unesco

Manuel de formation mondial : l'IA et l'état de droit pour le pouvoir judiciaire

Publié en 2024 par l'Organisation des Nations Unies pour l'éducation, la science et la culture
7, place de Fontenoy, 75352 Paris 07 SP, France

ISBN

Ce manuel de formation a été préparé par :

Dr Miriam Stankovich, Spécialiste principale des politiques numériques au Center for Digital Acceleration (Bethesda, Maryland, États-Unis).

La section sur les biais d'IA et l'égalité des sexes a été développée par Ivana Feldfeber (cofondatrice et directrice exécutive de DataGénero), Yasmín Quiroga (cofondateur de DataGénero et secrétaire du tribunal pénal n° 10 de Buenos Aires, Argentine) et Marianela Ciolfi Felice (professeure adjointe en design d'interaction à la KTH University, Suède, et conseillère chez DataGénero). La section sur les opportunités : l'IA et le pouvoir judiciaire sur le continent africain a été rédigée par le professeur Vukosi Marivate (Université de Pretoria, Afrique du Sud).

Conseillers académiques :

Prof. Joan Barata Mir (Senior Legal Fellow à Justitia, Danemark-États-Unis), Prof. Maria Fasli (Université d'Essex, Royaume-Uni), Prof. Els de Busser (Université de Leiden, Pays-Bas) et Prof. Vukosi Marivate (Université de Pretoria, Afrique du Sud).

Correcteurs de l'UNESCO :

Cedric Wachholz, Jaco Du Toit, Bhanu Neupane, Rosa María González, Natalia Zuazo, Misako Ito, Mehdi Benchelah.

Correcteurs externes :

Jhalak M. Kakkar (Directeur exécutif, Centre pour la gouvernance de la communication, Université nationale de droit de Delhi et professeur invité, NLU Delhi), Nidhi Singh (Chargé de programme, Centre pour la gouvernance de la communication, NLU Delhi), juge Jean Aloise Ndiaye (Cour suprême du Sénégal), Dr. Alexandre Barbosa (chef du Centre régional d'études sur le développement de la société de l'information, Cetic.br | NIC.br), Luiz Costa (Observatoire brésilien de l'intelligence artificielle, OBIA), Ameen Jauhar (chef d'équipe, ALTR, Centre Vidhi pour la politique juridique), Nathalie Smuha (professeure adjointe à la faculté de droit de la KU Leuven et boursière Emile Noël à la faculté de droit de l'Université de New York), Lee Tiedrich (membre distingué du corps professoral, technologie éthique à l'Université Duke et GPAI et expert en IA de l'OCDE), Marc Rotenberg (président et fondateur du Centre pour l'IA et la politique numérique), Alfonso Peralta Gutiérrez (juge de première instance et enquête criminelle, Grenade, Espagne), Murali Sagi (directeur général adjoint de la Commission judiciaire de NSW, Nouvelle-Galles du Sud), Anthony Wong (président de l'IFIP, Fédération internationale pour le traitement de l'information), Saurabh Karn (fondateur et scientifique principal chez OpenNyAI et fondateur de Sampatti Card) et Prof. Keith R. Fisher (Distinguished Fellow, National Judicial College, États-Unis), Niki Iliadis (Director, AI and the Rule of Law at TFS, The Future Society), Amanda Leal (Associate, AI Governance at TFS), Nicolas Miailhe (Founder & President of TFS), Prof. Srikrishna Deva Rao (vice-chancelier de l'Université de droit NALSAR, Hyderabad), M. Pranav Verma (professeur adjoint à la National Law School of India University, Bengaluru), Dr. Ravi Srinivas (professeur adjoint à l'Université de droit NALSAR).

Dr Naveen Thayyil (professeur agrégé à l'IIT, Delhi), Neela Badami (associée chez Samvad Partners), Dr Shouvik Kumar Guha (professeur agrégé à l'Université nationale des sciences juridiques du Bengale occidental, Kolkata), Rohan George (associé chez Samvad Partners), Nehaa Chaudhari (associée chez Ikigai Law), Pallavi Sondhi (associé principal chez Ikigai Law), Ajey Karthik (associé chez Ikigai Law) et Namratha Murugeshan (associée chez Ikigai Law), Jaideep Reddy (avocat spécialisé en technologie chez Trilegal et professeur invité à la National Law School of India University, Bengaluru).

Direction et coordination de projet :

Prateek Sibal, spécialiste de programme, Politiques numériques et transformation numérique, UNESCO.
Charline d'Oultremont, consultante, Politiques numériques et transformation numérique, UNESCO.
Giovanni Imperiali, stagiaire, Politiques numériques et transformation numérique, UNESCO.
Gustavo Fonseca Ribeiro, consultant en politiques numériques et transformation numérique à l'UNESCO, a contribué à l'organisation d'ateliers pilotes pour le manuel de formation.

Graphisme de la couverture : Nube Consulting
Composition : Nube Consulting + Aliens

Imprimé par : UNESCO

Le manuel de formation mondial a été développé dans le cadre du projet financé par la Commission européenne « Aider les États membres à mettre en œuvre la recommandation de l'UNESCO sur l'éthique de l'IA grâce à des outils innovants ».

