Good In Tech

Confronter la proposition de réglementation européenne aux incidents éthiques de l'IA

aout 2021

SOMMAIRE

Confronter la proposition de réglementation européenne aux incidents éthiques de l'IA

- **Introduction**
  - Résumé
  - Présentation de la base de données
  - Présentation de la proposition de la Commission européenne
- **Approche sélective**
  - L'IA ''forte'' catastrophiste ciblée par la Commission européenne
  - Le texte de la Commission comme barrière à l'entrée des IA chinoises
  - Les cas à haut risque ne se vérifient empiriquement qu'en partie
  - La protection des données personnelles et les cas de diffamation omis en partie dans la liste des cas à haut risque
  - Les incidents économiques et financiers absents du texte
- **Approche par gestion des risques**
  - Les incidents politiques impliquant des deepfakes sous-estimés
  - Les incidents systémiques sont difficilement quantifiables
  - La difficulté à anticiper à priori les risques implique la mise en place de contrôles à posteriori par le texte
- **Approche par autorégulation**
  - Les cas les plus graves sont généralement divulgués par des acteurs tiers
  - Une fois l'incident révélé, les fournisseurs adoptent des postures très diverses selon les cas d'usage

Les incidents sensibles pour les acteurs sont ceux qui touchent leur crédibilité technique ou leur image de marque

Les acteurs ont une approche utilitariste concernant les biais de genre que les biais racistes

Les acteurs s'excusent plus souvent si l'incident touche un client qu'un citoyen

Annexes

Tableau de classification des cas selon la proposition de la Commission européenne

Présentation de la méthodologie utilisée pour la base de données

Bibliographie

Auteurs

Théo Mercadal, Chargé de mission au sein de la Chaire Good in Tech

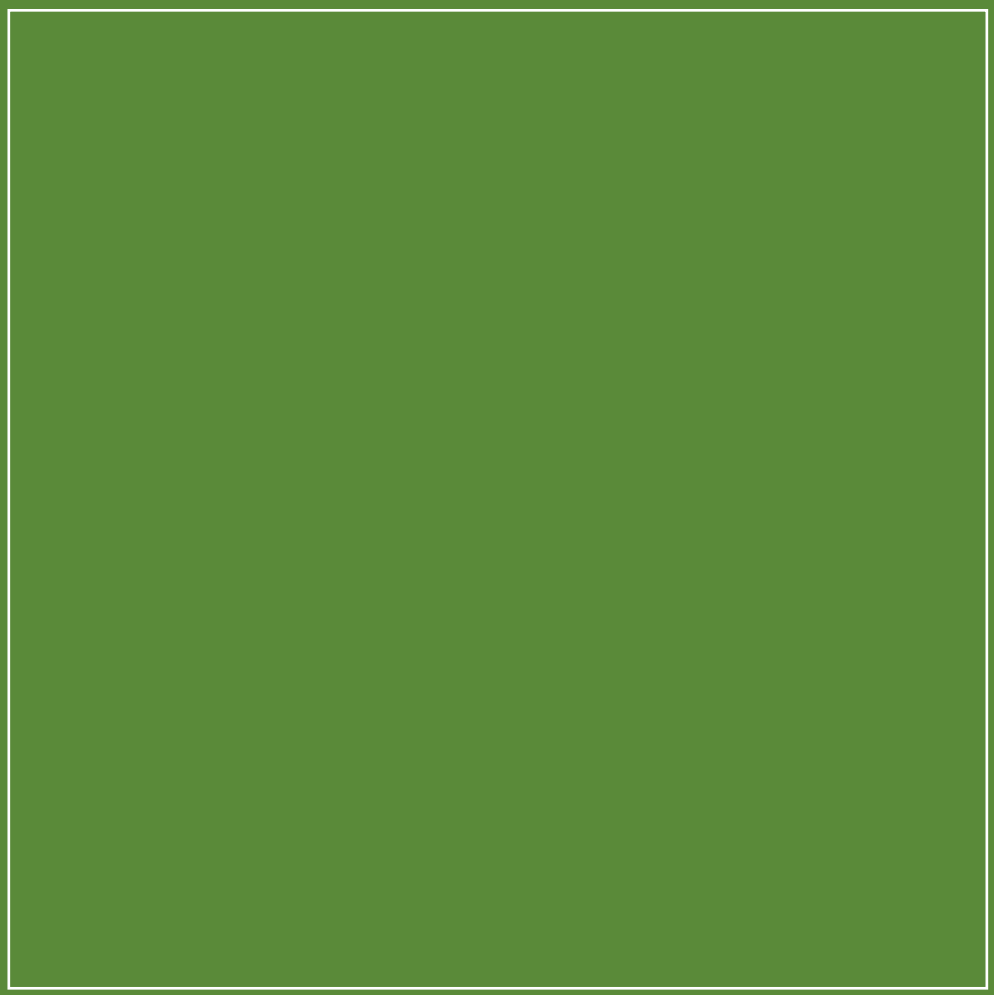
Sous la coordination de :

Jean-Marie John-Mathews, Data scientist, Coordinateur de la Chaire Good In Tech

Christine Balagué, Professeure à Institut Mines-Télécom Business School, Fondatrice et co-titulaire de la Chaire Good in Tech

Contact : jean-marie.john-mathews@imt-bs.eu

INTRODUCTION



Résumé

En avril 2021, la Commission européenne a rendu public une proposition visant à réglementer l'usage de l'intelligence artificielle (IA) sur le marché européen. Les principaux buts du texte sont de donner un cadre légal homogène à l'emploi de cette technologie dans l'UE et de réduire au maximum les procédures et législations pour encourager l'innovation dans le secteur. Pour cela, le texte vise à protéger les citoyens européens contre les excès de l'IA pour favoriser l'émergence d'une IA européenne compétitive par rapport aux concurrents chinois et américains.

Ce texte va être discuté et vraisemblablement amendé sur de nombreux points par le Parlement européen et le Conseil de l'Europe avant d'entrer en vigueur. Mais il semble important de déterminer dès maintenant les points forts et faibles du projet actuel pour contribuer à la discussion autour de la régulation de l'IA en Europe.

La démarche du présent rapport est de mettre en perspective les dispositions de la proposition à partir d'une connaissance empirique des incidents de l'IA recensés ces dernières années. Pour cela, nous nous baserons sur une base de données recensant 567 incidents éthiques en lien avec l'IA ayant eu lieu entre 2010 et 2021. Cette base a été constitue à partir des travaux de Charles Pownall et de l'Artificial Intelligence Incident Database (AIID), auxquels ont été rajoutées plusieurs métriques pour affiner l'analyse du texte européen. En mettant en parallèle les dispositions du texte avec des cas observés empiriquement en Europe et ailleurs dans le monde, le but est de déterminer si la proposition est en prise avec les pratiques réelles des acteurs de l'IA et dans quelles mesures elles peuvent être améliorées.

A partir de cette comparaison, deux principales constats sur la proposition de la Commission vont être dressés :

- la classification des emplois de l'IA en usage à risque limité, à haut risque ou risque inacceptable encadre imparfaitement les incidents empiriquement observés. En prenant le parti de fixer une telle pyramide des risques, le texte ne peut prévoir tous les futurs cas problématiques de cette technologie en perpétuelle mutation, et n'a pas vocation à le faire. La Commission introduirait donc par cette proposition un contrôle à postériori de l'IA consacrant un droit à l'oubli.
- la création d'un cadre favorable à l'autorégulation des acteurs pour contrôler les cas à haut risque n'est pas adaptée au comportement actuel des acteurs de l'IA et empêche une bonne application du texte normatif.

A l'issu de chaque partie, des recommandations indicatives seront proposées pour résoudre les points faibles du textes pointés par le présent rapport, dans le but d'alimenter les débats autour de la proposition de la Commission.

La base de données des incidents éthiques de l'IA

La base de données utilisée par ce rapport a été constituée à partir des bases de Charles Pownall et de l'AIID [1]. Chaque ligne de la base correspond à un incident ayant eu lieu avec l'IA qui a été relayé par la presse.

Plusieurs métriques créées par Charles Pownall ont été réutilisées sans modifications (type de problème éthique causé, pays concerné), d'autres ont été adaptées à l'étude du texte de la Commission (secteur touché, manière dont l'incident a été révélé), et certaines ont été créées pour ce rapport (niveau de gravité de l'incident, type d'acteur touché, type de réaction de l'acteur ayant été provoqué l'incident et classification de l'incident selon les critères de la proposition de la Commission).

Cette base ne vise pas à être exhaustive, mais permet d'avoir un aperçu satisfaisant des différents types d'incidents provoqués par l'IA. Elle ne cite que des incidents ayant été repris par la presse anglo-saxonne. Elle compte donc proportionnellement plus de cas ayant eu lieu aux Etats-Unis et au Royaume-Uni, et plus de cas ''business to consumer'' que ''business to business''.

Principaux axes de la proposition de la Commission européenne

La proposition de la Commission européenne du 21 avril 2021 a pour but :

- de donner un cadre légal homogène à l'IA dans l'UE pour plus de clarté
- de rationnaliser et réduire au strict minimum les procédures et législations pour encourager l'innovation et rester compétitif par rapport aux Etats-Unis et la Chine [2]
- de protéger les citoyens des excès de l'IA

Elle s'applique sur le territoire de l'UE, mais aussi à l'encontre des fournisseurs de systèmes d'IA vivant à l'extérieur du marché commun et important leurs algorithmes dans cet espace économique.

La proposition crée une pyramide de risques qui vise à interdire les risques inacceptables de l'UE, réglementer strictement les IA à haut risque, et ne contrôler que très faiblement les autres IA à risque limité.

Dans le cas des IA à haut risque, une procédure de vérification de conformité est mise en place. Une chaine de responsabilité entre le fournisseur, l'importateur, le distributeur et l'utilisateur est créée où chacun doit avertir l'acteur précédent s'il identifie un problème sur le système d'IA commercialisé. Le fournisseur est à la source de la responsabilité pénale en cas d'incident. Des obligations en terme de transparence vis-à-vis des consommateurs sont aussi demandées.

[1] l'ensemble de la méthodologie utilisée pour constituer la base de donnée est développée dans l'annexe du présent rapport, ainsi que le lien internet pour y accéder.

[2] Thibout Charles, ''*La compétition mondiale de l'intelligence artificielle*'', Pouvoirs, 2019

La proposition met également en place des mécanismes de contrôle de l'application du texte par les acteurs en prévoyant différents types de sanctions. Pour favoriser l'innovation, elle crée la possibilité de créer des ''bacs à sable'' (possibilité de déroger à certaines lois pour une durée limitée dans le cadre d'un programme de recherche).

Définition juridique de l'IA selon la Commission européenne [3]

'artificial intelligence system' (AI system) means software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommandations, or decisions influencing the environnements they interact with. (art 3 § 1)

Artificial intelligence technics and approaches :

(a) Machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning ;
(b) Logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems;
(c) Statistical approaches, Bayesian estimation, search and optimization methods (Annex I)

L'une des principales caractéristiques du texte de la Commission européenne est qu'il adopte une approche par risque. Contrairement à la plupart des guidelines précédentes régissant le domaine de l'IA, le débat éthique est très peu présent dans la proposition [4]. Le but est de remplacer autant que possible la subjectivité du débat moral et éthique de l'utilisation de systèmes par une quantification de la dangerosité de ces outils pour ses utilisateurs. Il s'agit non plus de déterminer si tel usage de l'IA est moral dans son ensemble par rapport aux sociétés humaines dans lesquelles il s'inscrit, mais de savoir s'il comporte un risque dans le temps et l'espace pour des individus précis, et si oui, de quelle intensité.

Le principal point positif de l'approche par risque est qu'elle permet de s'extraire de l'approche théorique propre à de nombreux travaux et rapports en éthique de l'IA ces derniers années. [5] [6]

[3] La définition de l'IA étant un débat non tranché à l'heure actuelle, le présent rapport utilisera celle utilisée par la Commission européenne. On remarquera que cette définition est elle-même prévue pour être évolutive suivant les mutations futures de cette technologie.
[4] le mot ''ethics'' n'apparait que 2 fois dans la proposition de la Commission européenne
[5] Hagendorff Thilo, ''*The Ethics of AI Ethics ; An Evaluation of Guidelines*'', Minds and Machines, 2020
[6] Mittelstadt Brent, ''Principles alone cannot guarantee ethical AI'', Nature Machine Intelligence, 2019

En effet, celle-ci s'interroge en général sur les possibilités de ''coder'' [7] un comportement éthique consensuel dans les systèmes d'IA, ou de déterminer à postériori par les bonnes pratiques des acteurs par des guidelines rarement contraignantes. Le but est de définir un corpus unitaire, exportable et réplicable de règles morales visant à trancher les différents dilemmes auxquels la machine peut être confrontée. Mais à l'image de la loi d'Asimov ou du dilemme du tramway, ces situations sont discutées, débattues, mais jamais formellement tranchées [8]. Ensuite, elle cherche à rendre ces dispositions applicables concrètement. Une fois que le dilemme a été tranché, est-il possible de le coder et de le rendre effectif par la machine ?

