization • A different organization
	What is the potential severity of the effect if one or more of these unintended outcome(s) were to occur?	• High / Medium / Low
	Does your organization have a review model in place including looking at aspects of diversity and complete representation to ensure alternative perspectives or viewpoints are taken into account in advance of the system operating?	• Governance board includes diverse and complete representation including members who represent each area of the organization as well as those with legal and financial responsibilities. • Governance board includes a minimum of one individual with reasonable experience and knoweldge in ethics. • There is a mechanism and review process for items raised by individuals or groups working on the project to present information such as potential issues, including but not limited to, risks (eg. biases, maturity of process, lack of fairness, etc) • There is a mechanism and review process for credible third parties to review project and comment on any issues, including but not limited, potential risks of the project (eg. biases, maturity of process, lack of fairness, etc) • If a third party (eg. government body, civil society organization, etc) or member of the public is aware of the project, there is a mechanism and review process to ensure their question is addressed in a timely manner. • No review model

	How does the system ensure that rights, values, and principles of the public have been protected through the data collection process?	• The system is equally available to all segments of the potential user base • The user is informed on the potential risks on human mental integrity (eg. nudging) when this product is being used. • Useful information about the design, testing, and development of the system (including, but not limited to, decision criteria, inputs for training the model) are provided in a clear and easy to access manner to the user. • Notification is provided in a clear and easy to understand way that there was a machine involved in the creation of a decision, content, advice, outcome, or action. • Results of ongoing testing, training, and monitoring of the system are open and available.
	Are the objectives of the system clear to the users?	▪ Yes, objectives are clear ▪ Yes, clear documentation about these objectives have been provided to the intended user ▪ No
	In which of the following areas has the algorithm been designed or trained in a way to avoid the creation, reinforcement, or reproduction of bias:	• Socio-economic capacity • Physical attributes • Level of education • Sexual orientation • Ethnicity • Cultural association(s) • Religious affiliation(s) • Mental health • Gender • Age • Degree of ability • other
	How is the privacy and intimacy of an individual or group protected in both the development and implementation of your system?	▪ The collection of private information only takes place if a user consents or for authorized surveillance purposes. ▪ Data related to personal thoughts and emotions are not used in situations where the system could case harm, especially in circumstances where moral judgements (eg. lifestyle choices) could be made. ▪ Users are able to disconnect or stop sharing information with the system at any point in time. ▪ Only users have the ability to set profile preferences, changes to the these preferences can easily be done at any time. ▪ Access to personal information is limited to only individuals who require it for the direct functioning of the system. ▪ Individuals have the ability to access their personal data including, but not limited to, the collection, use, and sharing of this data at any time. ▪ Individuals have the ability to donate their personal data to research organizations. ▪ Data integrity is assured. The system does not use private data to imitate or alter a persons appearance, voice, or other

		individual characteristics in order to damage one's reputation or manipulate other people. ▪ The system does not curtail people's real or percieved liberty.
	If your system have learning capabilities, what protections are in place for safeguarding user privacy?	• Not applicable • Testing has been completed to ensure that the model is not memorizing or exposing sensitive data • Use of data is minimized while maintaining
	Was training for diversity and inclusion completed by all individuals working on the design and development of this system?	• Yes • Yes, only those who worked on the design and input aspects • No
	Are psychological, behavioral, geographical or any other societal inferences used for targeting (or other predictions)?	• No • Yes, however we have engaged with social scientists, humanists, and other relevant experts to understand and account for various perspectives, including, but not limited to, how the system will function over time. • Yes, however, goals have been set to ensure the system works fairly across anticipated use cases, including across different regions, with different languages, with different segmentations
	Can the deployment of this system restrict an individual from access to a specific business product or service offering?	• Yes • No
	Will the user have the ability to challenge decisions/recommendations made by the system?	
	Does this system impact an end-user or consumer's economic interest, health, access or mobility, licensing and permit issuance, or otherwise impact their lives?	
	Does training data rely on decisions/outcomes previously made by individuals to influence the outcomes? Was this data reviewed for bias?	
	The decision or action made by the system has the potential to adversely impact:	• Not applicable • Health and well-being of an individual or group • Economic interests of an individual or group • Ongoing sustainability of an environmental ecosystem
	Are the objectives of the system clear to the operators?	• Yes • No