BREF RÉSUMÉ

L'intelligence artificielle : une nouvelle frontière pour le pouvoir judiciaire

Qu'est-ce que l'intelligence artificielle (IA) ? Comment ça marche ? Et plus important encore, comment trouve-t-elle sa place dans un contexte judiciaire ? Des technologies telles que l'IA existent depuis des décennies, mais elles n'ont commencé à être utilisées dans divers cadres judiciaires et exécutifs que très récemment. Bien que l'IA ait un immense potentiel pour le système judiciaire, pour aider les juges à prendre de meilleures décisions, améliorer leur efficacité, l'accessiblité, et pour aider à détecter et à prévenir la criminalité, elle soulève également des questions importantes parmi les acteurs judiciaires, alors qu'ils se préparent à un avenir où l'IA sera de plus en plus utilisée dans les systèmes judiciaires.

En 2022, l'UNESCO a lancé deux évaluations des besoins. Tout d'abord, dans le cadre de l'enquête de l'UNESCO sur l'évaluation des besoins en matière d'intelligence artificielle en Afrique, 90 % des 32 pays interrogés ont demandé un soutien au renforcement des capacités du pouvoir judiciaire en matière d'IA. Dans le même temps, une deuxième enquête mondiale sur les acteurs judiciaires dans 100 pays a mis en avant la nécessité de mieux comprendre l'utilisation de l'IA dans l'administration de la justice et ses implications juridiques plus larges sur les sociétés.

Cet ouvrage, « Manuel de formation mondial : l'IA et l'état de droit pour le pouvoir judiciaire », répond à ces besoins et fournit aux acteurs judiciaires (juges, procureurs, avocats, juristes, facultés de droit et établissements de formation judiciaire) les connaissances et les outils nécessaires pour comprendre les avantages et les risques de l'IA dans leur travail. Ce manuel de formation aidera les acteurs judiciaires à atténuer les risques potentiels de l'IA pour les droits humains en fournissant des conseils sur les lois, principes, règles et jurisprudences internationales pertinents en matière de droits humains qui sous-tendent l'utilisation éthique de l'IA.

unesco

« Les guerres prenant naissance dans l'esprit des hommes, c'est dans l'esprit des hommes que doivent être élevées les défenses de la paix. »

Manuel de formation mondial : l'IA et l'état de droit pour le pouvoir judiciaire

AVANT-PROPOS

Les juges jouent un rôle crucial dans la protection des droits civils : ils sont créateurs de jurisprudence lors d'affaires individuelles, ce qui permet aux pays de progresser dans un domaine particulier du droit. Des affaires récentes ont montré que le pouvoir judiciaire peut s'appuyer sur le droit international relatif aux droits humains, les garanties constitutionnelles et les lois sur la protection des données, pour se prémunir contre les systèmes d'IA discriminatoires et biaisés. Pour que les juges jouent efficacement ce rôle déterminant, nous devons les aider à renforcer leurs connaissances et leur compréhension du fonctionnement des systèmes d'IA et de l'application possible du droit international relatif aux droits humains à ces systèmes.

Depuis 2014, l'Initiative mondiale des juges de l'UNESCO implique plus de 34 800 acteurs judiciaires de plus de 160 pays autour de thèmes tels que la liberté d'expression, l'accès à l'information et la sécurité des journalistes. Cette initiative contribue à renforcer la capacité des opérateurs judiciaires à relever les défis émergents pour le pouvoir judiciaire et à protéger les droits humains fondamentaux ainsi que la liberté d'expression.

En 2022, l'Initiative mondiale des juges a lancé un programme sur l'IA et l'état de droit, dans le but d'engager les parties prenantes des systèmes judiciaires dans une discussion mondiale et opportune sur les applications de l'intelligence artificielle et son impact sur l'état de droit. Cela fait suite à la Recommandation de l'UNESCO sur l'éthique de l'intelligence artificielle, un plan détaillé pour la mise en place de régimes réglementaires fondés sur des valeurs et des principes universellement acceptés, adoptée par les 193 États membres de l'UNESCO, en novembre 2021. La recommandation a mis en avant la valeur des « systèmes d'IA pour améliorer l'accès à l'information et aux connaissances » et la nécessité de « renforcer la capacité du pouvoir judiciaire à prendre des décisions relatives aux systèmes d'IA conformément à l'état de droit et aux normes et au droit internationaux ».

À la suite d'une enquête mondiale impliquant des acteurs judiciaires du Réseau mondial des anciens élèves de l'Initiative des juges, l'UNESCO et ses partenaires ont mis au point un cours en ligne ouvert à tous sur l'IA et l'état de droit (MOOC) en sept langues, en 2022. Le MOOC présente les bonnes pratiques sur la façon dont les tribunaux statuent sur les affaires liées à l'IA, conformément aux droits humains et aux normes éthiques, et explore les opportunités et les risques de l'adoption de l'IA par les systèmes judiciaires.

Dans la lignée de ce MOOC, le « Manuel de formation mondial : l'IA et l'état de droit pour le pouvoir judiciaire » vise à former les acteurs judiciaires sur la manière de veiller à ce que le développement de l'IA atteigne son plein potentiel conformément à l'état de droit. En effet, alors que nous nous efforçons d'élaborer de nouvelles lois pour régir l'IA elle-même, il est impératif que nous soutenions les juges, les procureurs et les fonctionnaires dotés de capacités accrues pour nous protéger des risques liés à l'IA.