Or cette vision de l'éthique de l'IA peine à définir un ensemble de règles acceptées par les acteurs et de nombreuses guidelines concurrents ou superposables sont régulièrement publiées sans réels effets sur les comportements des fournisseurs de système d'IA. Cela est autant du au manque de dispositions contraignantes dans ces textes qu'au manque de consensus sut LA bonne éthique de l'IA.

A l'inverse, la proposition de la Commission européenne prend le parti de se détourner de cette approche en cherchant à quantifier le niveau de risque de leurs usages. Autrement dit, un système d'IA doit être réglementé s'il représente une menace quantifiable, physique ou morale pour les individus. Cette approche permettrait d'éviter les dilemmes insolubles évoqués plus haut. Par exemple, le texte ne dit pas comment une voiture autonome doit agir. Elle constate que cette machine représente un risque pour les passagers et les individus circulant autour d'elle. En conséquence elle détermine qui doit être tenu responsable de ses agissements en cas d'incident et comment le système d'IA gérant ses mouvements doit être contrôlé et régulé.

Dans ce cadre, le texte crée une pyramide des risques. Les risques inacceptables sont interdits sur le marché européen, les hauts risques sont strictement controlés, et les autres sont exemptés de la plupart des dispositions contraignantes de la proposition. Si cette approche est pertinente pour les raisons énoncées plus haut, elle comporte plusieurs limites que nous allons discuter dans ce rapport :

- la sélection des usages de l'IA à interdire ou contrôler déterminés par le texte surestime certains dangers et en oublie d'autres. (approche sélective)
- Les seuils de menace séparent les différents risques sont difficilement quantifiables pour certains incidents, qui n'entrent pas dans les conditions du texte malgré le préjudice qu'ils peuvent représenter pour les sociétés concernées. (approche par gestion des risques)
- Les contrôles des usages à haut risque se basent en partie sur une auto-régulation des acteurs, qui ne s'observe pas dans les faits et hypothèque une application efficace du texte. (approche par auto-régulation)

[7] Floyd Juliet, ''*La quête culturelle : Revisiter le test de Turing*'', Cités, 2019

[8] Etienne Klein ''Par définition, une question éthique, c'est une question qui n'a pas de solution'', conférence ThinkerView, 2018 (https://www.youtube.com/watch?v=KlwtT8cAAKI)

1. Approche sélective

La proposition de la Commission européenne met en place une classification sectorielle des applications de systèmes d'IA en fonction des risques qu'ils représentent pour les sociétés. Les ‘’risques inacceptables’’ doivent être interdits du marché commun de l'Union, les ‘’hauts risques’’ doivent être encadrés strictement par des procédures détaillées dans la suite du texte et les autres ‘’risques limités’’ n'entrant pas dans les 2 premières catégories sont exonérés de la plupart des contrôles.

La détermination de certains secteurs à contrôler et d'autres non est centrale dans l'efficacité du texte, car elle permet de cibler efficacement ou non les principales menaces que peut représenter l'IA pour les citoyens européens. La liste proposée par la proposition a d'ailleurs été critiquée à plusieurs reprisées car elle aurait oublié ou minoré certains aspects dangereux de l'IA [10].

Pour vérifier la pertinence des catégories déterminées par le texte, celles-ci ont été confrontées aux incidents recensés empiriquement, de façon à déterminer si la proposition couvre efficacement l'ensemble des cas de figure. Une typologie des différentes situations a ensuite été dressée pour déterminer les forces et faiblesses de cette pyramide des risques.

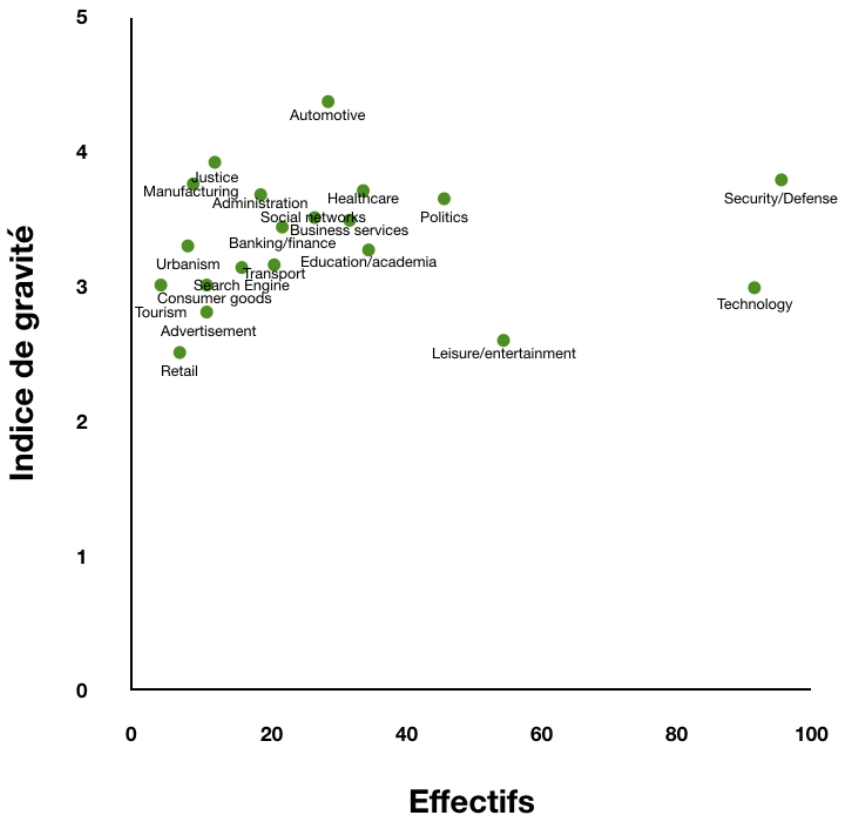


Figure 1 : Secteurs touchés par des incidents par leur gravité

[10] Ienca Marcello et Malgieri Gianclaudio, ‘’The EU regulates AI but forget to protect our mind’’, European Law Blog, 2021

La figure 1 permet de donner une idée générale des types de secteurs provoquant le plus d'incidents et les plus graves. On remarque notamment que les usages de l'IA se rapportant à la sécurité et la défense provoquent de très nombreux incidents, à la gravité importante [11]. Les cas de voitures autonomes ont la moyenne de gravité la plus haute, alors que les usages récréatifs de l'IA provoquent des incidents globalement moins graves. Si d'autres conclusions peuvent être tirées à partir de ce graphique, nous allons à présent comparer les résultats obtenus par l'observation empiriques des incidents éthiques de l'IA à l'encadrement qui est fait de ces technologies par le texte de la Commission européenne.

IA à risque inacceptable ciblées par la Commission européenne

The following artificial intelligence practices shall be prohibited :

(a) AI system that deploys subliminal techniques beyond a person's consciousness in order to materially distort a person's behaviour in a manner that causes or is likely to cause that person or another person physical or psychological harm;

(b) AI system that exploits any of the vulnerabilities of a specific group of persons due to their age, physical or mental disability, in order to materially distort the behaviour of a person pertaining to that group in a manner that causes or is likely to cause that person or another person physical or psychological harm;

(c) AI systems by public authorities or on their behalf for the evaluation or classification of the trustworthiness of natural persons over a certain period of time based on their social behaviour or known or predicted personal or personality characteristics, with the social score leading to either or both the following:

  (a) Detrimental or unfavorable treatment of certain natural persons or whole groups thereof in social context which are unrelated to the contextes in which the data was originally generated or collected;

  (b) Detrimental or unfavorable treatment of certain natural persons or whole groups thereof that is unjustified or disproportionate to their social behaviour or its gravity.

(d) The use of 'real-time' remote biometric identification systems in publicly accessible spaces for the purpose of law enforcement, unless and in as far as such use is strictly necessary for one of the following objectives.

  (a) The targeted search for specific potential victims of crimes, including missing children;

  (b) The prevention of a specific, substantial and imminent threat to the life or physical safety of natural persons or of a terrorist attack;

[11] Les différentes typologies utilisées dans la Figure 1 (secteurs et note de gravité de l'incident) sont expliquées dans l'annexe et sur la base de donnée.

> (c) the detection, localisation, identification or prosecution of a perpetrator or suspect of a criminal offence referred to in Article 2(2) of Council Framework Decision 2002/584/JHA62 and punishable in the Member State concerned by a custodial sentence or a detention order for a maximum period of at least three years, as determined by the law of that Member State

Les usages de l'IA interdits par le texte de la Commission européenne peuvent être regroupés en 2 types :

- les IA ''fortes'' (possédant une conscience et une volonté propre) qui influencent et nuisent au comportement des individus humains avec lesquels elles sont en interaction (alinéas (a) et (b)).
- Certains types d'IA de surveillance sociale et politique principalement utilisés en Chine (alinéas (c) et (d)).

L'IA forte catastrophiste ciblée par la Commission européenne

Les systèmes d'IA très spécifiques ciblés par les alizés (a) et (b) ne se retrouvent pratiquement pas sur la base de donnée. Des IA manipulant consciemment des individus ou groupes d'individus pour modifier leurs perceptions et comportements ne semblent pas exister aujourd'hui, ou sont du moins pas assez perfectionnées pour parvenir aux résultats escomptés. Ainsi, ces dispositions viseraient des types d'IA décrites dans les scénarios les plus alarmistes d'une technologie asservissant l'homme, mais qui ne semblent pas d'actualité.

On voit donc que les limitations de ces dispositions (l'IA doit exprimer une volonté claire et identifiable de nuire à l'individu, l'agissement de l'individu menant à l'incident doit clairement être corrélé à l'influence de l'IA) diminuent grandement son application à des cas concrets actuels. Certaines critiques [12] affirment même que ces dispositions ont plus un but rhétorique qu'une réelle volonté de cibler certains abus. Quelque soit la volonté des législateurs de la Commission européenne, il apparait que les cas ciblés par ces alinéas sont marginaux voire inexistants par rapport à l'utilisation actuelle de l'IA, même s'ils représentent en théorie des incidents particulièrement dangereux.

[12] Veale Michael et Zuiderveen Borgesius Frederik, ''Demystifying the Draft EU Artificial Intelligence Act'', SocArXiv Papers, 2021

Le texte de la Commission européenne comme barrière à l'entrée des IA chinoises

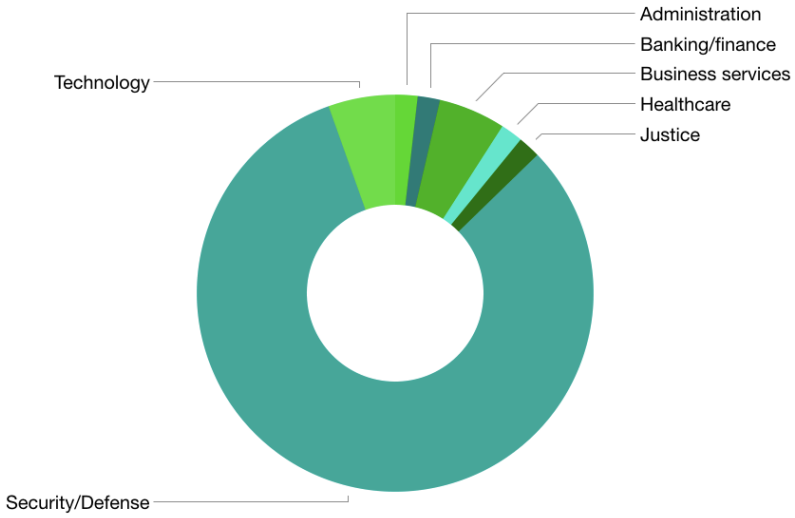


Figure 2 : Secteurs touchés par des incidents à risque inacceptable

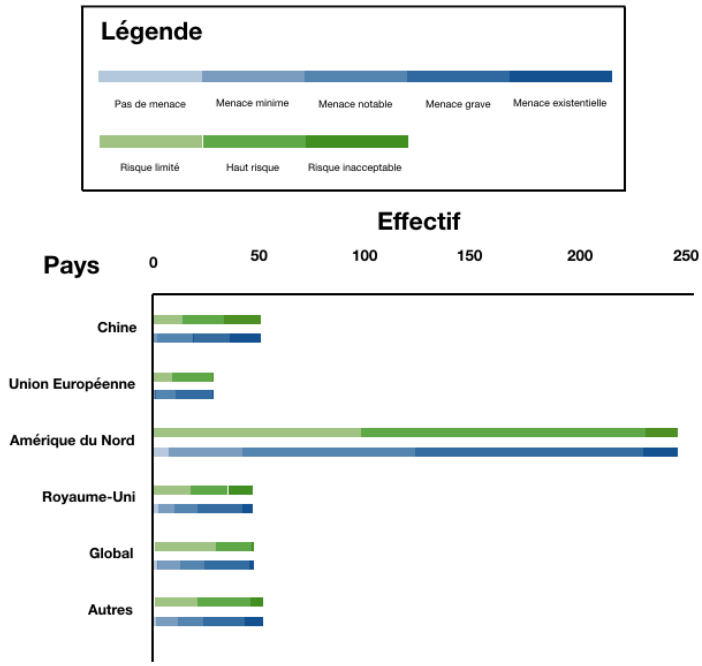


Figure 3 : Pays touché en fonction de la gravité de l'incident

Les graphiques montrent que parmi les cas considérés comme à haut risque par la Commission européenne, ceux concernant la Chine et les usages de sécurité et défense sont largement surreprésentés.