	All datasets being used for this system or model throughout each phase of the lifecycle have been reviewed for currentness/ timeliness following industry best practices (equivalent to ISO 25012)?	• Data was not reviewed for currentness/ timeliness • Dataset is relevant given the time and context it is being used in • The decay of the dataset is understood by system developer and operator • The lifespan of the dataset has been documented clearly in the corresponding metadata
3. Explainability and Fairness	**Is it possible to discover how your system renders a decision or performs a function?**	• System is transparent, it is possible to know for certain how and why the system made a particular decision, or in the case of a robot, acted the way it did. • System is opaque, it is possible through post-decision or post-action, to analyze through various processes (eg. counterfactual or repeatability testing) to draw an accurate conculsion on how the decision was made or the action was taken. • System is a black box, but detailed records of the design processes and decision making have been kept throughout the entire process. • System is a black box, it is possible to make best guesses on why a decision was rendered or a decision was taken, but it's not certain.
	With which of the following groups are you willing to share information about how your system renders a decision or performs a function in an easy to understand way? Information should include, but not be limited to, data sources, development processess, and consulted stakeholders.	• The public • Users who make a special request • Certification bodies for verification of the system • Authorities, if required • None of the above
	Are the terms of reference easy to understand by the intended audience?	• Yes • Not sure • No
	What information is provided about how the system operates?	• Simple counterfactual explanations are provided to users • Easy to understand visuals • Access to support team for conversational explanations • Not available
	Are the terms and conditions of the system easy to understand?	• Terms and conditions for the system use plain language and are easy to understand • Terms and conditions take less than thirty minutes to read and understand by target audience • Terms and conditions are comprehensive and as a result difficult to turn into plain language and take more than thirty minutes to read
	To what extent is it possible for the operator of the system to access the training data? Select all that apply.	• Full access to data, including automated decisioning data, manual changes to data, manual processing of the decision, etc. • Access to all automatically collected data is available

		• Access to all key decision parameters • Some gaps in special cases of the decision (think manual workflows that fall off of the automated workflow where somebody makes a change or makes a one-off decision/edit to the data) • Access to data is limited • No access to data
	Is it possible to retrain your model and compare results when training data or other factors are changed?	▪ Yes ▪ No
	Can you gather test data for all categories of interest?	
	Can you design a new model or are you constrained to an already trained model?	• Yes, a new model can be trained • No, the existing model must be used, but can be augmented • No, the existing model must be used and can't be augmented
	Would it be possible for an independent user to change the vectors (parameters) of the model so it is more accurate for their context?	• Yes • No • Not applicable
	To what extent was user testing completed to ensure the goals and objectives of the users are being met? **Select all that apply.**	• User experience testing has been done since the start of the project • User experience testing will be done at every major interval of the project • User experience testing was done before the launch of the project • Users are able to test how different inputs affect the model output • Users are able to provide feedback on their representation • No user experience testing was done
4. Robustness	**What type(s) of testing have been conducted to ensure quality development of the system?** Select all that have been completed.	• None yet • Rigorous unit tests to test each component of the system in isolation. • Integration tests to understand how individual ML components interact with other parts of the overall system. • Proactive detection to determine drift by testing the statistics of the inputs to the AI system to ensure they are not changing in unexpected ways. • A diverse dataset covering most states of the input variables with many individual examples was used to test the system and ensure that it continues to behave as expected.

		• Edge cases and extreme scenarios have been studied and experimented with to ensure acceptable system performance under all intended use cases of the system • Iterative user testing to incorporate a diverse set of user needs in the development cycles. • Quality engineering principles like poka-yoke have been used to build quality checks into a system so that unintended failures either cannot happen or trigger an immediate response.
	Is the entire supply chain secure?	• Yes • Not sure • No
	Is there tracking of system performance in place, are there established criteria where the system is considered not fit for purpose (eg the usage is outside of what is considered appropriate for the system to provide a decision/ outcome)?	• Yes • No
	What safeguards have you put in place to ensure your system is robust enough to deal with the edge cases and extreme scenarios (eg. load inputs, adversarial attacks) to adequately mitigate erroneous outcomes to the best extent possible?	• None • Have ensured that there is a mitigation plan in place for any individual, group, or organization who has an incentive to make the system misbehave. • The unintended consequences resulting in a mistake from the system have been assessed and mitigated to the best extent possible. • A rigorous threat model to understand all possible attack vectors has been implemented. • The system has been tested against adversarial attacks. • Third party advesarial testing of the system was completed. • Ongoing research is being conducted to ensure the latest tools are being applied. • Preventative and precautionary measures have been taken.
	What contingencies are incorporated in the model and or system? **Select all that apply.**	• • Established procedures for certain functions are clearly outlined and understood • There are established procedures for each type of recognized exception (eg denial of service, manual processing, etc.) • Regular review of system failures occuring are considered and added to subsequent documentation • Other
5. Data Quality	**How was the data being used for the system collected?**	• • Collected by same organization as development of the system • Purchased by a dependable third-party • Open dataset with known owner • Open dataset with unknown origin • Third-party aquisition with an unknown supply chain • Unclear

	Does data used for training or use of the system include personal information?	• No • Yes : age • Yes, name • Yes, address • Yes, gender • Yes, ethnicity • Yes, status • Yes, degree of accessibility • Yes, level of education • Yes, income • Yes, marital status • Other
	Does the system require the use of sensitive data for development or implementation?	• Yes • The use of sensitive data is minimized • No, the use of sensitive data is necessary, however, measures are in place to mitigate the use and sharing of sensitive data • No, the use of sensitive data is necessary for the operation of the system
	What steps are followed to ensure data used in training or implementation of this system is handled with care?	• Standards for cryptography or security are followed. • Data is encrypted when in transit or rest. • Anonymize and aggregate incoming data using best-practice data scrubbing pipelines. • Provides users with clear notification when personal information is being collected, used, or shared. • When and where possible, system leverages on-device processing. • Where possible, independent data trusts or data collaboratives are used to ensure users privacy.
	Has training and implementation data been reviewed for quality? **Select all that apply.**	• No • Yes, data has been reviewed by internal team • Yes, data has been reviewed by a qualified third-party • Yes, data is compliant with ISO 8000 or equivalent industry recognized standard
	Has your system been trained on data that accurately represents your entire user base?	• No • Yes, sample data used for training and testing accurately reflects user base including, but not limited to, age, location, gender, ethnicity, status, degree of accessibility, level of education, income. • Yes, training data has been ensured that it is fit for purpose. • Yes, sample data is accurate for current point in time.