Table des matières

Liste des acronymes 14

Pourquoi ce manuel de formation ? 16

Glossaire 20

Module 1 - Introduction à l'IA et à l'état de droit 24

1. Comprendre l'IA et ses éléments constitutifs 25
2. Pourquoi les données sont-elles importantes dans le contexte de l'IA ? 36
3. Les systèmes d'IA en tant que « boîtes noires » 39
4. Principe de l'humain dans la boucle 43
5. Pourquoi la cybersécurité est-elle importante dans le contexte de l'IA ? 46
6. Activités 49
7. Ressources 52

Module 2 - L'adoption de l'IA dans le système judiciaire 54

1. Quelles sont les applications de l'IA dans le système judiciaire ? 55
2. Études de cas sur le déploiement de l'IA dans le système judiciaire 78
3. Activités 83
4. Ressources 86

Module 3 - Défis juridiques et éthiques de l'IA 88

1. Qu'est-ce que l'éthique de l'IA ? 89
2. Qu'est-ce que le biais d'IA ? 94
3. Pourquoi la transparence et la responsabilité algorithmique sont-elles importantes dans les contexte du pouvoir judiciaire ? 109
4. Pleins feux sur l'identification biométrique, la technologie de reconnaissance faciale et les deepfakes 113
5. Activités 122
6. Ressources 127

Module 4 - Les droits humains et l'IA 128

1. Introduction aux droits humains et à l'IA 129
2. Sélectionner les droits humains impactés par le déploiement de l'IA 136
3. Approches de la gouvernance de l'IA 184
4. Activités 191
5. Ressources 194

Ressources suggérées par l'UNESCO 196

Comment utiliser ce manuel de formation ? 199

Annexe I - Évaluation de l'impact éthique de l'UNESCO pour les systèmes d'IA 202

Annexe II - Exemples d'activités supplémentaires 204

Annexe III - Programme de formation – modèle 207

Liste des acronymes

ACLU	Union américaine pour les libertés civiles
ADM	Prise de décision automatisée
AGR	Reconnaissance automatique du genre
CAHAI	Comité ad hoc du Conseil de l'Europe sur l'intelligence artificielle
CE	Commission européenne
CEDH	Convention européenne des droits de l'homme
ChatGPT	Transformateur pré-entraîné génératif
COMPAS	Profilage de la gestion des délinquants correctionnels pour les sanctions alternatives
DUDH	Déclaration universelle des droits de l'homme
EFF	Electronic Frontier Foundation (ONG)
ESI	Informations stockées électroniquement
FAIR	Trouvable, accessible, interopérable et réutilisable
FTC	Commission fédérale du commerce
GAN	Réseaux antagonistes génératifs
HUDERAF	Droits de l'homme, démocratie et cadre d'assurance de l'état de droit
IA	Intelligence artificielle
IdO	Internet des objets
IFIP	Fédération internationale pour le traitement de l'information
ISO	Organisation internationale de normalisation
LAPD	Service de police de Los Angeles
LLM	Modèle de langage de grande taille
MIT	Institut de technologie du Massachusetts

ML	Apprentissage automatique
NDAS	Solution nationale d'analyse de données
NIST	Institut national des normes et de la technologie
OCDE	Organisation de coopération et de développement économiques
ONG	Organisation non gouvernementale
ONU	Organisation des Nations Unies
PIDCP	Pacte international relatif aux droits civils et politiques
PIDESC	Pacte international relatif aux droits économiques, sociaux et culturels
RGPD	Règlement général sur la protection des données
SPC	Cour populaire suprême chinoise
STF	Cour suprême brésilienne
SUPACE	Portail de la Cour suprême (Inde) pour l'aide à l'efficacité des tribunaux
TAL	Traitement automatique du langage naturel
TAR	Examen assisté par la technologie
TRC	Tribunal de résolution civile
TRF	Technologie de reconnaissance faciale
UCL	Université catholique de Louvain
UE	Union européenne
UIT	Union internationale des télécommunications
UNESCO	Organisation des Nations Unies pour l'éducation, la science et la culture

POURQUOI CE MANUEL DE FORMATION ?

Ce manuel de formation fournit aux opérateurs judiciaires les connaissances et les outils nécessaires pour comprendre les avantages et les risques de l'intelligence artificielle (IA) dans leur travail. Le manuel de formation aidera les opérateurs judiciaires à atténuer les risques potentiels de l'IA pour les droits humains, en fournissant des conseils sur les instances, les principes et les règlements pertinents du droit international relatif aux droits humains, ainsi que sur la jurisprudence émergente qui sous-tend l'utilisation responsable de l'IA.

Le manuel de formation répond à la recommandation de l'UNESCO sur l'éthique de l'IA, adoptée par 193 pays en 2021, qui recommande que « les États membres renforcent la capacité du pouvoir judiciaire à prendre des décisions relatives aux systèmes d'IA conformément à l'état de droit… ».

Qu'allez-vous apprendre ?