Les alinéas (c) et (d) ciblent des usages de l'IA bien plus actuels, et en général retrouvés en Chine. Tous les cas de notation sociale effectués par des pouvoirs publics recensés concernent des cas chinois. Certains incidents ayant eu lieu dans le secteur privé sont aussi présents aux Etats-Unis, mais ne sont pas concernés par cet alinéa.

De plus, les cas d'utilisation abusive d'algorithmes de reconnaissance faciale pour des fins répressives sont proportionnellement plus nombreux en Chine que dans les pays occidentaux. Seul un cas de ce type a été recensé sur le territoire de l'Union, qui concerne l'utilisation répressive de la reconnaissance faciale dans les métros parisiens pour vérifier le port du masque par les passagers. On peut nuancer ce constat par le fait que la base est principalement centrée sur des pays anglophones, donc que certains cas ayant eu lieu dans le reste de l'Europe ont pu être omis dans la présente étude.

Donc les risques inacceptables déterminés par la Commission européenne sont amenés à avoir une faible application sur le territoire de l'Union à court terme, mais pourraient avoir pour but de constituer des barrières à l'entrée face aux usages les plus dangereux de l'IA. L'utilisation par la Chine de l'IA pour développer des outils de surveillance socio-politique généralisés apparait comme une menace pour le modèle démocratique européen. Pour cette raison, le texte a certainement voulu se prémunir d'une éventuelle importation de certains de ces algorithmes sur le sol européen par certains Etats membres.

On remarque donc plus largement que les systèmes d'IA interdits dans le marché commun de l'UE sont très spécifiques et pratiquement aucun usage actuel de l'IA n'est visé par ces dispositions. Ce constat est en accord avec la volonté de la Commission de contrôler le moins possible le secteur pour favoriser l'innovation car une proportion infime des usages de systèmes d'IA actuels pourraient être interdits par ce texte. Ces dispositions ont pour but de développer un ''modèle européen de l'IA'' distinct des usages chinois ou américains de cette technologie en créant des barrières à l'entrée contre les aspects les plus saillants et menaçants de leur emploi de l'IA.

IA à haut risque ciblées par la Commission européenne

High-risk AI systems pursuant to Article 6(2) are the AI systems listed in any of the following areas:

1. Biometric identification and categorization of natural persons
2. Management and operation of critical infrastructures
3. Education and vocational training
4. Employment, workers management and access to self-employment
5. Access to and enjoyment of essential private services and public services and benefits
6. Law enforcement
7. Migrations, asylum and border control management (Annex III)

Sont aussi ciblés les systèmes d'IA utilisés comme composants de sécurité de produits inclus dans l'annexe II (art 6 § 1)

Tout nouveau cas peut être considéré comme à haut risque si ; ''the AI systems pose a risk of harm to the health and safety, or a risk of adverse impact on fundamental rights, that is, in respect of its severity and probability of occurence, equivalent or greater than the risk of harm or adverse impact posed by the high-risk AI systems referred to Annex III'' (art 7 § 1)

Dans les annexes II et III de la proposition, la Commission européenne fournit une longue liste des usages de l'IA devant être réglementés. Or l'application de ces critères aux cas recensés empiriquement montre que certaines catégories d'incidents sont surreprésentés et d'autres oubliés par le texte

Les cas ciblés à haut risque ne se vérifient empiriquement qu'en partie

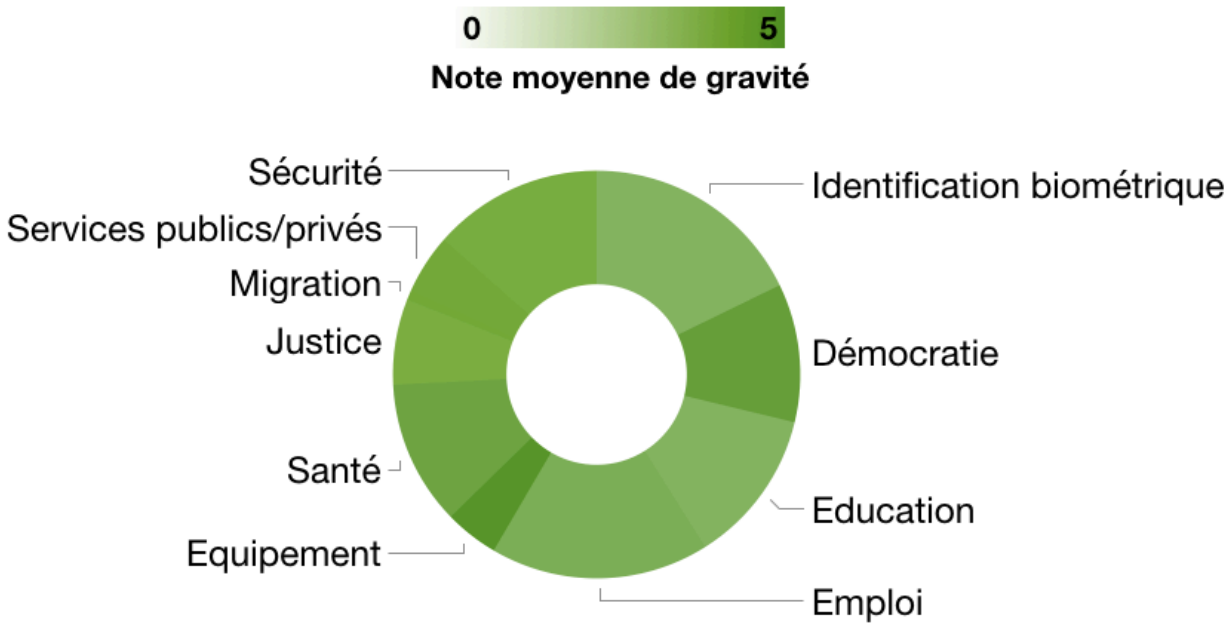


Figure 4 : Proportion des IA à haut risque et gravité moyenne [13]

On note que les 7 types d'IA mis en avant par la proposition de la Commission européenne visent des proportions inégales d'incidents recensés dans la base. Parmi les 571 incidents recensés, 241 ont été classés en usage à haut risque. Seuls 2 cas se rapportent aux contrôles des migrations et frontières (un détecteur de mensonges utilisé aux frontières de l'UE sur les migrants illégaux, et la collaboration entre Google et le gouvernement américain pour surveiller par l'IA la frontière mexico-américaine). De plus, les cas se rapportant à des IA comme composants de sécurité et ceux concernant l'accès aux services publics et privés essentiels ne représentent qu'une dizaine de cas chacun. Les cas de safety et de reconnaissance faciale sont ceux qui couvrent le plus d'incidents parmi ceux recensés.

[13] Les métriques utilisées se réfèrent au Tableau 1 de l'Annexe, et consistent à constituer des sous-groupes à partir de la liste d'IA à haut risque déterminée par le texte de la Commission européenne

Mais si certaines dispositions du texte ciblent des usages de l'IA peu communs, elles restent pertinentes au regard de la menace qu'ils représentent pour les utilisateurs (la note moyenne pour les cas de migrations et d'accès aux services publics et privés est de 4/5 et de 4,7/5 pour les cas de composants de sécurité). Donc si le texte de la Commission peut donner l'impression qu'il réglementer certains aspects très peu représentatifs de l'emploi d'IA, ceux-ci sont légitimes au regard de la menace qu'ils incarnent. Il semblerait toutefois possible de déterminer des catégories plus grandes incluant plus de cas à risque pour ces quelques dispositions.

Recommandations :

- modifier la disposition sur le contrôle des IA utilisés aux frontières car elle englobe très peu de cas trop spécifiques. Il pourrait être souhaitable de la remplacer une disposition ciblant un nombre plus large d'usages d'IA portant préjudice aux droits humains fondamentaux.

La protection des données personnelles et les cas de diffamation omis dans certains de la liste des incidents à haut risque de la Commission européenne

L'article 7 § 1 (cité plus haut) indique que tout cas présentant une menace pour la santé, les droits fondamentaux [14] ou la sécurité des individus a vocation à être classé comme à haut risque. Autrement dit, tout usage de l'IA portant préjudice à l'une de ces 3 dimensions est encadré ou doit être encadré par la liste des incidents à haut risque du texte de la Commission. Cependant, cette liste encadre les usages de l'IA par secteur, et non par valeur mise en péril par leur utilisation, donc il parait possible que certains cas portant atteinte aux droits fondamentaux d'utilisateurs ne soient pas encore formellement compris dans la liste des incidents à haut risque.

En l'occurence, si aucun cas parmi ceux recensés empiriquement porte atteinte à la sécurité ou à la santé d'individus tout en n'étant pas inclus dans la liste à des usages à haut risque, 80 cas peuvent remplir cette condition concernant une atteinte des droits fondamentaux. Ils concernent principalement des cas de diffamations et de manque de transparence sur les données personnelles des utilisateurs. Par ailleurs, d'autres études ont montré des manques du texte de la Commission européenne concernant la reconnaissance d'émotion ou les détecteurs de mensonge [15].

[14] Les droits fondamentaux de l'Union européens sont le respect de la dignité humaine, la liberté, la démocratie, l'égalité, l'état de droit et le respect des droits de l'homme, y compris des droits des personnes appartenant à des minorités (issu de l'art 2 du traité sur l'UE)

[15] Ienca Marcello et Malgieri Gianclaudio, ''The EU regulates AI but forget to protect our mind'', European Law Blog, 2021

Facebook retranscrirait les notes vocales envoyées sur Messenger

Après les révélations d'un lanceur d'alerte, Facebook a été accusé en 2019 de retranscrire des messages vocaux envoyés sur l'application Messenger pour qu'ils soient réutilisés pour entrainer des IA de reconnaissance vocale. Les utilisateurs concernés ne sont pas tenus au courant de la réutilisation de leurs conversations et n'ont pas exprimé de consentement clair en ce sens. L'entreprise a annoncé la fin de ces pratiques quelques semaines après ces révélations.

Google condamné pour diffamation sur son moteur de recherche

En 2012, Google est condamné par un tribunal japonais suite à la plainte d'un individu affirmant qu'il a perdu son emploi à cause des recommandations du moteur de recherche. Lorsque son nom était inséré sur le site, des liens faisant référence à des groupes criminels apparaissaient, ce qui a abouti à une plainte pour diffamation.

Recommandations :

- faire explicitement référence aux textes régulant la protection des données personnelles dans le titre IV abordant les obligations de transparence des fournisseurs de système d'IA.
- Inclure dans les cas à haut risque les systèmes d'IA provoquant des diffamations manifestes et graves envers des individus.

Les incidents économiques et financiers absents du texte

Enfin, le texte de la Commission ne fait aucune mention de l'utilisation de l'IA sur les marchés financiers. Pourtant son usage par les banques et institutions financières est fréquent. Plusieurs incidents concernant de tels acteurs sont aussi présents parmi ceux recensés, autant à l'échelle micro que macro.

Banqueroute du fond d'investissement Knight Capital Group

En 2012, le fond d'investissement Knight Capital Group perd d'immenses sommes d'argent suite à plusieurs mauvais placements décidés par une IA utilisée par l'entreprise. Le problème serait venu d'une erreur de manipulation d'un des employés, et aurait provoqué une baisse notable des valeurs des acteurs de centaines d'entreprises du NYSE. Le groupe a été racheté par Getco LLC pour éviter la faillite en décembre de la même année.