		• Yes, sample data has been tested for accuracy • Yes, sample data is complete.
	How is the data being collected, used, and stored being managed?	• Through a third-party data service, terms and conditions are unknown. • Through a third-party data service, terms and conditions ensure users data should be protected from theft, misuse, or data corruption
	What are the recourse mechanisms for a decision or action that does not meet the objectives of the system?	• Decision is able to be reversed • Update will be made to the process • If an issue with the system has been identified that could cause harm to an individual or community, it is immediatly taken off-line until it is remediated • Users are notified • None
	Do you have clear policies around storing and handling testing data to evaluate model performance such that there is no data leakage to intentionally or unintentionally influence results from model performance and the evaluation of the overall model?	• Yes • No
	not group	• Checks to the underlying population • Analysis of new use cases • Analysis of exceptional cases • None • Other
	How was the creation of the dataset funded?	• All data was collected and created by internal researchers • • It was funded in part or whole by a third party organization including a government or academic institution • It was funded in part or whole by a third party organization including a research institute, company, or non-governmental organization • Not applicable
	If data was collected within your organization (or the collection methodology is known), which of the following best practices were followed during data collection? Select all that apply.	• Data was directly observable (eg. raw text, movie ratings) • Reported directly by subjects (survey response) • • Indirectly inferred/ derived from other data (eg. part-of-speech tags, model-based guesses for age or language) • Data that was reported by subjects or indirectly inferred/ derived from other data that was validated and/ or verified. • If data was automatically collected the mechanisms and procedures were validated • Data collection practices are unknown

	What is the complexity of the dataset?	• The dataset is simple, there are a limited number of instances (eg. documents, photos, people, countries) and types of instances (eg. movies, users, and ratings; people and interactions between them; nodes and edges) that comprise the full dataset. • The dataset is complex, there are a large number of instances (eg. documents, photos, people, countries) and types of instances (eg. movies, users, and ratings; people and interactions between them; nodes and edges) that comprise the full dataset.
	Is your system trained and tested on data that accurately represents the use of the system? That is, is the data fit for purpose? Select all that apply	• Yes, sample data used for training and testing accurately represents the use of the system • • Data has been sampled from a related dataset • Data being used to train and test the system is aligned, however, some components of the dataset needed to be synthesized to augment for the intended purpose. • Not applicable
	Is the data being used to train the system "raw" or has it been processed?	• Data is "raw" no further processing has been done from direct collection. • Data has been processed from it's "raw" state to be interpreted for human consumption, data is representative of "raw" state. • Data has been processed from it's "raw" state for other purposes.
	How are data being produced by your system or model being re-used or re-interpreted?	• The data is used to recalibrate and retrain the same system or model • Data are not re-used or re-interpreted • The data from this model or system is collected and made available for reuse within the same organization • The data from this model or system is collected and made available for reuse outside of the organization
	What is the lifecycle of the dataset being used for this system or model?	• Data being used for training, operating, and maintaining the system or model is disposed of after it is no longer needed for operations • Data being used for training, operating, and maintaining the system or model is retained after operations as it may be used for different models or systems
	All datasets being used for this system or model throughout each phase of the lifecycle have been reviewed for completeness following industry best practices (equivalent to ISO 8000 Part 140 and ISO 25012)? Select all that apply.	• Data was not reviewed for completeness • All values in the dataset have a specific context or use • There are no missing values that are pertinent for the intended use of the dataset • Comprehensive metadata including pertinent collection information including who collected the data, ownership, the description of the dataset, the provenance, etc. • There is a data dictionary to support and help interpret the dataset

		• Data attributes within dataset conform to a data specification
	All datasets being used for this system or model throughout each phase of the lifecycle have been reviewed for accuracy following industry best practices (equivalent to ISO 8000 Part 130 and ISO 25012)?	• Data was not reviewed for accuracy • All values within the dataset represent the true intent of a sepcific event, context, or use • Data has been validated against applicable reference data where possible • Dataset contains applicable labels that are representative of the data attributes contained in the dataset • Data was collected in a way that accurately represents the intended purpose, population, event, etc
	All datasets being used for this system or model throughout each phase of the lifecycle have been reviewed for credibility/ provenance following industry best practices (equivalent to ISO 8000 Part 130 and ISO 25012)? **Select all that apply.**	• Data was not reviewed for credibility/ provenance • All values included in the dataset are true, accurate, and or authentic • • The history of the data is known and clear • Owner of data has been identified and can be contacted if needed
	All datasets being used for this system or model throughout each phase of the lifecycle have been reviewed for consistency following industry best practices (equivalent to ISO 25012)?	• Data was not reviewed for consistency • Dataset values are free from contradition and are coherent with other attributes within the dataset • Dataset values are free from contradiction and are coherent with other related datasets both produced by your organization or another credible organization
	All datasets being used for this system or model throughout each phase of the lifecycle have been reviewed for confidentiality following industry best practices (equivalent to ISO 25012)? **Select all that apply.**	• Data was not reviewed for confidentiality • Data is only accessible and interpretable by authorized users in a specific context of use. • Data that could be deemed offensive is either removed during use, or an appropriate warning is provided? • • Data is anonymized to the greatest extent possible to ensure that no individual directly or indirectly can be identified
	All datasets being used for this system or model throughout each phase of the lifecycle have been reviewed for traceability following industry best practices (equivalent to ISO 25012)? Select all that apply.	• Data was not reviewed for traceability • Dataset includes an audit trail of access to the data including any changes made • Dataset values contain unique identifiers where needed