Après avoir étudié le manuel de formation, les opérateurs judiciaires pourront :

- Définir l'IA et la prise de décision automatisée (ADM), et les comprendre comme des systèmes socio-techniques.
- Comprendre les problèmes clés liés aux préjugés algorithmiques et à la discrimination (tels que les préjugés de genre, les préjugés raciaux et d'autres formes de préjugés qui se recoupent), et expliquer pourquoi ils sont importants dans les contextes judiciaires.
- Comprendre et expliquer l'impact de l'IA sur les droits fondamentaux suivants : vie privée, liberté d'expression, accès à l'information, protection contre la discrimination, droit d'accès au tribunal, procès et audiences équitables et impartiaux, et procédure régulière.
- Examiner les affaires juridiques liées à l'utilisation de l'IA, en s'appuyant sur leur connaissance des initiatives réglementaires récentes et de la jurisprudence relative aux biais algorithmiques, à l'utilisation inappropriée des algorithmes dans la prise de décision.
- Appliquer des outils tels que l'évaluation de l'impact éthique de l'UNESCO pour comprendre l'impact éthique des systèmes d'IA.

Le manuel de formation comprend quatre modules qui complètent un programme de formation sur l'IA, les droits humains et l'état de droit pour le pouvoir judiciaire. Ce manuel de formation fournit également les connaissances nécessaires non seulement aux juges, mais également aux autres acteurs impliqués dans le processus de règlement des différends, y compris les avocats et les arbitres.

- ***Module 1 : Introduction à l'IA et à l'état de droit***

  Le module 1 présente au lecteur les principaux concepts liés à la gouvernance algorithmique, aux droits humains et à l'état de droit dans le contexte du développement de l'IA. Le module définit des termes tels que l'IA, les algorithmes,

les systèmes algorithmiques, et décrit leurs caractéristiques clés et leurs éléments constitutifs. Le module 1 traite également de l'importance des données et de la cybersécurité dans le contexte de l'IA, et donne un aperçu des principaux risques associés à l'IA, tels que les boîtes noires.

- ***Module 2 : Adoption de l'IA dans le système judiciaire***

  Le module 2 traite de l'adoption de l'IA dans le système judiciaire. Cette session décrit les utilisations de l'IA dans le système judiciaire, telles que la découverte électronique et l'examen des documents, l'utilisation de l'IA générative pour aider à la rédaction de documents, l'analyse prédictive et le soutien de l'ADM, les outils d'évaluation des risques, le règlement des différends, la reconnaissance et l'analyse linguistiques, les fichiers numériques et la gestion des cas. Le module met ensuite en évidence des études de cas sur le déploiement de l'IA dans le système judiciaire de différents pays, et décrit les opportunités et les défis liés à ces cas d'utilisation.

- ***Module 3 : Défis juridiques et éthiques du déploiement de l'IA***

  Le module 3 présente les principaux défis juridiques et éthiques liés à l'IA dans le système judiciaire, et résume les questions juridiques liées à l'identification biométrique et à la technologie de reconnaissance faciale. Le module 3 aborde en détail les défis liés à l'IA et à l'éthique sur la base de la Recommandation de l'UNESCO 2021 sur l'éthique de l'intelligence artificielle.[1]

- ***Module 4 : Droits humains et IA***

  Le module 4 présente une analyse approfondie des droits humains touchés par l'IA, tels que (i) le droit d'accéder à un tribunal, à un procès équitable et à une procédure régulière, (ii) le droit à un recours effectif, (iii) le droit à la protection contre la discrimination, (iv) la liberté d'expression et l'accès à l'information, et (v) le droit à la vie privée et à la protection des données. Le module 4 donne également un aperçu des principales approches de gouvernance de l'IA, fondées sur les risques et sur les droits humains.

À qui s'adresse ce manuel de formation ?

Le principal public visé par ce manuel de formation sont les juges, les procureurs, les avocats d'État, les avocats publics, les universités de droit et les établissements de formation judiciaire.

Comment utiliser ce manuel de formation pour enseigner ?

Ce manuel peut être adapté aux besoins spécifiques de chaque programme de formation judiciaire. Le nombre d'heures et la durée du programme de formation dépendront de la méthodologie choisie par le programme de formation judiciaire. Le programme peut être enseigné comme un programme d'apprentissage en ligne, en classe ou hybride, et il peut être offert comme un cours intensif ou régulier d'un programme de premier cycle, de troisième cycle ou de formation continue, en fonction de la disponibilité des formateurs et/ou de la répartition géographique des apprenants inscrits pour un cours spécifique, et du niveau d'accessibilité et de connectivité.

Il est préférable d'enseigner le programme comme un effort organisé pour transférer les connaissances et développer les compétences et les attitudes qui encouragent les

1 UNESCO (2021). Recommendation on the Ethics of Artificial Intelligence, disponible sur : https://unesdoc.unesco.org/ark:/48223/pf0000380455

actions axées sur la promotion et la protection des droits humains en relation avec l'IA. Par conséquent, les éléments suivants sont recommandés pour toute formation basée sur le manuel de formation :

- Transfert de connaissances : Dans le contexte de ce manuel de formation, le terme « connaissances » fait référence aux normes relatives aux droits humains et aux mécanismes de protection qui sont pertinents pour l'IA, pour le groupe cible d'apprenants. Par exemple, dans le contexte d'un cours adressé à des juges, la connaissance peut se référer aux normes relatives aux droits humains pour décider des affaires impliquant l'utilisation de l'IA.

- Développement des compétences : Une compréhension de base des normes applicables en matière de droits humains peut être insuffisante pour permettre aux apprenants de traduire ces normes en comportement réel. Les capacités sont affinées par la pratique, l'application et la réflexion. C'est un processus qui peut être initié au cours de la formation par le biais de diverses activités, mais qui peut devoir être poursuivi après le cours de formation, y compris par le biais de programmes de suivi correctement planifiés. Par exemple, la capacité de mener une évaluation des risques des systèmes d'IA pour déterminer s'ils doivent être déployés en premier lieu, plutôt que d'assumer le déploiement et de tenter ensuite d'atténuer les préjudices ex post.