‘Flash crash’ de la livre sterling

En octobre 2016, la livre sterling connait une brutale chute de sa valeur avant de retrouver son niveau initial. Les causes de ce ‘’flash crash’’ sont multiples, mais les algorithmes de trading sont pointés du doigt pour avoir amplifié la chie de la devise. En interagissant entre eux, ils auraient créé une forme de chambre d’écho et aggravé la perte de valeur de la livre.

Ces deux incidents montrent que l’utilisation de systèmes d’IA dans le domaine de la finance sont susceptibles d’avoir des répercussions importantes pour le quotidien des citoyens européens. Si ces technologies n’ont pas jusqu’à présent représenté des menaces à la santé, à la sécurité et aux droits fondamentaux des citoyens européens, ils peuvent influencer de manière directe leur niveau de vie.

Cependant, ce type d’IA semble plus difficile à réglementer que les autres IA citées dans le texte de la Commission européenne, car leur empreinte sur la société est plus systémique, globale et diffuse, et les préjudices directs sur tel ou tel individu sont plus difficiles à prouver. La responsabilité de tel acteur ou algorithme est noyée dans de multiples opérations sur l’ensemble du marché et est complexe à quantifier. Donc les cas économiques ne sont pas compatibles avec l’approche par risque telle qu’elle est mise en place dans ce texte.

Pour ces raisons, il semble nécessaire d’intégrer ce type d’IA dans le texte de réglementation de la Commission en raison des menaces qu’il provoque, directement sur les investisseurs, comme indirectement sur des sociétés entières en cas de crises économiques. Pourtant, la nature de l’emploi de ces algorithmes nécessiterait une réglementation particulière en raison de la difficulté de déterminer les responsabilités de telle ou telle action de vente ou d’achat dans la dynamique globale du marché.

Recommandations :

- inclure les usages d’IA dans le secteur financier et bancaire dans la proposition de la Commission européenne. Les inclure parmi les cas à haut risque, ou selon d’autres procédures spécifiques en raison des particularités de ce type de systèmes d’IA

2. Approche par gestion des risques

A l'image des incidents économiques évoqués plus haut, l'un des points faibles du texte de la Commission européenne est qu'il ne semble pas encadrer efficacement certains incidents touchant l'ensemble d'une société, dont le préjudice envers tel ou tel individu est complexe à déterminer. En cherchant à quantifier le niveau de risque que représente un système d'IA en terme de sécurité, santé et droits fondamentaux des individus, le texte crée un cadre ou les incidents provoquant une menace équivalente pour un ensemble d'utilisateurs donnés sont regroupés sur la même échelle de la pyramide des risques. Des seuils, ou niveaux de préjudice sont créés en fonction des dommages causés aux utilisateurs. Mais parmi les cas recensés, certains causes des torts mineurs à l'échelle micro, mais peuvent constituer d'importantes menaces lorsque leurs conséquences individuelles sont agrégées à l'échelle macro. La difficulté est qu'il est complexe de déterminer une causalité claire entre l'incident et une conséquence sur des individus donnés, car son influence est brouillée dans un ensemble d'interactions plus nombreuses.

Les incidents politiques impliquant des deepfakes sous-estimés

Réglementation des deepfakes par la Commission européenne

''Users of an AI system that generates or manipulates image, audio or video content that appreciably resembles existing persons, objects, places or others entities or events and would falsely appear to a person to be authentic or truthful ('deep fake'), shall disclose that the content has been artificially generated or manipulated.
However, the first subparagraph shall not apply where the use is authorized by law to detect, prevent, investigate and prosecute criminal offences or it is necessary for the exercice of the right to freedom of expression and the right to freedom of the arts and sciences guaranteed in the Charter of Fundamental Rights of the EU, and subject to appropriate safeguards for the rights and freedoms of third parties" (art 52 § 3)

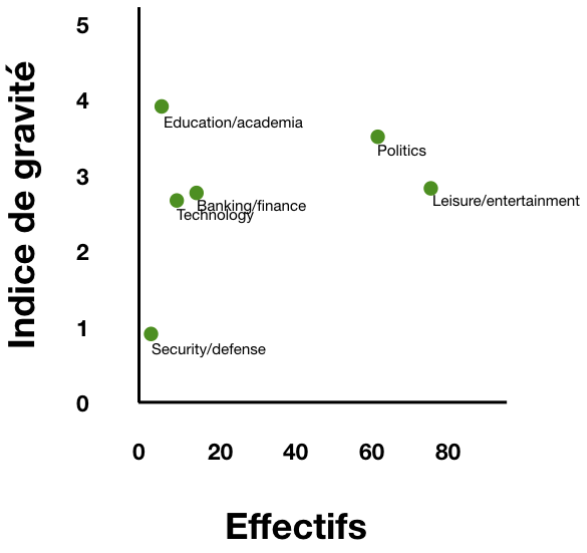


Figure 5 : Secteurs touchés par les incidents impliquant des deepfakes en fonction de leur gravité

Les deepfakes sont abordés à part des autres usages de l'IA dans la proposition de la Commission européenne. Leur réglementation n'est déterminée que brièvement dans l'article 52, dans la section visant les obligations de transparence pour certains systèmes d'IA. Les créateurs de deepfakes doivent indiquer qu'il s'agit d'images artificielles, sauf si des lois nationales autorisent leur divulgation dans le cadre de libertés diverses, ou dans le cadre d'investigations policières. Ces règles ne s'appliquent qu'aux deepfakes représentants des personnes et objets existants, et les contenus générants des individus totalement fictifs ne sont pas concernés. Au final, on s'aperçoit que les limitations apportées à cette technologie sont très faibles dans le texte.

Or il est important de noter que les deepfakes (autant audio, picturaux que vidéos) représentent près de 10% de la totalité des incidents recensés. Le graphique montre que ceux-ci ont dans la grande majorité une dimension récréative ou politique. Les premiers sont en général relativement anodins (une campagne publicitaire réutilisant l'image d'une actrice décédée, une émission télévisée présentée par un journaliste fictif …). Les seconds sont parfois particulièrement graves et ont d'importantes conséquences économiques ou politiques sur les personnes ciblées.

Un deepfake pornographique diffame un ministre malaisien

Une vidéo fait son apparition sur les réseaux sociaux en 2020 montrant le ministre de l'économie malaisien Muhammad Haziq avoir une relation sexuelle avec un autre homme politique du pays. L'homosexualité étant durement réprimée dans le pays, il est arrêté et risque la peine de mort. Sa défense est d'affirmer que la vidéo est un deepfake.

Un deepfake anti-Biden pendant la campagne présidentielle américaine

Dans le cadre de la campagne présidentielle américaine, le représentant républicain Steven Scalise partage sur les réseaux sociaux un deepfake du candidat démocrate Joe Biden affirmant devant une journaliste qu'il souhaite couper les fonds de la police. Si l'homme politique ne cache pas la nature artificielle de la vidéo postée, cette manoeuvre a inquiété les commentateurs politiques sur la croissance de l'utilisation de ces technologies comme outil de communication politique

De manière générale, le dommage créé par un deepfake n'est pas issu d'un dysfonctionnement de l'IA, comme c'est le cas dans la plupart des autres incidents, mais de l'usage qui en est fait à postériori pour servir un but politique, économique ou autre. Le problème n'est pas la fidélité scientifique d'une vidéo créée par deepfake par rapport à la réalité, mais l'existence même de cette copie artificielle. Que ces deepfakes soient à dimension satirique ou non, ils peuvent porter un préjudice tout aussi important pour la personne visée qu'une recréation parfaitement crédible d'une situation préjudiciable. En semant la confusion entre ce qui est vrai et ce qui ne l'est pas mais pourrait l'être, les deepfakes sont en passe de devenir des outils puissants de désinformation et peuvent constituer à terme de sérieuses menaces au débat démocratique européen.

Recommandations :

- établir un système de suivi équivalent à la procédure des systèmes d'IA à haut risque pour déterminer la traçabilité et la transparence des deepfakes (créateur, date de publication etc).
- Étendre la régulation des deepfakes aux contenus totalement artificiels (ne réutilisant pas des représentations d'individus existants).
- Supprimer les limitations de l'article 52 § 3 pour les deepfakes artistiques et policiers.

Les incidents systémiques sont difficilement quantifiables

Au-delà du cas des deepfakes, les incidents généraux, qui touchent de manière indiscriminée les individus sont peu ciblées et réglementées par le texte.

Le langage secret de chatbox de Facebook

En 2017, Facebook organise une expérience qui consiste à faire converser entre elles deux chatbox. Après quelques échanges, les deux IA commencent à utiliser leur propre langage en modifiant la grammaire anglaise et la rendre plus compréhensible pour elles. Perdant le contrôle de l'expérience, les scientifiques de l'entreprise y mettent fin.

Les biais de '80 Millions Tiny Images'

En 2020, des chercheurs montrent que la base de données ''80 Million Tiny Images'' comporte des biais nombreux qui aboutissent à des IA racistes et défectueuses en raison d'erreurs de labeling. Alors que ce dataset est utilisé comme données d'entrainements par de nombreuses IA, son usage a été retiré par le MIT.

Ces deux cas montrent qu'il est parfois difficile d'estimer les dommages d'une IA défectueuse sur une population donnée, malgré l'existence manifeste d'un incident éthique. Ici, la détermination de l'existence d'un préjudice envers un individu précis est très difficile à mettre en forme, donc ces cas ne peuvent pas être traités par le texte de la Commission européenne [16].

Cependant, le pari de la proposition est justement d'autoriser de tels cas pour ne pas entraver la recherche technologique dans le secteur de l'IA dans son ensemble même si elle flirte parfois avec les règles éthiques et les limites légales. Le but est qu'à moyen et long terme, de telles recherches contribuent à un enrichissement de la connaissance et que l'absence de limitations légales et morales de tels travaux permettent d'éviter que les chercheurs européens

[16] Veale Michael et Zuiderveen Borgesius Frederik, ''Demystifying the Draft EU Artificial Intelligence Act'', SocArXiv Papers, 2021

prennent du retard sur leurs concurrents chinois et américains. Le texte prévoit d'ailleurs la mise en place de ''bacs à sable'', soit la possibilité pour une durée de temps définie de déroger aux lois existantes pour mener des expérimentations en IA.

Cet objectif des législateurs européens est légitime, mais il reste souhaitable que de telles recherches fassent l'objet d'un suivi continu pour éviter les excès les plus graves.

La difficulté à anticiper a priori les risques implique la mise en place de contrôles à postériori par le texte

Les arguments ci-dessus ont permis de pointer du doigt certains manques de la pyramide des risques mise en place par la proposition de la Commission européenne. Par rapport aux incidents éthiques de l'IA observés empiriquement, le texte normatif surreprésente certaines menaces et en occulte d'autres. En supposant que les manques pointés par ce rapport soient pris en compte par les législateurs européens, il reste très probable qu'une étude équivalente à celle menée ici trouve d'autres failles et menaces oubliés par le texte dans quelques années. En effet, l'IA est un ensemble de technologies en mutation trop rapide et diverse pour que l'ensemble de ses usages actuels et futurs puissent être prévus à priori. Il parait donc impossible de déterminer une pyramide des risques parfaitement exhaustive, prévoyant tous les incidents possibles et les encadrant efficacement. [17]

Certains mécanismes inclus dans la proposition pourraient permettre de contourner cette objection, mais ils mériteraient d'être approfondis, car leur application conditionne en grande partie l'efficacité de la proposition dans son ensemble.

Ensemble des mécanismes d'un contrôle à postériori de la proposition

Création d'une base de donnée des usages à haut risque

1. The Commission shall, in collaboration with the Member States, set up and maintain a EU database containing information referred to in paragraph 2 concerning high-risk AI systems referred to in Article 6(2) which are registered in accordance with Article 51
2. The data listed in Annex VIII shall be entered into the EU database by the providers. The Commission shall provide them with technical and administrative support.
3. Information contained in the EU database shall be accessible to the public.
4. The EU datable shall contain personal data only insofar as necessary for collecting and processing information in accordance with this Regulation. That information shall include the names and contact details of natural persons who are responsible for registering the system and have the legal authority to represent the provider.