	All datasets being used for this system or model throughout each phase of the lifecycle have been reviewed for compliance following industry best practices (equivalent to ISO 25012)? Select all that apply.	• Data was not reviewed for compliance • Data complies with all applicable standards • Data complies with all applicable conventions • Data complies with all applicable legal rules (laws, regulations, etc)
	All datasets being used for this system or model throughout each phase of the lifecycle have been reviewed for accessibility including portability and recoverability following industry best practices (equivalent to ISO 25012)? select all that apply:	• Data was not reviewed for accessibility • Data can be accessed by users or operators as needed • • Data is able to be installed, replaced, or moved from one system to another preserving the existing quality in a specific context or use. • Data is adaptable and cab be swapped out by another supplier or newer revisions/ updates • Data is able to be preserved even in the event of a failure

ANNEXE N – UTRECHT DATA SCHOOL et UNIVERSITÉ D'UTRECHT, 2020

Extrait de : Utrecht Data School et Université d'Utrecht. "Data Ethics Decision Aid (DEDA)", (2020).

Section 1 : Data related considerations		
Sections	**Thèmes**	**Questions associées**
Collecting	**Algorithms**	• Does the project make use of an algorithm, or some form of machine learning or neural network? If not, go to "Source". • Is there someone within the team that can explain how the algorithm in question works? • Is there someone who can provide an explanation that is accessible to the wider public?
	Source	• Where do the data(sets) come from? • In what ways have you checked the quality of the data? • Do the data have an expiration date?
Using	**Anonymization**	• Should the data be anonymized, pseudonymized or generalized? • Who has access to the encryption key to de-pseudonymize the data?
	Visualization	• How will the results of the project be presented? • Are the results suitable for visualization?
Storing	**Access**	• Who has access to the data and under what conditions? • How is the access monitored?
	Sharing, reusing and repurposing	• Are any of the data suitable for reuse? If so, under what conditions and for what (new) purpose(s) could they be reused? • Are there any obligations (not) to make the data publicly available? If you were to provide open access to parts of the data, what opportunities and risks might arise?

Section 2 : General considerations	
Thèmes	**Questions associées**
Responsibility	• Which laws and regulations apply to your project? • Who is ultimately responsible for the project? • Are the duties and responsibilities of that person clear, with regard to this project? • Is the project suitable for cooperation with (commercial) partners? If so, which parties could that be?
Communication	• What is the communication strategy with regard to this project? Are all parties involved in agreement as to this strategy? • What communication strategies are there for cases in which something goes wrong, and who is responsible for them?
Transparency	• Does the project risk generating public concern or outrage? • How transparent are you about this project towards citizens? • Do citizens have the opportunity to raise objections to the results of the project? • Can citizens opt out of their involvement in the project? If so, when and how can they do this?
Privacy	• Is there a data protection officer or data privacy officer involved in this project?

	• Have you conducted a PIA (Privacy Impact Assessment) or DPIA (Data Protection Impact Assessment)? • Does this project make use of personal data? If not, continue to "Bias". • Do the data provide insight into the personal lives of citizens?
Bias	• As a member of the project, what outcome do you expect? • Is there anything about this project that makes you uneasy? If so, discuss these concerns. • Will the results of the analysis be evaluated by a human before being implemented? • Is there a risk that your project could contribute to discrimination against certain people or groups? • Are all relevant citizens adequately represented within your data(sets)? Which ones are missing or under-represented? • Is there a *feedback loop* in the model that might have negative consequences? • Are you gathering the information that is appropriate for the purposes of your project? • Is there a risk that the project will unintentionally create incentives for undesirable behaviour? • *Function creep*: can you imagine a future scenario in which the results of your project could be mis(used) for alternative purposes? • Do your answers change when you consider possible long-term effects? Why?

ANNEXE O – ARNOLD et al, 2019

Extrait de Arnold et al, "FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity", (2019).

To illustrate how these questions could be answered, we provide two examples for fictitious AI services: a fingerprint verification service (Appendix B) and a trending topics service (Appendix C). However, given that the examples we provide are fictitious, we would expect an actual service provider to answer these questions in much more detail. For instance, they would be able to better characterize an API that actually exists. Our example answers are mainly to provide additional insight about the type of information we would find in a FactSheet.