- Développement des attitudes : Cela implique l'acquisition et le renforcement d'attitudes positives envers les droits humains et la primauté du droit, afin que les apprenants prennent des mesures pour promouvoir et protéger les droits humains dans leur vie quotidienne et leurs responsabilités professionnelles dans le jugement des violations des droits humains impliquant les processus d'ADM et d'IA.[2]

Le contenu de la formation a été mis à disposition en ligne en tant que ressource ouverte et peut être mis à jour régulièrement en créant un référentiel en ligne de présentations que les formateurs peuvent consulter et réutiliser sous licences ouvertes Creative Commons (Attribution 4.0 International).[3]

2 Tout changement conduisant à un meilleur respect des droits humains - changements au niveau des apprenants individuels, de leur organisation/groupe et de la communauté/société au sens large - qui peut être attribué de manière plausible à l'effort de formation doit être pris en compte dans une évaluation de l'impact de la formation.

3 Voir : https://creativecommons.org/licenses/by/4.0/

Glossaire

- **Agrégation de données :** Collecte d'une quantité importante d'informations à partir d'une base de données et présentation dans un format plus facile à gérer.
- **Algorithme :** Série d'instructions pour effectuer des calculs ou d'autres tâches, que ce soit en mathématiques ou en informatique. Dans le cas de l'IA, un algorithme fournit les instructions qui permettent à un ordinateur d'apprendre à apprendre de son environnement et d'effectuer un ensemble de tâches.
- **Analyse prédictive :** Catégorie générale d'outils et de modèles statistiques, par exemple les systèmes de ML, qui utilisent et analysent des données historiques pour créer des prédictions sur l'avenir, afin de guider la prise de décision. Ces prédictions peuvent être à faible risque (recommandation de film), à risque moyen (acceptation de demande de prêt) ou à risque élevé (prédiction du défendeur le plus susceptible d'adopter un comportement particulier).
- **Apprentissage automatique (ML) :** Ensemble de techniques qui permettent aux machines d'apprendre automatiquement en utilisant des modèles et des déductions, plutôt que des instructions directes d'une personne. Les techniques de ML demandent souvent aux machines d'arriver à un résultat en fournissant de nombreuses instances de résultats corrects. Cependant, elles peuvent également spécifier un ensemble de directives et laisser la machine les découvrir seule dans les données.
- **Apprentissage automatique supervisé :** Processus qui consiste à fournir à un système d'apprentissage automatique un ensemble de données déjà étiquetées ou classifiées, que le système peut utiliser pour apprendre à effectuer une tâche particulière avec précision, selon les instructions données. Le système de ML est chargé avec un ensemble de données et le résultat attendu. Dans la phase de formation, le modèle de ML ajuste ses variables pour connecter les entrées à la sortie correspondante. La création d'un algorithme d'apprentissage supervisé réussi requiert une équipe de spécialistes investis, pour évaluer et examiner les résultats. Cela implique des scientifiques des données qui examinent minutieusement les modèles produits par l'algorithme pour vérifier leur précision par rapport aux données sources et identifier toute inexactitude causée par l'IA.
- **Biais d'IA :** Différence systématique dans le traitement de certains objets, personnes ou groupes (par exemple, stéréotypes, préjugés ou favoritisme) par rapport à d'autres, par les algorithmes d'IA.
- **Datafication :** Processus qui désigne la prolifération des outils numériques utilisés pour intégrer, analyser et afficher des modèles de données.
- **Deepfake :** Toute forme de média (vidéo, audio ou autre) qui a été modifié, ou entièrement ou partiellement créé à partir de zéro.