[17] Ganascia Jean-Gabriel et Powers Thomas, ''The Ethics of the Ethics of AI'', The Oxford Handbook of Ethics of AI, 2020

5. The Commission shall be the controller of the EU database. It shall ensure to providers adequate technical and administrative support." (art 60)

Amendement de la liste des usages à haut risque
''The Commission shall assess the need for amendment of the list in Annex III once a year following the entry into force of this Regulation'' (art 80 § 1)

On a vu que les cas inacceptables regroupent des usages très spécifiques et ne constituent pas le coeur du contrôle des systèmes d'IA mis en place par le texte. La régulation introduite par la proposition de la Commission européenne est essentiellement centrée sur les cas à haut risque. En complément des mesures mises en avant précédemment, on note que le texte prévoit la création d'une base de donnée recensant tous les cas à haut risque recensés (avec la documentation produite par les fournisseurs de ces systèmes d'IA) et un amendement systématique de la liste des usages concernés par cette disposition.

On peut supposer que ces outils servent à consolider les contrôles à postériori des usages à haut risque. En créant une base regroupant tous les documents concernant les usages à haut risque, la Commission vise à harmoniser les démarches et comportements des acteurs, tout en favorisant l'émergence d'une émulation collective. Les acteurs produisant des IA sont incités à révéler eux-mêmes les incidents en alimentant la base de données, ce qui pourrait permettre une autocorrection de l'ensemble des acteurs du domaine. Un comité de l'intelligence artificielle indépendant est créé par les instances européennes pour aider et conseiller les acteurs dans leurs procédures. Enfin, la Commission se réserve le droit de modifier tous les ans la liste des IA à haut risque en cas d'émergence de nouveaux usages de l'IA problématiques et non prévus par la proposition initiale.

Tous ces mécanismes vont dans le sens d'un principe de droit à l'erreur. La Commission européenne acte le fait qu'elle ne parviendra pas à déterminer tous les cas d'usage problématique de l'IA dès la première version du texte à cause de l'extrême diversité de ses emplois et de ses mutations rapides. Dès lors, les acteurs ont le droit de se tromper, de produire des incidents éthiques, à condition qu'ils le reconnaissent et qu'ils s'inscrivent dans un cadre général de correction des comportements en alimentant la base de donnée de la Commission et en l'aidant à améliorer la législation. Les amendes fixées par la proposition ne ciblent d'ailleurs pas les acteurs ayant provoqué des incidents, mais ceux qui ne les déclarent pas à la Commission.

Ce mécanisme de contrôle à postériori des incidents est donc central pour l'efficacité globale du texte normatif, car c'est lui qui détermine les bases de la coopération des acteurs de l'IA avec les nouveaux contrôles introduits par la proposition. Mais on peut déplorer que ce processus n'est pas explicitement formulé dans le texte. Contrairement à la pyramide des risques, qui est

longuement détaillée dans l'introduction du texte puis à travers la succession logique des articles normatifs, le contrôle à postériori n'est jamais clairement évoqué, et chaque mesure qui le compose est décrite séparément les unes des autres, sans référence à la cohérence d'ensemble.

Autrement dit, la philosophie générale du contrôle à postériori semble pertinente pour encadrer les usages d'une technologie aussi mouvante que l'IA, mais celle-ci ne semble pas assez clairement explicitée dans le texte, ce qui risquerait de fragiliser la mise en application de ce mécanisme.

Recommandations :

- rendre plus explicite le mécanisme de droit à l'oubli en insistant sur la cohérence et l'interdépendance de toutes ces étapes ; création d'une base de données / obligation de rendre compte des incidents éthiques pour l'alimenter / création d'un comité de l'intelligence artificielle pour aider les procédure.
- Faire éventuellement figurer de manière explicite le principe de droit à l'erreur.

3. Approche par autorégulation

Ensemble des mécanismes mettant en place une autorégulation des acteurs de l'IA par la proposition

Définition juridique du fournisseur de systèmes d'IA :
‘’’Provider’ means a natural or legal person, public authority, agency or other body that develops an AI system or that has an AI system developed with a new to placing it on the market or putting it into service under its own name or trademark, whether for payment or for free of charge” (art 3 § 2)

Obligations du fournisseur de systèmes d'IA :
- (en cas d'identification de problèmes sur le système d'IA concerné), ‘’provider shall immediately inform the national competent authorities of the Member States in which it made the system available (art 22)
- ‘’Such notification shall be made immediately after the provider has established a causal link between the AI system and the incident or malfunctioning or the reasonable likelihood of such a link, and, in any event, not later than 15 days after the providers becomes aware of the serious incident or of the malfunctioning” (art 62 § 1)
- Liste de la documentation nécessaire que doit délivrer le fournisseur décrivant le système d'IA mis sur le marché européen (art 13)

Obligations des autres acteurs faisant transiter le système d'IA :
- (en cas d'identification de problèmes sur le système d'IA concerné), ‘’the importer shall inform the provider of the AI system and the market surveillance authorities to that effect’’ (art 26)
- ‘’where the system presents a risk within the meaning of Article 65(1), the distributor shall inform the provider or the importer of the system” (art 27).

L'un des principaux points de la proposition de la Commission européenne est de créer les conditions d'une forme d'autorégulation des acteurs du domaine de l'IA. Il introduit la personnalité juridique du fournisseur, à la source de toute responsabilité concernant les programmes qu'il produit et vend. Il doit de plus fournir les documents facilitant l'enregistrement du système d'IA sur le sol européen, coopérer avec les éventuels contrôles, et il est le premier visé en cas d'incidents provoquées par ses produits. Par ailleurs, une chaine de responsabilité entre les différents acteurs intervenant dans la commercialisation du système d'IA sur le marché européen est créée. Le principe est que chacun doit avertir l'acteur précédent de la chaine s'il détecte un problème sur le fonctionnement du produit et vérifier le bon suivi de la procédure d'homologation du système d'IA par le fournisseur.

L'un des objectifs du texte annoncés par par la Commission européenne est de réduire la bureaucratie et l'intervention des instances de contrôle dans le domaine de l'IA pour réduire les barrières à l'innovation technologique et au développement économique du secteur. Ce principe est assuré par la détermination de certains domaines à risque (voir par ailleurs), mais également ici par la création d'un cadre favorable à l'autorégulation des acteurs. Ainsi, il

demande explicitement, et sous menace d'amendes conséquentes, à chaque acteur de la chaine fournisseur-importateur-utilisateur de contrôler les performances des systèmes d'IA.

Notre thèse est ici de montrer que l'autorégulation de leurs pratiques n'a lieu que lorsque les gains réputationnels sont plus élevés que les pertes économiques et de productivité d'une modification de leur emploi de l'IA ciblée par les critiques. Autrement dit, les fournisseurs d'IA auraient une forte tendance à recourir à des comportements utilitaristes. Ces acteurs évalueraient les coûts et bénéfices que représenterait chacune de leurs actions (divulguer ou non un incident, s'excuser ou non lorsqu'il a été révélé …) et ce type de raisonnement n'incite que très rarement ces acteurs à aller dans le sens d'une autorégulation.

Ce constat menace le principe d'autorégulation tel qu'il est mis en place dans la proposition de la Commission européenne. Soit le manque d'auto-correction des acteurs est comblé par des contrôles efficaces et nombreux de la part d'instances publiques aux budgets conséquents, soit le texte sera peu ou pas appliqué, de la même façon que les autres guidelines portant sur les pratiques éthiques de l'IA

Les cas les plus graves sont généralement divulgués par des acteurs tiers

S'il y avait une autorégulation efficace de la chaine fournisseur-importateur-utilisateur, on pourrait s'attendre à ce que la plupart des incidents éthiques recensés aient été divulgués par ces acteurs. Or ils occupent une place marginale dans la révélation de nouveaux incidents, et ce sont des acteurs tiers (médias, ONG, recherches académiques …) qui effectuent en général cette action.

On s'aperçoit que dans les cas recensés, les entreprises et pouvoirs publics utilisant des systèmes d'IA ne communiquent en premier que très rarement, et qu'ils ont tendance à le faire principalement dans les cas les moins graves.

Les acteurs tiers (médias, recherches académiques et ONG) jouent un role central dans la régulation du secteur, en divulguant la majorité des incidents concernant les systèmes d'IA. Si leur action n'est pas évoquée dans la proposition de la Commission européenne, elle est essentielle pour la mise en place de pratiques éthiques dans le secteur de l'IA. En attaquant le capital réputationnel des acteurs à l'origine des incidents éthiques, les médias et autres permettent de rééquilibrer légèrement la balance des coûts-bénéfices d'une dissimulation d'un problème de système d'IA. Mais cette action des médias, recherches académiques et ONG n'est pas suffisante pour faire évoluer les pratiques des fournisseurs de systèmes d'IA et autres.

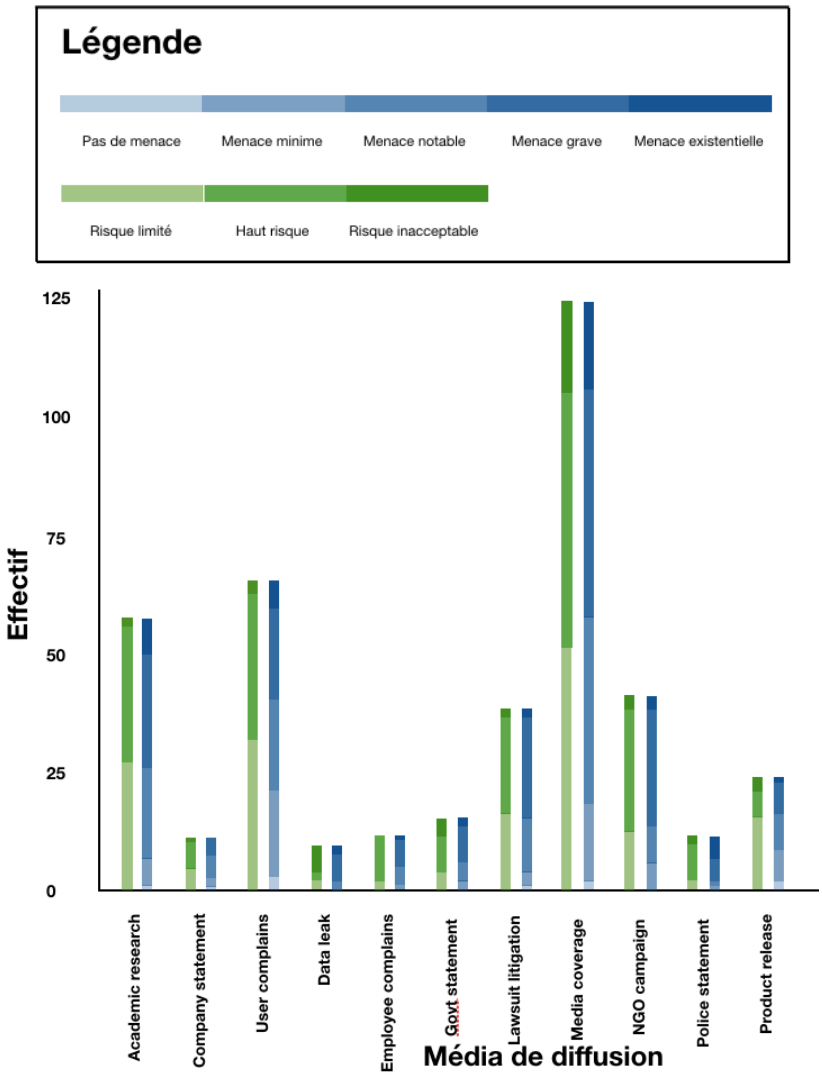


Figure 6 : Répartition des médias de diffusion selon leur gravité

En effet, lorsqu'un incident est détecté par un des acteurs de la chaine, sa révélation est à double tranchant. D'une part, elle peut permettre de renforcer son image de probité et d'honnêteté, tout en provoquant des coûts réputationnels (l'acteur concerné peut paraitre incompétent en raison des limites techniques des systèmes d'IA utilisés ou vendus) et économiques (l'arrêt de l'utilisation de l'IA concernée par l'incidence peut entrainer des pertes de productivité ou d'efficacité). D'autre part, ne pas révéler l'incidence peut permettre de continuer à utiliser le système d'IA problématique tant qu'aucun autre acteur de la société n'ait perçu ou révélé ses limites. Si elle est révélée, le coût réputationnel de la dissimulation de l'incident peut toujours être atténué par sa communication ultérieure (par exemple en s'excusant, ou en rejetant énergiquement les critiques). Donc l'acteur n'a intérêt à divulguer un incident que si les conditions suivantes sont réunies :

- la probabilité que l'incident soit révélé par un acteur tiers est très forte
- Une révélation de l'incident par un acteur tiers représente un important coût réputationnel

- Une modification de l'IA concernée ne provoque pas de coûts économiques et fonctionnels importants.