Statement of purpose

The following questions are aimed at providing an overview of the service provider and of the intended uses for the service. Valid answers include "N/A" (not applicable) and "Proprietary" (cannot be publicly disclosed, usually for competitive reasons).

General

- Who are "you" (the supplier) and what type of services do you typically offer (beyond this particular service)?
- What is this service about?
  - Briefly describe the service.
  - When was the service first released? When was the last release?
  - Who is the target user?
- Describe the outputs of the service.
- What algorithms or techniques does this service implement?
  - Provide links to technical papers.
- What are the characteristics of the development team?
  - Do the teams charged with developing and maintaining this service reflect a diversity of opinions, backgrounds, and thought?
- Have you updated this FactSheet before?
  - When and how often?
  - What sections have changed?
  - Is the FactSheet updated every time the service is retrained or updated?

Usage

- What is the intended use of the service output?
  - Briefly describe a simple use-case.
- What are the key procedures followed while using the service?
  - How is the input provided? By whom?
  - How is the output returned?

Domains and applications

- What are the domains and applications the service was tested on or used for?
  - Were domain experts involved in the development, testing, and deployment? Please elaborate.
- How is the service being used by your customers or users?
  - Are you enabling others to build a solution by providing a cloud service or is your application end-user facing?
  - Is the service output used as-is, is it fed directly into another tool or actuator, or is there human input/oversight before use?
  - Do users rely on pre-trained/canned models or can they train their own models?
  - Do your customers typically use your service in a time critical setup (e.g. they have limited time to evaluate the output)? Or do they incorporate it in a slower decision making process? Please elaborate.
- List applications that the service has been used for in the past.

- Please provide information about these applications or relevant pointers.
- Please provide key performance results for those applications.

- Other comments?

Basic Performance

The following questions aim to offer an overall assessment of the service performance.

Testing by service provider

- Which datasets was the service tested on? (e.g., links to datasets that were used for testing, along with corresponding datasheets)

  - List the test datasets and provide links to these datasets.
  - Do the datasets have an associated datasheet? If yes, please attach.
  - Could these datasets be used for independent testing of the service? Did the data need to be changed or sampled before use?

- Describe the testing methodology.

  - Please provide details on train, test and holdout data.
  - What performance metrics were used? (e.g. accuracy, error rates, AUC, precision/recall)
  - Please briefly justify the choice of metrics.

- Describe the test results.

  - Were latency, throughput, and availability measured?
  - If yes, briefly include those metrics as well.

Testing by third parties

- Is there a way to verify the performance metrics (e.g., via a service API)?

  - Briefly describe how a third party could independently verify the performance of the service.
  - Are there benchmarks publicly available and adequate for testing the service.

- In addition to the service provider, was this service tested by any third party?

  - Please list all third parties that performed the testing.
  - Also, please include information about the tests and test results.

- Other comments?

Safety

The following questions aim to offer insights about potential unintentional harms, and mitigation efforts to eliminate or minimize those harms.

General

- Are you aware of possible examples of bias, ethical issues, or other safety risks as a result of using the service?

  - Were the possible sources of bias or unfairness analyzed?
  - Where do they arise from: the data? the particular techniques being implemented? other sources?
  - Is there any mechanism for redress if individuals are negatively affected?

- Do you use data from or make inferences about individuals or groups of individuals. Have you obtained their consent?

- How was it decided whose data to use or about whom to make inferences?
- Do these individuals know that their data is being used or that inferences are being made about them? What were they told? When were they made aware? What kind of consent was needed from them? What were the procedures for gathering consent? Please attach the consent form to this declaration.
- What are the potential risks to these individuals or groups? Might the service output interfere with individual rights? How are these risks being handled or minimized?
- What trade-offs were made between the rights of these individuals and business interests?
- Do they have the option to withdraw their data? Can they opt out from inferences being made about them? What is the withdrawal procedure?

Explainability

- Are the service outputs explainable and/or interpretable?

  - Please explain how explainability is achieved (e.g. directly explainable algorithm, local explainability, explanations via examples).
  - Who is the target user of the explanation (ML expert, domain expert, general consumer, etc.)
  - Please describe any human validation of the explainability of the algorithms

Fairness

- For each dataset used by the service: Was the dataset checked for bias? What efforts were made to ensure that it is fair and representative?

  - Please describe the data bias policies that were checked (such as with respect to known protected attributes), bias checking methods, and results (e.g., disparate error rates across different groups).
  - Was there any bias remediation performed on this dataset? Please provide details about the value of any bias estimates before and after it.
  - What techniques were used to perform the remediation? Please provide links to relevant technical papers.
  - How did the value of other performance metrics change as a result?

- Does the service implement and perform any bias detection and remediation?

  - Please describe model bias policies that were checked, bias checking methods, and results (e.g., disparate error rates across different groups).
  - What procedures were used to perform the remediation? Please provide links or references to corresponding technical papers.
  - Please provide details about the value of any bias estimates before and after such remediation.
  - How did the value of other performance metrics change as a result?

Concept Drift

- What is the expected performance on unseen data or data with different distributions?

  - Please describe any relevant testing done along with test results.

- Does your system make updates to its behavior based on newly ingested data?