- **Discrimination par procuration :** Dans les systèmes d'IA, se produit lorsqu'une caractéristique apparemment neutre est substituée par une caractéristique interdite.
- **Étiquetage des données :** Dans l'apprentissage automatique (ML) des données, processus de reconnaissance de données brutes (images, fichiers texte, vidéos, etc.), auxquelles on ajoute une ou plusieurs étiquettes pertinentes et utiles pour offrir un contexte à un modèle de ML, afin qu'il en tire des leçons. Les étiquettes peuvent indiquer si une photographie contient un oiseau ou une automobile, si des mots ont été prononcés dans un enregistrement audio ou si une radiographie montre une tumeur. De nombreux cas d'application, notamment la vision par ordinateur, le traitement du langage naturel et la reconnaissance vocale, nécessitent un étiquetage des données.
- **Fiducie de données :** Organisation indépendante qui agit en tant que fiduciaire pour les fournisseurs de données et réglemente la bonne utilisation de leurs données.
- **Humain dans la boucle (HITL) :** Processus dans lequel un système d'IA est étroitement surveillé par un humain, qui est responsable de prendre toutes les décisions finales. Ceci est particulièrement important dans des domaines comme les soins de santé, où l'IA peut fournir un soutien inestimable dans la formulation de recommandations pour le traitement du cancer, le traitement de la septicémie, la planification chirurgicale, etc. Bien que les outils d'IA puissent aider les prestataires de soins de santé à prendre des décisions éclairées de manière rapide et précise, la responsabilité ultime des soins aux patients incombe toujours à l'expert humain.
- **IA en tant que « boîte noire » :** Le terme « boîte noire » désigne un système technologique intrinsèquement opaque, dont le fonctionnement interne ou la logique sous-jacente ne sont pas correctement compris, ou dont les résultats et les effets ne peuvent être expliqués.
- **IA explicable (XAI) :** Systèmes, algorithmes et modèles capables d'expliquer la raison d'être de leurs décisions, de caractériser les forces et les faiblesses de leur processus décisionnel et de comprendre comment ils se comporteront à l'avenir.
- **IA générative :** IA composée d'algorithmes d'apprentissage automatique (ML) conçus pour créer du nouveau contenu, notamment de l'audio, du code, des images, du texte, des simulations et des vidéos.
- **Modèle de diffusion :** Modèle génératif plus avancés que les réseaux antagonistes génératifs (voir définition ci-dessous) sur la synthèse d'images. Plus récemment, les modèles de diffusion ont été utilisés dans DALL-E 2, le modèle de génération d'images d'OpenAI et Imagen de Google.
- **Prise de décision automatisée (ADM) :** Utilisation de « résultats produits par des algorithmes pour prendre des décisions ».
- **Réseau neuronal :** Type de technique de ML qui permet aux ordinateurs d'apprendre à effectuer des tâches en analysant des exemples d'entraînement. En règle générale, ces exemples sont pré-étiquetés. Par exemple, un système de reconnaissance d'objets peut recevoir des milliers d'images étiquetées d'objets tels que des voitures, des maisons et des tasses à café. Grâce à l'analyse, il peut identifier des motifs dans les images qui correspondent à des étiquettes spécifiques. Un réseau neuronal est conçu pour ressembler vaguement à la structure du cerveau humain, avec des

milliers ou des millions de nœuds de traitement interconnectés. Ces nœuds sont généralement organisés en couches et les données les traversent dans une seule direction, ce qui les rend « feed-forward ». Chaque nœud reçoit des données des nœuds de la couche située en dessous et envoie des données aux nœuds de la couche située au-dessus.

- **Réseaux antagonistes génératifs (GAN) :** Approche non supervisée de l'apprentissage profond qui peut générer du matériel hyperréaliste. Les GAN sont utilisés pour des techniques d'apprentissage en profondeur non supervisées, telles que la génération d'images réalistes ou d'ensembles de données d'image, la traduction de texte en image et d'image en texte, le vieillissement des visages et la création d'émojis.
- **Sandbox réglementaire :** Outil réglementaire permettant aux entreprises de tester et d'expérimenter des produits, des services ou des commerces nouveaux et innovants, sous la supervision d'un régulateur, pendant une période limitée.
- **Traitement du langage naturel (TAL) :** Technique de ML qui analyse de grandes quantités de données textuelles ou vocales humaines (transcrites ou acoustiques) pour des propriétés spécifiques, telles que le sens, le contenu, l'intention, l'attitude et le contexte.
- **Valeur de hachage :** Valeur renvoyée par une fonction de hachage, utilisée pour convertir des données numériques de taille arbitraire en une chaîne de sortie avec un nombre de caractères de taille fixe.


Technology
Innovation
SYSTEM
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec

Module 1

Introduction à l'IA et à l'état de droit

Le module 1 présente la gouvernance algorithmique, les droits humains et l'état de droit. Il discute des définitions de l'IA, des algorithmes et des systèmes algorithmiques, en décrivant leurs caractéristiques clés et leurs éléments constitutifs. Ce module souligne l'importance des données et de la cybersécurité dans le contexte du déploiement de l'IA dans le système judiciaire. Il donne un aperçu des principaux risques associés au déploiement de l'IA dans le système judiciaire, tels que les boîtes noires, et explique le principe de l'humain dans la boucle.

Qu'allez-vous apprendre ?

Après avoir terminé ce module, les participants seront en mesure de :

- Comprendre et expliquer les concepts clés liés à l'IA, à la gouvernance algorithmique et à l'état de droit.
- Définir et expliquer l'IA, les algorithmes, les systèmes algorithmiques, en décrivant leurs caractéristiques clés et leurs éléments constitutifs.
- Comprendre et reconnaître les risques associés à l'IA, tels que les boîtes noires et la cybersécurité.
- Comprendre l'importance du principe de l'humain dans la boucle dans le cycle de vie de l'IA.
- Comprendre pourquoi les données sont importantes dans le contexte de l'IA.

Qu'est-ce qu'un système d'IA ?

Selon l'UNESCO, les systèmes d'IA sont des systèmes qui ont la capacité de traiter les données et les informations d'une manière qui ressemble à un comportement intelligent, comprenant généralement des aspects de raisonnement, d'apprentissage, de perception, de prédiction, de planification ou de contrôle.[4] En d'autres termes, les systèmes d'IA sont des technologies de traitement de l'information qui intègrent des modèles et des algorithmes qui produisent une capacité d'apprentissage et d'exécution de tâches cognitives menant à des résultats tels que la prédiction et la prise de décision, dans des environnements matériels et virtuels. Les systèmes d'IA sont conçus pour fonctionner avec différents degrés d'autonomie au moyen de la modélisation et de la représentation des connaissances, et en exploitant les données et en calculant les corrélations. Les systèmes d'IA peuvent inclure plusieurs méthodes, telles que (sans s'y limiter) :

- l'apprentissage automatique, notamment l'apprentissage profond et l'apprentissage par renforcement ;
- le raisonnement machine, notamment la planification, l'ordonnancement, la représentation et le raisonnement des connaissances, la recherche et l'optimisation.