Une fois l'incident révélé, les fournisseurs adoptent des postures très diverses selon le type d'incident

Alors qu'on pourrait penser qu'après que l'incident ait été divulgué, l'acteur de la chaine fournisseur-importateur-utilisateur concerné a tout intérêt à reconnaitre les limites de son système d'IA pour minimiser le coût réputationnel de la révélation de ses défaillances. Une promesse de résoudre les problèmes pointés par les critiques permettrait de garder une certaine probité et volonté d'améliorer l'éthique du produit d'IA visé. Mais cette posture entraine aussi un coût, lié à la modification substantielle de l'IA, qui peut s'avérer très élevé si l'incident touche le coeur de la stratégie commerciale ou économique de l'acteur concerné, ou si elle est techniquement impossible. Dans ces cas, un rejet des critiques extérieures et un déni de l'incident paraissent tout aussi pertinents.

Donc une posture conciliante de la part de ces acteurs une fois l'incident révélé ne va pas de soi, et dans les faits, elle n'est adoptée que dans certains cas précis.

On remarque que les incidents sont reconnus par les acteurs à leur origine principalement quand elles sont considérées comme peu grave, facilement corrigibles ou trop sensibles pour leur image de marque.

Ainsi, un même acteur peut adopter des réactions opposées pour un incident similaire selon les thèmes socio-politiques qu'il peut englober, et donc les effets qu'il peut avoir sur son image de marque.

Les biais sexistes de l'IA de recrutement d'Amazon

En 2018, des critiques de la presse ciblent une IA utilisée par Amazon depuis 2015 pour sélectionner les CV des candidats à des emplois dans l'entreprise. L'algorithme aurait été préentrainé à partir de données recensant le profil des embauchés des 10 dernières années, majoritairement des hommes, et l'IA aurait reproduit le biais dans ses recommandations, au dépens des femmes candidates. Suite aux révélations, Amazon a affirmé faire tout son possible pour améliorer cette IA, mais n'a pas souhaité réagir sur les critiques plus larges dénonçant l'emploi de cette technologie pour assurer les services de recrutement de l'entreprise.

L'IA d'Amazon licencie des malades du Covid-19

En pleine épidémie de covid, Bloomberg relaie des accusations d'employés d'Amazon qui disent avoir été licenciés suite à leur infection à la maladie. L'IA gérant les salariés de l'entreprise et licenciant les moins productifs aurait considéré que leurs congés maladie étaient abusifs et a refusé leurs demandes d'aménagement.

Les employés concernés ont tenté en vain de joindre des individus humains du service des ressources humaines, avant que certains d'entre eux ne soient licenciés. Amazon s'est excusé et a dédommagé les employés une fois que l'incident ait été rendu public. d'aménagement.

L'IA d'Amazon anti-syndicalisation

En 2020, des enquêtes de la presse accusent WholeFood, une filiale d'Amazon, de classer par une IA ses magasins selon la probabilité de syndicalisation de ses salariés. Le but est d'accepter certaines revendications des employés ayant le plus de chance de se syndiquer pour couper l'herbe sous le pied de tout mouvement social avant qu'il ne se structure. L'entreprise n'a pas confirmé l'emploi d'une telle IA, et a rappelé que son but était d'assurer les meilleures conditions de travail possible pour ses salariés.

La gestion des employés d'Amazon par une IA

Amazon est régulièrement critiquée pour l'emploi d'une IA pour gérer la productivité de ses employés, et notamment licencier les moins efficaces. L'entreprise rejette régulièrement ces dénonciations en affirmant que le but de tels algorithmes est d'améliorer la qualité du travail de l'ensemble de ses salariés et qu'ils sont bienveillants et non punitifs.

A partir de ces cas particuliers, on peut supposer qu'Amazon ne souhaite pas être une entreprise perpétuant des inégalités homme-femme en son sein, ou créant une différenciation absurde entre malades du covid et individus sains. Ici, le maintien du statu quo est plus coûteux qu'une modification de leurs pratiques. Concernant les critiques sur la gestion de ses employés, la situation est inversée, car il serait moins coûteux d'organiser une campagne de communication pour garder l'image de l'entreprise à un niveau acceptable que de réformer totalement le fonctionnement de l'entreprise et l'organisation de ses salariés.

Plus largement, l'étude des incidents recensés permet de déterminer quels sont les types d'incidents susceptibles de provoquer des attitudes plus conciliantes de la part des fournisseurs, et lesquels entrainent des positions de déni

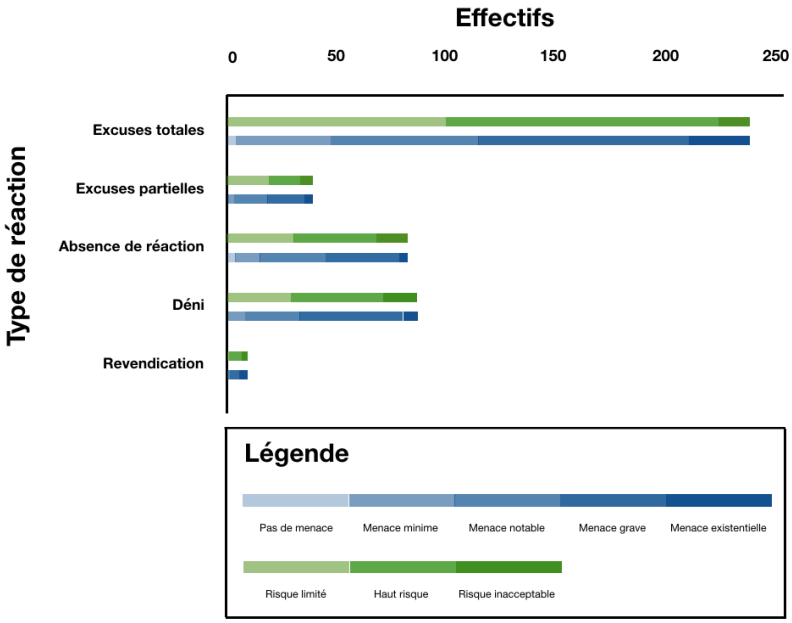


Figure 6 : Type de réaction des fournisseurs de système d'IA après révélation de l'incident en fonction de leur gravité

Les incidents sensibles pour les acteurs sont ceux qui touchent leur crédibilité technique ou leur image de marque

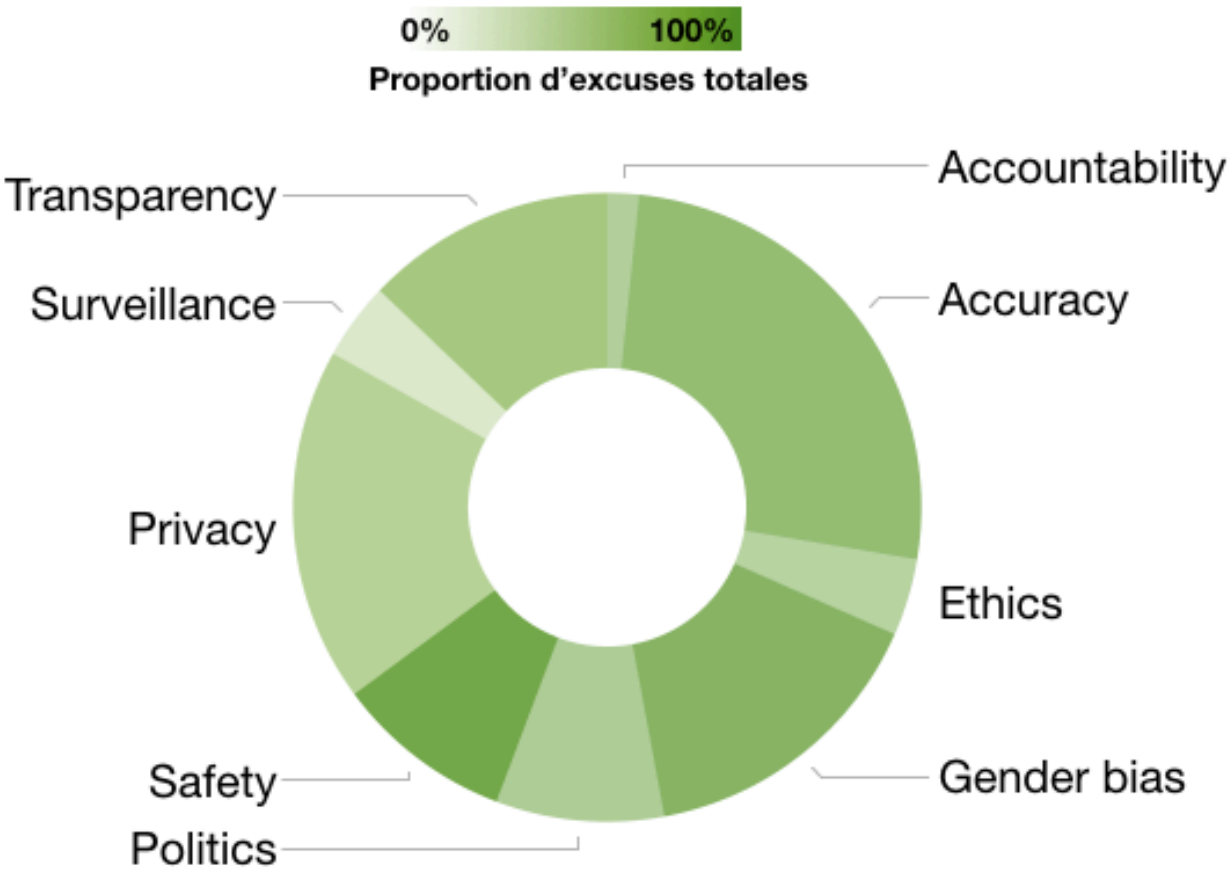


Figure 7 : Répartition des types d'incidents selon la proportion d'excuses lors de leur révélation

On constate que les trois types d'incident provoquant le plus d'excuses de la part des acteurs sont ceux touchant la performance technique des systèmes d'IA (accuracy et safety) et l'image politico-sociale de l'entreprise ou de l'administration publique concernée (biais/discriminations).

Les cas d'accuracy concernent des systèmes d'IA qui retournent des résultats manifestement erronés pour des raisons variées. Ces IA sont généralement dénoncées pour leurs performances et non pas leur usage, donc le coût économique d'une prise en compte des critiques sera moindre. Il est en effet plus facile pour l'acteur de modifier le fonctionnement d'un système d'IA pour améliorer ses performances que d'accepter la fin de son usage pour des raisons éthiques. C'est entre autre pour ces raisons que ces incidents provoquent des attitudes plus conciliantes de la part des acteurs. Les cas de safety sont similaires et concernent principalement des accidents de voitures autonomes.

Les biais de Google Home

Avec la commercialisation du speaker Google Home, de nombreux utilisateurs se sont plaints de bugs de l'appareil, qui ne parvient pas à identifier certains accents ou certaines voix. L'entreprise s'est immédiatement excusée en affirmant qu'elle travaille à l'amélioration de ses données d'apprentissage pour corriger ces différentes erreurs.

Google Home écouterait ses utilisateurs

Un média belge accuse le logiciel Google Assistant, notamment utilisé par les speakers de l'entreprise, d'espionner les utilisateurs en les enregistrant en permanence, mais aussi de réutiliser les enregistrements pour améliorer les algorithmes de reconnaissance vocale de l'entreprise. Google embaucherait en effet des individus pour retranscrire les sons récupérés par les speakers pour permettre à d'autres IA de s'entrainer sur ces supports, à l'insu des utilisateurs. Après la révélation de cet incident, l'entreprise a nié de tels agissements et affirmé que tout enregistrement effectué par ses speakers l'a été par erreur.

A partir de ces incidents, on voit que lorsqu'il est question d'incompréhensions du système d'IA, Google promet des corrections rapides de l'algorithme. La correction est réalisable techniquement, et mettre sur le marché un outil à reconnaissance vocale aux performances aléatoires ne peut que nuire à la vente de ce produit et à la crédibilité de l'entreprise. A l'inverse, lorsqu'il est question de dénonciations de l'écoute systématique des utilisateurs par le speaker pour des fins commerciales ou autres, Google adopte une position bien plus intransigeante. Revenir sur cette fonctionnalité du produit remettrait en cause les fondements de sa rentabilité économique pour l'entreprise, donc sacrifier une partie de sa réputation pour maintenir le speaker sur le marché tel quel semble l'alternative la moins coûteuse.