- Is the new data uploaded by your users? Is it generated by an automated process? Are the patterns in the data largely static or do they change over time?
- Are there any performance guarantees/bounds?
- Does the service have an automatic feedback/retraining loop, or is there a human in the loop?

- How is the service tested and monitored for model or performance drift over time?
  - If applicable, describe any relevant testing along with test results.
- How can the service be checked for correct, expected output when new data is added?
- Does the service allow for checking for differences between training and usage data?
  - Does it deploy mechanisms to alert the user of the difference?
- Do you test the service periodically?
  - Does the testing includes bias or fairness related aspects?
  - How has the value of the tested metrics evolved over time?
- Other comments?

Security

The following questions aim to assess the susceptibility to deliberate harms such as attacks by adversaries.

- How could this service be attacked or abused? Please describe.
- List applications or scenarios for which the service is not suitable.
  - Describe specific concerns and sensitive use cases.
  - Are there any procedures in place to ensure that the service will not be used for these applications?
- How are you securing user or usage data?
  - Is usage data from service operations retained and stored?
  - How is the data being stored? For how long is the data stored?
  - Is user or usage data being shared outside the service? Who has access to the data?
- Was the service checked for robustness against adversarial attacks?
  - Describe robustness policies that were checked, the type of attacks considered, checking methods, and results.
- What is the plan to handle any potential security breaches?
  - Describe any protocol that is in place.
- Other comments?

Lineage

The following questions aim to overview how the service provider keeps track of details that might be required in the event of an audit by a third party, such as in the case of harm or suspicion of harm.

Training Data

- Does the service provide an as-is/canned model? Which datasets was the service trained on?

- List the training datasets.
- Where there any quality assurance processes employed while the data was collected or before use?
- Were the datasets used for training built-for-purpose or were they repurposed/adapted? Were the datasets created specifically for the purpose of training the models offered by this service?

- For each dataset: Are the training datasets publicly available?
  - Please provide a link to the datasets or the source of the datasets.
- For each dataset: Does the dataset have a datasheet or data statement?
  - If available, attach the datasheet; otherwise, provide answers to questions from the datasheet as appropriate [to insert citation]
- Did the service require any transformation of the data in addition to those provided in the datasheet?
- Do you use synthetic data?
  - When? How was it created?
  - Briefly describe its properties and the creation procedure.

Trained Models

- How were the models trained?
  - Please provide specific details (e.g., hyperparameters).
- When were the models last updated?
  - How much did the performance change with each update?
  - How often are the models retrained or updated?
- Did you use any prior knowledge or re-weight the data in any way before training?
- Other comments?

ANNEXE P – FLORIDI et al, 2020

Extrait de : Floridi et al, "How to Design AI for Social Good: Seven Essential Factors", à la p. 20, (2020).

Table 1 Summary of seven factors supporting AI4SG and the corresponding best practices

Factors	Corresponding best practices	Corresponding ethical principle
Falsifiability and incremental deployment	Identify falsifiable requirements and test them in incremental steps from the lab to the "outside world"	Nonmaleficence
Safeguards against the manipulation of predictors	Adopt safeguards which (i) ensure that non-causal indicators do not inappropriately skew interventions, and (ii) limit, when appropriate, knowledge of how inputs affect outputs from AI4SG systems, to prevent manipulation	Nonmaleficence
Receiver-contextualised intervention	Build decision-making systems in consultation with users interacting with and impacted by these systems; with understanding of users' characteristics, the methods of coordination, the purposes and effects of an intervention; and with respect for users' right to ignore or modify interventions	Autonomy
Receiver-contextualised explanation and transparent purposes	Choose a Level of Abstraction for AI explanation that fulfils the desired explanatory purpose and is appropriate to the system and the receivers; then deploy arguments that are rationally and suitably persuasive for the receiver to deliver the explanation; and ensure that the goal (the system's purpose) for which an AI4SG system is developed and deployed is knowable to receivers of its outputs by default	Explicability
Privacy protection and data subject consent	Respect the threshold of consent established for the processing of datasets of personal data	Nonmaleficence; autonomy
Situational fairness	Remove from relevant datasets variables and proxies that are irrelevant to an outcome, except when their inclusion supports inclusivity, safety, or other ethical imperatives	Justice
Human-friendly semanticisation	Do not hinder the ability for people to semanticise (that is, to give meaning to, and make sense of) something	Autonomy

***

ANNEXES DE LA SECTION B

ANNEXE Q – Matrice de formalisation retenue – Grille analytique (Yves Boisvert, 2021)

Référentiel (Principes ou valeurs éthiques)	***A***	***B***	***C***	***D***	***…***
Les enjeux, risques ou dilemmes éthiques (propres à un principe ou à une valeur éthique)					
Les meilleures pratiques (pour gérer ces enjeux, risques ou dilemmes)					
Les entraves (qui empêchent d'adopter les meilleures pratiques)					
Les stratégies (qui permettent de dénouer les impasses créées par les entraves)					

ANNEXE R – Chartes éthiques encadrant les systèmes d'intelligence artificielle : *informations complémentaires*