Il est important de noter qu'une telle définition doit s'affiner au fil du temps, pour accompagner les évolutions technologiques. En outre, le terme « IA » est souvent utilisé de manière interchangeable avec « apprentissage automatique » (ML), alors que l'IA est un domaine beaucoup plus large qui va bien au-delà du ML, comme la représentation des connaissances, la planification et le raisonnement.[5]

En plus de la description ci-dessus, le tableau 1 présente un aperçu de la façon dont différentes organisations définissent l'IA de manière pragmatique, en fonction de l'ensemble des tâches ou des fonctions que la technologie peut entreprendre (OCDE, ISO), ou en fonction des idéaux humanistes qu'elles cherchent à imprégner dans toutes sortes de systèmes axés sur les données pour s'assurer qu'ils contribuent à l'amélioration de la société (CE, UIT).

4 UNESCO (2021). Recommendation on the Ethics of Artificial Intelligence, disponible sur :https://unesdoc.unesco.org/ark:/48223/pf0000380455

5 OCDE (2019). Artificial Intelligence in Society, disponible sur : https://www.oecd.org/publications/artificial-intelligence-in-society-eedfee77-en.htm; Leslie D., Burr C., Aitken M., Cowls J., Katell M., and Briggs, M. (2021). Artificial intelligence, human rights, democracy, and the rule of law: a primer, Le Conseil de l'Europe, disponible sur : https://ssrn.com/abstract=3817999 ou http://dx.doi.org/10.2139/ssrn.3817999

Tableau 1. Définitions de l'IA dans les organisations internationales et multilatérales

Organisme	Définition de l'IA
OCDE[6]	L'IA est un système basé sur une machine qui peut, pour un ensemble donné d'objectifs définis par l'homme, fournir des prédictions, des recommandations ou des décisions influençant des environnements réels ou virtuels. Lorsqu'elle est appliquée, l'IA a sept cas d'utilisation différents, également appelés modèles, qui peuvent coexister au sein du même système d'IA.
ISO[7]	Système d'ingénierie qui génère des résultats tels que du contenu, des prévisions, des recommandations ou des décisions pour un ensemble donné d'objectifs définis par l'homme.
EC[8]	L'IA comprend des systèmes qui affichent un comportement intelligent en analysant leur environnement et en prenant des mesures - avec un certain degré d'autonomie - pour atteindre des objectifs spécifiques.
ITU[9]	L'IA fait référence à la capacité d'un ordinateur ou d'un système robotique informatisé à traiter des informations et à produire des résultats d'une manière similaire au processus de pensée des humains dans l'apprentissage, la prise de décision et la résolution de problèmes. D'une certaine manière, l'objectif des systèmes d'IA est de développer des systèmes capables de s'attaquer à des problèmes complexes de manière similaire à la logique et au raisonnement humains.

Les systèmes d'IA dans notre vie quotidienne

L'IA fait déjà partie de notre quotidien, que nous nous en rendions compte ou non. Dans votre boîte de réception, par exemple : certains e-mails se retrouvent dans le dossier de spams, tandis que d'autres sont classés dans les catégories « social » ou « promotion ». Comment est-ce possible ? Savez-vous que Google met en œuvre des algorithmes d'IA pour catégoriser et filtrer automatiquement les e-mails ? Ces algorithmes sont des programmes formés pour identifier des éléments spécifiques qui indiquent qu'un e-mail est peut-être un spam. Lorsque l'algorithme reconnaît ces éléments, il marque l'e-mail et le déplace vers le dossier correspondant. Les algorithmes sont constamment améliorés. Si vous trouvez un e-mail légitime dans votre dossier spams, vous pouvez informer Google qu'il a été étiqueté à tort comme spam. Ces commentaires aident à améliorer la précision de l'algorithme.[10]

Les chatbots des services client sont un autre exemple d'IA présents dans nos interactions quotidiennes. Lorsque vous tapez votre question, le chatbot utilise un algorithme pour reconnaître les mots-clés et déterminer l'assistance dont vous avez besoin. Sur la base des informations existantes et nouvellement acquises, le modèle d'apprentissage automatique génère une réponse appropriée. Plus le chatbot interagit avec des clients, plus il reçoit des données supplémentaires et peut s'améliorer.[11] Parmi d'autres exemples, nous pouvons citer le moteur de recommandation de Netflix qui suggère films et émissions de télévision en fonction de vos préférences, ou encore des assistants vocaux, comme Siri et Alexa, qui traitent des requêtes simples.

6 OCDE (2019). Artificial intelligence and responsible business conduct, disponible sur : https://mneguidelines.oecd.org/RBC-and-artificial-intelligence.pdf
7 ISO (2021). ISO/IEC DIS 22989, disponible sur : www.iso.org/standard/74296.html
8 Commission européenne (2018). Communication Artificial Intelligence for Europe, disponible sur : https://digital-strategy.ec.europa.eu/en/library/communication-artificial-intelligence-europe
9 UIT (2018). Policy Considerations for AI Governance, disponible sur : www.itu.int/en/ITU-T/studygroups/2017-2020/03/Documents/Shailendra%20Hajela_Presentation.pdf
10 Voir : https://dig.watch/technologies/artificial-intelligence
11 Bravo K. (2023). How Does AI actually work?, disponible sur : https://blog.mozilla.org/en/internet-culture/how-does-ai-work/

Questions de réflexion

1. Que vous évoque le terme « IA » ? Énumérez librement vos connotations et comparez-les avec un pair. Avez-vous eu des idées similaires ? Comment ces idées se reflètent-elles éventuellement dans les discours publics dominants sur l'IA ?