Les acteurs prennent plus au sérieux les cas de biais de genre que les biais racistes

Par ailleurs, les incidents concernant les biais et discriminations suscitent des réactions fortes de la part des acteurs de la chaine fournisseur-importateur-utilisateur. Mais on remarque étonnamment que si les biais de genre recueillent généralement des excuses et attitudes conciliantes, l'inverse se produit pour les incidents de biais racistes. Comme vu plus haut, une entreprise préfère revoir son algorithme plutôt que d'être présentée comme sexiste, mais cela ne semble pas être le cas concernant les accusations de racisme.

Une des raisons pour expliquer ce résultat serait de remarquer que les algorithmes amenés à différencier les individus selon leur genre ou leur origine ethnique sont utilisés dans des secteurs différents. Les algorithmes produisant des biais de genre sont surtout présents sur les réseaux sociaux et dans les algorithmes de recrutement. Ceux produisant des biais racistes sont plutôt liés à des systèmes de surveillance et à la justice prédictive. Or nous allons voir que les incidents donnant lieu à des liens commerciaux et économiques entre vendeur et client sont traités différemment des incidents plus ‘’politiques’’.

L'acteur s'excuse plus souvent si l'incident touche un client qu'un citoyen

Type de réaction	Customer	Private company	General public	Individual	Employee	Social group	Citizen	Total
Excuses totales	23,57 %	12,14 %	3,93 %	20,71 %	5 %	17,86 %	14,64 %	100 %
Excuses partielles	19,57 %	0 %	10,87 %	10,87 %	2,17 %	41,3 %	15,22 %	100 %
Absence de réactions	8,7 %	6,96 %	9,57 %	28,7 %	6,96 %	16,52 %	20 %	100 %
Déni	9,68 %	6,45 %	8,60 %	15,05 %	19,35 %	19,35 %	21,52 %	100 %
Revendi-cation	7,69 %	0 %	0 %	7,69 %	7,69 %	30,77 %	46,15 %	100 %
Total	13,34 %	7,06 %	7,4 %	19,71 %	10,07 %	20,4 %	20,74 %	100 %

Figure 8 : Tableau croisé des types de réactions des acteurs à l'origine de l'incident et du type d'acteur touché par l'incident

Lorsqu'une entreprise est épinglée par les médias pour une pratique non éthique, le premier risque est que cette révélation entraine une défiance des clients et donc des pertes économiques. Or l'étude des cas recensés empiriquement montre que lorsque les clients sont touchés par un incident, les acteurs à son origine sont beaucoup plus enclins à s'excuser pour limiter le coût commercial d'une telle divulgation.

A l'inverse, les incidents touchant des citoyens recueillant des réactions moins conciliantes. Ce constat doit être pondéré par le fait que les cas concernés touchent principalement des Etats autoritaires et non-occidentaux (les incidents touchant les Ouïgours par exemple), où les gouvernements peuvent se permettre plus facilement d'adopter des attitudes allant à l'encontre des droits fondamentaux des citoyens.

Mais la même tendance reste observable au sein des pays occidentaux, bien que moins extrême. Dans le cas d'incidences liées à l'usage de la reconnaissance faciale dans des lieux publics, les réactions sont fortement corrélées au type d'utilisateur touché.

Les dysfonctionnement du système anti-vol de Walmart

Un groupe d'employés anonymes de Walmart s'est plaint via la presse de l'utilisation d'une IA anti vol dans les supermarchés de l'entreprise. Celle-ci serait inefficace voire contre-productive, car elle ne parviendrait pas à identifier les voleurs et rallongerait les files d'attente. Suite à ces critiques, Walmart a supprimé l'usage de cet outil

L'utilisation de reconnaissance faciale par la police londonienne

La police londonienne a annoncé l'utilisation de la reconnaissance faciale sur ses dispositifs de vidéosurveillance à partir de 2020. Cette déclaration a suscité d'importantes réactions dans la population et parmi les associations de défense des droits de l'homme. Le chef de la police a assuré qu'un tel usage de l'IA était légal et était dans l'intérêt des citoyens, et a rejeté tout scepticisme envers l'emploi de cette technologie dans la ville.

Ces 2 exemples illustrent le fait que lorsqu'un lien commercial existe entre l'acteur touché et celui responsable de l'incident, ce dernier accepte plus facilement de renoncer à ses pratiques les plus invasives. A l'inverse, quand les individus sont touchés en tant que citoyens, les autorités à l'origine de l'incident préfèrent rappeler la légalité de leur action, ou nier l'emploi de telles technologies par leurs services plutôt que de renoncer à leur emploi.

De plus, lorsque les incidents touchent des individus en raison de leur appartenance à des groupes sociaux (ethnique, religieux, socio-économique …) certains acteurs décident de donner des excuses limitées [17]. Celles-ci consistent à refuser d'endosser la responsabilité intégrale de l'incident tout en reconnaissant l'erreur du système d'IA. Par exemple, si un moteur de recherche produit des biais racistes dans ses résultats, l'entreprise le développant s'excuse tout en rappelant que ces résultats sont aussi liés aux recherches des utilisateurs antérieurs. Autrement dit, les biais des IA seraient aussi le reflet des biais des sociétés humaines [18]. Il n'est pas question ici de déterminer si cette justification est pertinente, mais de remarquer qu'elle est

[17] Nurock Vanessa, ''L'intelligence artificielle a-t-elle un genre ? '', Cités, 2019

[18] Bronner Gérald, ''Apocalypse cognitive'', Presses Universitaires Françaises, 2021

surtout présente dans certains cas précis, où les liens commerciaux et économiques entre fournisseur et utilisateur de l'algorithme sont faibles ou inexistants.

Les précédents arguments suggèrent un manque de robustesse de la proposition pour installer une autorégulation efficace, permettant autant de protéger les initiatives innovantes que les droits fondamentaux des utilisateurs. Le comportement utilitariste des acteurs de l'IA va en effet régulièrement à l'encontre d'une attitude responsable éthiquement. Pour remédier à cette situation plusieurs axes peuvent être envisagés.

Recommandations :

- renforcer les contrôles des instances de régulation nationales et européennes. Cette pratique va a l'encontre de l'objectif de réduction de la bureaucratie européenne prônée par le texte, mais semble essentielle pour sanctionner efficacement les contrevenants et rééquilibrer les calculs coût-bénéfice des acteurs en faveur d'un comportement plus éthique envers les consommateurs.
- Créer des recours légaux pour que les citoyens dénoncent collectivement certains abus d'acteurs de l'IA (éventuellement en utilisant une procédure équivalente à celle existante pour la protection des données (RGPD)). Cet outil permet d'effectuer un filtre des cas à traiter pour les instances de régulation, en présélectionnant les incidents les plus sensibles. Il peut aussi permettre de relayer des incidents plus rapidement en laissant les individus se mobiliser et saisir les autorités dès l'apparition du problème éthique.

-

ANNEXES

Tableau de classification des cas selon la proposition de la Commission européenne (Annexe 1)

Type de cas	Type d'IA	Sous-Type d'IA	Article concerné	Métrique
Unacceptable risk	IA fortes		Art 5 § 1 - alinéas (a) et (b)	
	Notation sociale par des acteurs publics		Art 5 § 1 - alinéas (c)	
	Système d'identification biométrique à distance, en temps réel et à visée répressive		Art 5 § 1 - alinéas (d)	
High risk	IA utilisée comme composant de sécurité d'un des produits cités en Annexe II	Aviation civile, véhicules à moteur, ascenseurs, dispositifs médicaux, équipements sous pression …	Art 6 § 1	Equipement
	Autres systèmes d'identification biométrique		Art 6 § 2 / Annexe III	Identification biométrique
	Gestion et exploitation des infrastructures essentielles	Gaz, électricité, trafic routier …	Art 6 § 2 / Annexe III	Equipement
	Education et formation professionnelle	IA déterminant l'accès à des services de formation et d'éducation	Art 6 § 2 / Annexe III	Education
		IA notant les tests éducatifs d'élèves ou d'individus en formation	Art 6 § 2 / Annexe III	
	Emploi, gestion des salariés et accès au travail indépendant	IA utilisée pour recruter ou sélectionner des individus à un emploi	Art 6 § 2 / Annexe III	Emploi

Type de cas	Type d'IA	Sous-Type d'IA	Article concerné	Métrique
High risk	Emploi, gestion des salariés et accès au travail indépendant	IA utilisée pour recruter ou sélectionner des individus à un emploi	Art 6 § 2 / Annexe III	Emploi
		IA déterminant le licenciement, la prolongation ou la promotion d'un employé	Art 6 § 2 / Annexe III	
	Accès aux services privés et publics essentiels	IA utilisée pour déterminer l'éligibilité d'un individu à un programme social	Art 6 § 2 / Annexe III	Service public/ privé
		IA déterminant la solvabilité d'un individu avant de lui fournir un crédit	Art 6 § 2 / Annexe III	
		IA déterminant la répartition des tâches entre employés dans les secteurs publics essentiels (soins, sécurité, pompiers …)	Art 6 § 2 / Annexe III	
	Application de la loi	IA déterminant le risque de récidive d'un individu	Art 6 § 2 / Annexe III	Justice
		IA reprenant le fonctionnement d'un polygraphe ou déterminant l'état émotionnel de la personne interrogée	Art 6 § 2 / Annexe III	
		IA détectant les deepfakes	Art 6 § 2 / Annexe III	
		IA organisant une justice prédictive	Art 6 § 2 / Annexe III	
		IA déterminant des correlations cachées dans le secteur judiciaire		

Type de cas	Type d'IA	Sous-Type d'IA	Article concerné	Métrique
High risk	Gestion des flux migratoires, des demandes d'asile et de contrôle aux frontières	IA reprenant le fonctionnement d'un polygraphe ou déterminant l'état émotionnel de la personne interrogée	Art 6 § 2 / Annexe III	Migrations
		IA utilisée par une autorité publique pour évaluer le risque que provoquerait l'admission d'une personne étrangère sur un territoire donné	Art 6 § 2 / Annexe III	
		IA utilisée pour vérifier des documents officiels	Art 6 § 2 / Annexe III	
		IA utilisée pour examiner une demande d'asile	Art 6 § 2 / Annexe III	
	Administration de la justice et démocratie	IA utilisée pour aider la justice à appliquer la loi	Art 6 § 2 / Annexe III	Démocratie
Limited risk	Tous les autres types d'IA			

Présentation de la méthodologie utilisée pour la base de donnée du rapport (Annexe 2)

Les conclusions du présent rapport ont été tirées de l'exploitation d'une base de données recensant 567 incidents éthiques de l'IA [19]. Cet outil est issu d'une synthèse des travaux de Charles Pownall [20] et de l'AIID [21]. Pour aboutir à la base de donnée finale, certains incidents ont pu être supprimés (doublons, incidents très anciens, trop généraux).

Les métriques présentes sur la base sont de 3 natures ;

- créées par Charles Pownall et réutilisées sans modifications
- créées par Charles Pownall et retravaillées pour être adaptées au présent rapport
- créées spécialement pour le présent rapport

Les métriques créées par Charles Pownall

Les années

Les incidents de la base ont été classés par année. Celle-ci représente le moment où l'incident a été révélé au grand public, qui peut être différent de celui de la date de déroulement de l'incident lui-même.

Objectifs du système de l'IA

Métrique déterminant en quelques mots clés les principaux objectifs que doit remplir le système d'IA touché par l'incident. Cette métrique a une dimension descriptive pour mieux s'approprier l'incident mais n'a pas été réutilisée en l'état dans les études statistiques du présent rapport.

Technologie utilisée

Métrique déterminant le type de technologie utilisée par le système d'IA

[19] https://docs.google.com/spreadsheets/d/1277janRl3ZPCJSpz7dCbwDYt5zB3w0Wp2X2Ya03VL_A/edit#gid=0

[20] https://docs.google.com/spreadsheets/d/1Bn55B4xz21-_Rgdr8BBb2lt0n_4rzLGxFADMIVW0PYI/edit#gid=888071280 (consulté le 15 mai 2021)

[21] https://incidentdatabase.ai (consulté le 15 mai 2021)

(reconnaissance faciale, NLP …). Dans de nombreux cas, une valeur nulle a été appliquée à cette colonne en raison de la difficulté de catégoriser ces systèmes d'IA selon des technologies prédéterminées. Pour cette raison, cette métrique a été peu utilisée dans les études statistiques du présent rapport.