Charte éthique	Contexte, auteurs	Autres remarques
Assessment List for Trustworthy Artificial Intelligence (ALTAI), Commission européenne (2020)	Liste d'évaluation finale pour une IA digne de confiance publiée à la suite de la publication des *Lignes directrices en matière d'éthique pour une IA digne de confiance par* le Groupe d'experts indépendants de haut niveau de la CE en 2019.	**Tensions entre les différentes exigences :** - Identifier les intérêts et les valeurs concernés par le SIA; - En cas de conflit, explicitement reconnaitre et évaluer en termes de risque pour la sécurité et les principes éthiques et les droits fondamentaux; - Motiver et documenter toute décision concernant le choix d'un compromis; - Prévoir des mécanismes accessibles pour assurer une réparation adéquate en cas d'impact négatif. **Droits fondamentaux** Grande place accordée à la protection des droits fondamentaux.
Recommandation du Conseil sur l'intelligence artificielle, OCDE (2019)	Recommandation constituant la 1re norme intergouvernementale sur l'IA qui complète les recommandations de l'OCDE en d'autres domaines (vie privée, sécurité numérique, etc.) et qui a pour but de définir une norme de mise en œuvre suffisamment flexible. Processus inclusif et participatif pour l'élaboration de la Recommandation : Groupe d'experts sur l'IA de l'OCDE composé de plus de 50 experts de disciplines et de secteurs différents + collaboration avec d'autres organes de l'OCDE.	**Principes éthiques** 5 principes complémentaires fondés sur des valeurs qui doivent être considérés comme un tout. Accent mis sur le respect des droits de l'homme et des valeurs démocratiques.
ALLEMAGNE		
From Principles to Practice : An interdisciplinary Framework to operationalise AI Ethics, AIEI Group, (2020)	Rapport préparé par des experts de différents domaines (informatique, philosophie, physique, ingénierie, sciences sociales) du AI Ethics Impact Group – consortium interdisciplinaire dirigé par VDE Association for Electrical, Electronic & Information Technologies et Bertelsmann Stiftung.	**Conflit entre des valeurs** Les auteurs proposent une méthode de hiérarchisation des valeurs en cas de conflit. ***AI Ethics Label*** L'approche VCIO (valeurs, critères, indicateurs et éléments observables) présentée dans le rapport permet d'attribuer une note à chaque valeur sur un **label éthique** de l'IA inspiré du label d'efficacité énergétique.
Trustworthy use of artificial intelligence: priorities from a philosophical, ethical,	**Auteurs** Publication par une équipe interdisciplinaire de scientifiques des universités de Bonn et de Cologne dans le cadre d'un projet dirigé	**Commentaires des auteurs sur les travaux du Groupe d'experts indépendants de haut niveau de la Commission européenne (CE) au moment de la publication :**

legal, and technological viewpoint as a basis for certification of artificial intelligence, Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS (2020)	par le *Fraunhofer Institute for Intelligent Analysis and Information Systems* et avec la participation de l'Office fédéral allemand de la sécurité de l'information (BSI). **But : certification** Développer de façon multidisciplinaire un système de certification de l'IA pour mettre en place un label de qualité "AI made in Europe" de manière à identifier les technologies fiables et sûres, à les développer de façon responsable d'un PDV éthique et juridique, à soutenir la concurrence et à contribuer à l'acceptation de l'IA dans la société. Certification dirigée par des experts de plusieurs domaines (apprentissage automatique, droit, philosophie, éthique, sécurité informatique) + participation d'acteurs issus du monde des affaires, de la recherche et de la société.	Les recommandations de la CE sont essentiellement de nature générale et n'abordent pas les aspects juridiques ni les exigences éthiques dans un but de certification. Ils écrivent; Pour les exigences d'opérationnalité, les 6 domaines d'audit sont décrits de manière plus spécifique et approfondie que les catégories de la CE; La publication adopte « une approche horizontale et verticale » par rapport aux recommandations de la CE en examinant l'éthique et le droit et en mettant les deux en relation. **Interdisciplinarité** Publication qui met l'accent sur les interactions entre la philosophie, l'éthique, le droit et l'informatique et sur la nécessité de l'apport de chacune de ces disciplines à la conception et à la certification d'applications d'IA. **6 domaines d'audit** - Formulés à partir de l'éthique, de la philosophie, du droit et des exigences informatiques; - Forment la base d'un " audit catalog" d'IA permettant aux auditeurs de vérifier la fiabilité des apps d'IA.
AUSTRALIE		
Australia's AI Ethics Framework, Gouvernement de l'Australie (2020)	**Contexte** Recherche sur le cadre éthique financée par le gouvernement australien en 2018; Travaux guidés par un comité composé d'experts de l'industrie, du gouvernement et d'organisations communautaires et de chercheurs de *Data61* et du *CSIRO*; Processus consultatif de différentes parties prenantes à travers l'Australie mené à la suite de la publication de *Australia's Ethics Framework: A Discussion Paper (2019)* : - Soumissions écrites (universités, gouvernements, entreprises, organisations non gouvernementales, particuliers, tables rondes, consultations ciblées; - Analyse des soumissions et collaboration avec des experts en IA d'entreprises, d'universités et de groupes communautaires pour analyser les commentaires et développer l'ensemble révisé des principes éthiques de l'IA (2020).	-

CANADA		
La Déclaration de Montréal pour un développement responsable de l'IA, Université de Montréal, (2018)	Initiative de l'Université de Montréal. Déclaration issue d'un processus délibératif et inclusif (citoyens, experts, responsables publics, parties prenantes de l'industrie, organisations de la société civile et ordres professionnels.)	**Principes** - Principes non hiérarchisés qui doivent faire l'objet d'une interprétation cohérente; - Possibilité selon le contexte d'attribuer plus de poids à un principe ou de considérer qu'un principe est plus pertinent; - Possibilité de traduire les principes éthiques en langage politique ou juridique. **Remarques** - S'intéresse à tous les êtres sensibles; - S'intéresse au principe d'intimité; - Ne comprend pas de principe distinct de transparence/explicabilité; - Reconnaissance de l'activité numérique des utilisateurs comme travail contribuant au fonctionnement des algorithmes et créant de la valeur.
ÉTATS-UNIS		
Ethical machines: The Human-centric use of artificial intelligence, Lepri et al. (2021)	Publication de Bruno Lepri, Nuria Oliver et Alex Pentland, tous affiliés au Data-Pop Alliance (New-York, É-U).	**Études rapportées par les auteurs par rapport à** : **La relation entre explicabilité et équité** (Dodge et al, 2019) • Le type d'explication impacte la perception des utilisateurs de l'équité d'un algorithme; • Différents types d'équité peuvent demander différents types d'explications; • Des différences individuelles déterminent les réactions des gens à différentes sortes d'explications. **Le choix d'utiliser l'IA ou non** (Salganik et al, 2020) • Les meilleures prédictions étaient peu précises et à peine meilleures que celles obtenues via des modèles simples; • Auteurs recommandent aux décideurs de déterminer si la précision prédictive réalisable via des approches d'apprentissage-machine est adéquate pour une situation donnée + déterminer si les modèles d'apprentissage automatique sont significativement plus précis que des analyses statistiques simples ou des décisions humaines.
Responsible AI Global Policy Framework, ItechLaw (2019)	Publication de iTechLaw ayant impliqué des experts en droit de la technologie, des chercheurs et des représentants de l'industrie en provenance de 16 pays.	Recommandations divisées par acteur.
ANGLETERRE		

An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations, Floridi et al. (2018)	Auteurs membres du AI4People. Présentation des opportunités et risques de l'IA pour la société, de 5 principes éthiques et de 20 recommandations servant de base à une bonne société d'IA.	**Autodétermination humaine** - L'IA peut être utilisée pour faciliter la mise en œuvre de ***solutions/choix* humains** mais non pour en imposer – l'IA doit être au service de l'autodétermination humaine. **Respect de la loi nécessaire mais insuffisant** - Approche éthique requise pour (1) permettre aux organisations d'identifier et d'exploiter de nouvelles opportunités socialement acceptables ou préférables et (2) d'anticiper ou minimiser les erreurs coûteuses par la prévention d'actions socialement inacceptables (même si juridiquement incontestables). **Confiance du public** — Les avantages d'une approche éthique ne peuvent pas fonctionner sans confiance du public. — Confiance dépend de : ⇨ Engagement du public; ⇨ Perception des risques et protection contre les risques (via assurance ou mécanismes de réparation); ⇨ Mécanismes de règlementation et de recours accessibles et compréhensibles; ⇨ Transparence sur la manière dont les technologies d'IA fonctionnent (explicabilité); ⇨ Développement de mécanismes/solutions en amont (voir recommandation #6). ***Recommandations spécifiques à la confiance du public** : **#7** (processus de recours), **#8** (paramètres pour mesurer la fiabilité + "trust comparison index"), **#18** (certification d'IA éthique via des "trust-labels").
How to Design AI for Social Good: Seven Essential Factors, Floridi et al. (2020)	Article de Floridi, Cowls et Taddeo. Présentation de 7 facteurs éthiques essentiels à considérer dans la conception et le déploiement de l'IA.	**Facteurs** - 7 facteurs éthiques non hiérarchisés; - 7 facteurs qui ne sont pas en eux-mêmes suffisants pour une AI4SG, mais qui doivent tous être examinés; - Possibilité de tensions entre les facteurs et au sein d'un même facteur – L'IA elle-même pourrait aider à fournir des outils pour évaluer la meilleure façon d'atteindre cet objectif; - Plusieurs circonstances pour lesquelles il serait moralement défendable de ne pas incorporer un facteur. Mais essentiel de : ▪ Considérer chaque meilleure pratique de façon proactive; ▪ Ne pas incorporer la meilleure pratique si et seulement s'il existe une raison claire, démontrable et moralement défendable pour laquelle elle ne devrait pas l'être.

		Opportunité AI4SG : - Pour chaque cas donné, déterminer si l'on est moralement obligé ou obligé de ne pas développer un AI4SG. *Circonstances pour lesquelles l'IA ne sera pas le moyen le plus efficace pour résoudre un problème social en particulier. Ex : autres solutions plus efficaces ("Not AI for Social Good") ou risques inacceptables liés au déploiement de l'IA ("AI for Insufficient Social Good").

***