Type de controverse

Métrique classifiant l'incident selon le type de controverse qu'il concerne (par exemple la sécurité, les biais et discriminations, la transparence …). Cette métrique a été réutilisée en l'état dans le présent rapport, mais certaines valeurs (pseudoscience, hypocrisie …) à l'effectif particulièrement faible ou pouvant comporter un jugement moral envers l'acteur utilisant le système d'IA ont été écartées

Conséquences sur l'acteur à l'origine de l'incident (opérationnelles, financières, légales, réputationnelles)

Métrique recensant les conséquences concrètes qu'ont pu avoir les incidents sur les acteurs à leur origine (amende, banqueroute, licenciements …). Si elle n'a pas été directement réutilisée dans les études statistiques du présent rapport, elle a été indirectement employée pour la mise en place de la métrique ''réaction de l'acteur à l'origine de l'incident'' abordée plus bas.

Les métriques de Charles Pownall retravaillées

Secteur

Dans sa base de données, Charles Pownall classifie les incidents selon les secteurs touchés. Cependant, sa métrique comporte certaines limites auxquelles il a été tenté de remédier. La métrique ''Technology'' regroupait près d'un tiers de l'ensemble des incidents de la base, donc il a été décidé de créer de nouvelles variables pour regrouper certains incidents reconnaissables (par exemple ''social networks'' et ''search engine''). De plus, de nombreux cas ''Education/academia'' concernaient des incidents révélés par des universitaires, mais ne se rapportant pas forcément à des incidents éducatif. Ces cas ont été classés dans d'autres catégories plus adaptées. Enfin, les cas touchant le secteur public étaient classés selon de très nombreuses sous-catégories regroupant chacune très peu de cas (par exemple Govt-culture, Govt-immigrations …). Il a été décidé de créer une seule métrique ''Administration'' et de regrouper certains incidents dans les autres catégories existantes (notamment ''Justice'' ou ''Security/defense'')

Après recodage de cette métrique, ainsi que d'autres modifications mineures évoquées ci-dessous, on abouti à un ensemble de 19 catégories décrites dans le tableau suivant.

Métrique	Description
Administration	Les incidentes provoqués par une utilisation d'un système d'IA par les pouvoirs publics (par exemple pour déterminer l'attribution de subventions, de documents officiels …). Pour les incidents impliquant une reconnaissance faciale répressive ou de la justice prédictive, les catégories ''Security/defense'' et ''Justice'' ont été jugées plus pertinentes.
Advertisement	Les incidents provoqués par une utilisation d'un système d'IA dans la publicité internet
Automotive	Les incidents provoqués par des voitures autonomes
Banking/finance	Les incidents provoqués par une utilisation d'un système d'IA dans le domaine de la finance ou par des organismes bancaires. Les cas de crypto monnaies ont été regroupés dans cette catégorie
Business services	Les incidents provoqués par une utilisation d'un système d'IA au sein d'une entreprise pour améliorer sa productivité (par exemple, les algorithmes de recrutement, de gestion des employés …)
Consumer goods	Les incidents provoqués par une utilisation d'un système d'IA faisant partie d'un bien de consommation (par exemple une brosse à dent ou un réveil connecté)
Education/ academia	Les incidents provoqués par une utilisation d'un système d'IA dans le secteur éducatif ou académique (par exemple, la notation d'étudiants par une IA, leur surveillance dans les établissements par reconnaissance faciale …)
Healthcare	Les incidents provoqués par une utilisation d'un système d'IA dans la santé
Justice	Les incidents liés à la justice prédictive concernant tous les processus légaux autres que les interpellations (remise de peines, détermination des sentences …)
Leisure/ entertainment	Les incidents provoqués par une utilisation d'un système d'IA dans un but récréatif

Métrique	Description
Manufacturing	Les incidents ayant eu lieu dans des usines et impliquant en général des robots industriels
Politics	Les incidents provoqués par une utilisation d'un système d'IA ayant une dimension politique préméditée de la part de l'acteur à leur origine
Retail	Les incidents provoqués par une utilisation d'un système d'IA dans la livraison ou la vente de biens de consommation. Les incidents impliquant des livreurs humains sont exclus de cette catégorie et ont été regroupés dans ‘’Transports''
Search Engine	Les incidents liés à l'utilisation d'un moteur de recherche, ou intégré à une application (par exemple LinkedIn ou Youtube)
Security/defense	Les incidents provoqués par une utilisation d'un système d'IA dans un but répressif par des pouvoirs publics (police principalement) comme privés. La surveillance d'opposants et l'usage de reconnaissance faciale dans les établissements scolaires sont exclus de cette catégorie et regroupés dans ‘’Politics’’ et ‘’Education/ academia’’
Social networks	Les incidents liés à l'utilisation d'un réseau social. Tous les incidents impliquant des entreprises comme Twitter ou Facebook ne sont pas forcément inclus dans cette catégorie
Technology	Les incidents concernant les usages de l'IA non inclus dans les autres catégories mentionnées. Sont concernés principalement les systèmes d'IA au stade expérimental ou à l'usage trop large pour pouvoir être inclus dans les autres catégories
Transports	Les incidents impliquant la gestion de transports publics et privés, et les activités de logistique
Urbanism/ real estate	Les incidents concernant la gestion de zones urbaines et de leur marché immobilier, ainsi que les projets de villes intelligentes

Pays

Dans la base de Charles Pownall, les pays touchés par les incidents sont mentionnés. Cependant, pour rendre plus compréhensibles les résultats des analyses statistiques menées dans ce rapport, il a été décidé de les regrouper en aires géographiques pour gommer les plus faibles effectifs

Médias

La plupart des catégories utilisées dans le rapport concernant le média utilisé pour divulguer l'incident sont issus de la base de données de Charles Pownall. Il a été décidé de supprimer les catégories regroupant moins de 10 incidents et de les intégrer dans es catégories plus grandes les plus adaptées. Par exemple, la nuance entre médias locaux et nationaux a été supprimée dans la base finale.

Les métriques créées directement pour le rapport

Classification européenne

La pyramide des risques instaurée par la proposition de la Commission européenne a été appliquée aux cas de la base de données. Les catégories et sous-catégories utilisées sont explicitées dans le tableau de l'annexe 1 du présent rapport.

Gravité de l'incident

Les incidents de la base ont été notés sur une échelle de 1 à 5 pour évaluer la menace qu'ils constituaient pour les individus touchés. Cette notation reste subjective et peut être débattue dans certains cas de figures. Les justifications de chaque note donnée sont présentes dans la feuille 2 de la base de donnée. Le barème appliqué est le suivant :

Note	Description
1	Menace inexistante
2	Menace minime
3	Menace notable
4	Menace grave
5	Menace existentielle

Etendue de l'incident

Les incidents de la base ont été notés sur une échelle de 1 à 5 pour évaluer le nombre d'individus touchés. Cette notation reste également subjective et peut être débattue dans certains cas de figure. Les justifications de chaque note sont présentes dans la feuille 2 de la base de donnée.

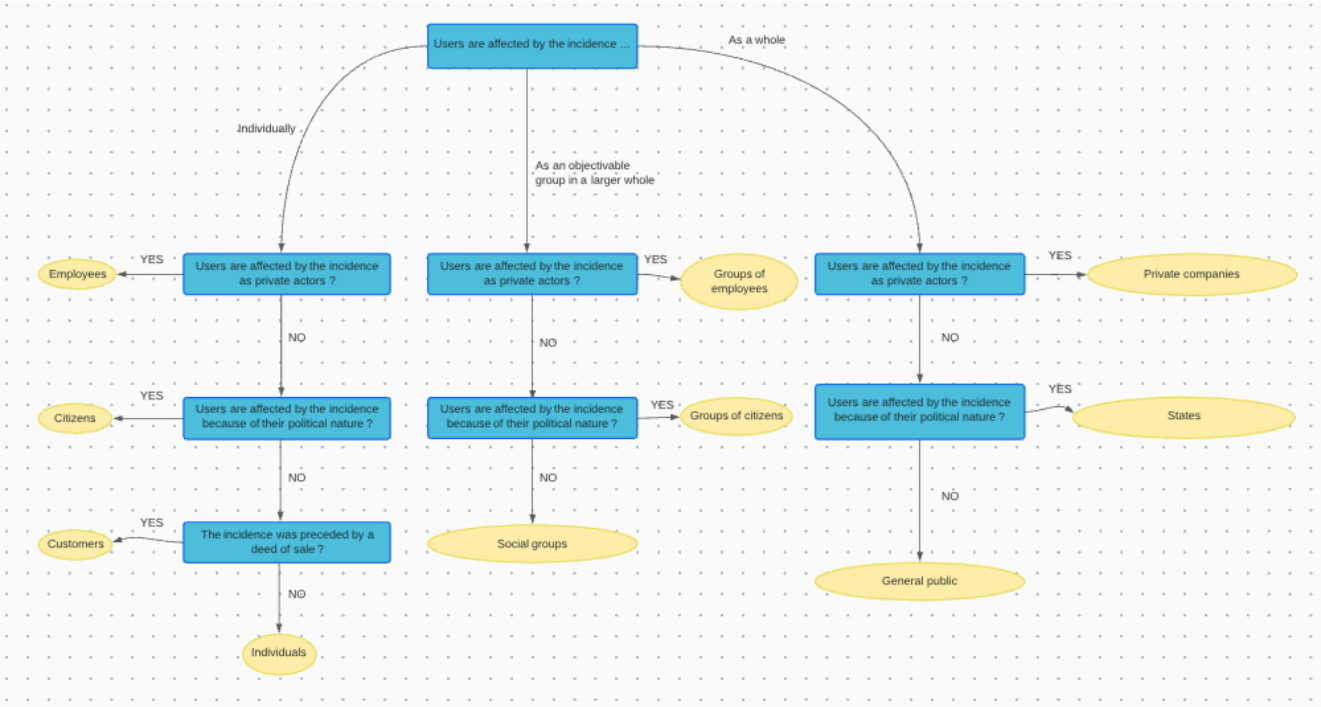
Le barème appliqué est le suivant :

Note	Description
1	1 seul individu touché
2	Une dizaine d'individus touchés
3	Un groupe important d'individus touchés
4	Un groupe de la taille de la population d'un Etat touché
5	L'ensemble de la population humaine susceptible d'être touchée

Type d'utilisateur touché

Les utilisateurs touchés par les incidents ont été catégorisés sur la base de donnée selon plusieurs dimensions (leur nombre, la dimension politique, sociale ou commerciale de leur rapport à l'acteur à l'origine de l'incident).

La classification de cette métrique suit la mindmap suivante :



Mindmap pour la catégorie ''Type d'utilisateur touché"

Réactions de l'acteur à l'origine de l'incident

Les incidents de la base ont été notés selon des indices allant de 1 à 5 pour caractériser le type de réaction adoptée par l'acteur à l'origine de l'incident une fois qu'il a été révélé. Cette notation se base sur certaines métriques de Charles Pownall et sur la lecture des articles de presse mis en lien pour chaque incident.
Le barème appliqué est le suivant :

Note	Description
1	Excuses totales —> l'acteur assume l'entière responsabilité de l'incident
2	Excuses mesurées —> l'acteur assume une partie de la responsabilité de l'incident
3	Absence de réactions
4	Déni —> l'acteur nie sa responsabilité dans l'incident
5	Revendication —> l'acteur assume la responsabilité de l'incident mais indique ne pas vouloir changer ses pratiques

Bibliographie

AI Ethics Impact Group, ‘’From Principles to practice, an interdisciplinary framework to operationalise AI ethics’’, AI Ethics Impact Group, 2020

Floyd Juliet, ‘’La quête culturelle : revisiter le test de Turing’’, Cités, vol 80, 2019, pp.15-30

Ganascia Jean-Gabriel et Powers Thomas, ‘’The Ethics of the Ethics of AI’’, The Oxford Handbook of Ethics of AI, 2020

Gueydier Pierre, ‘’Intelligence artificielle et travail des données’’, Revue d’éthique et de théologie morale, vol 307, 2020, pp.29-41

Hagendorff Thilo, ‘’The Ethics of AI Ethics : An Evaluation of Guidelines’’, Minds and Machines, 2020

Mittelstadt Brent, ‘’Principles alone cannot guarantee ethical AI’’, Nature Machine Intelligence, 2019

Nurock Vanessa, ‘’L’intelligence artificielle a-t-elle un genre ? ‘’, Cités, vol 80, 2019, pp.61-74

Thibaut Charles, ‘’La compétition mondiale de l’intelligence artificielle’’, Pouvoirs, vol 170, 2019, pp.131-142

Veale Michael et Zuiderveen Borgesius Frederick, ‘’Demystifying the Draft EU Artificial Intelligence Act’’, SocArXiv Papers, 2021

Villani Cédric, ‘’Les enjeux politiques de l’intelligence artificielle’’, Pouvoirs, vol 170, 2019, pp.5-18