2. Imaginez le développement technologique des trois prochaines décennies dans au moins l'un des environnements suivants : maison/famille, école, soins de santé. Quels processus ont été automatisés ? Comment l'automatisation a-t-elle affecté le comportement, les interactions sociales et les expériences des gens ?

Invitez les participants à la formation à regarder les vidéos suivantes.

Source : BBC, https://youtu.be/fvtrRGmv7aU

Source : OCDE, https://youtu.be/6Y_ysDHn4uU

Qu'est-ce qu'un algorithme ?

Un algorithme fait référence à une série d'instructions pour effectuer des calculs ou d'autres tâches, que ce soit en mathématiques ou en informatique. Dans le cas de l'IA, un algorithme fournit les instructions qui permettent à un ordinateur d'apprendre à apprendre de son environnement et d'effectuer un ensemble de tâches.[12]

Alors qu'un algorithme général peut être simple, les algorithmes d'IA sont plus complexes.

Les algorithmes d'IA sont conçus pour apprendre des données d'entraînement, qui peuvent être étiquetées ou non étiquetées. L'algorithme utilise ces informations pour améliorer ses capacités et effectuer ses tâches. Certains algorithmes d'IA sont capables d'apprentissage continu et peuvent incorporer de nouvelles entrées de données pour affiner leur processus, tandis que d'autres nécessitent l'intervention d'un programmeur pour optimiser leurs performances.[13]

> Les algorithmes fonctionnent en prenant un ensemble d'entrées, telles que l'âge, la zone de résidence, l'état matrimonial ou le revenu d'une personne, et en les exécutant à travers un ensemble d'étapes qui génèrent un ou plusieurs résultats, ou décisions, pour cette personne ou ce groupe, telles que l'admissibilité à un programme d'aide financière ou à l'école publique à laquelle un enfant est affecté. Les algorithmes sont utilisés dans divers secteurs et objectifs, de la prise de décisions en matière de soins de santé, à l'éligibilité aux prestations publiques, la planification des infrastructures, l'allocation budgétaire, entre autres domaines, avec divers degrés de complexité et d'intrants.

La prise de décision automatisée (ADM) fait référence à l'utilisation de « résultats produits par des algorithmes pour prendre des décisions ».[14]

Vagues de développement de l'IA

Les systèmes d'IA de la première vague étaient des systèmes experts ou basés sur des règles, où un ordinateur suivait une programmation spécifique pour générer des résultats. Cependant, les systèmes d'IA de la deuxième vague, basés sur l'apprentissage automatique, apprennent des données de formation et infèrent des règles pour prédire des résultats spécifiques. Les systèmes d'IA de la troisième vague combinent les avantages des deux vagues précédentes et ont des capacités supplémentaires de pouvoir répondre au contexte dans lequel ils sont utilisés et de fournir aux utilisateurs des explications sur leur processus de prise de décision.[15]

Les sections ci-dessous expliquent et se concentrent sur (i) les systèmes experts et la programmation traditionnelle et (ii) l'apprentissage automatique.

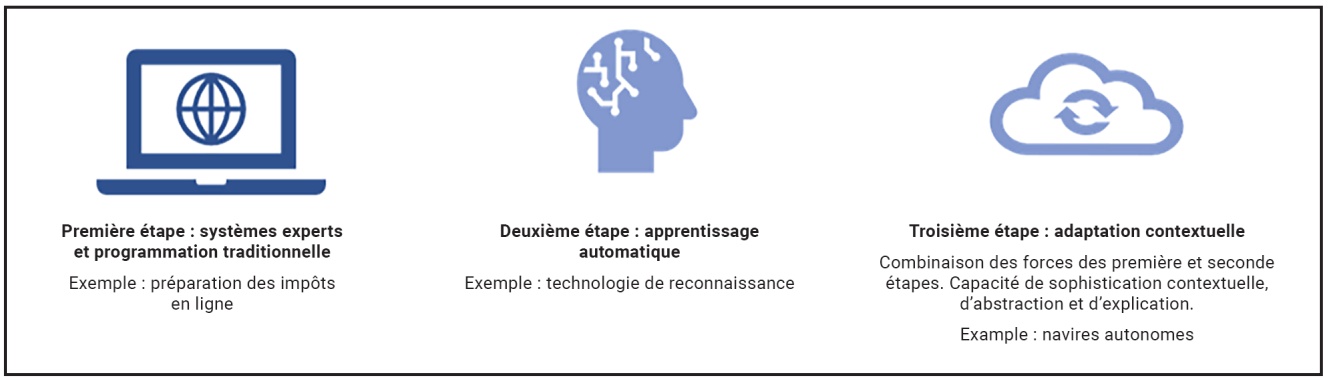
12 Bell F., Bennett Moses L., Legg M., Silove J., Zalnieriute M. (2022). AI Decision-Making and the Courts: A Guide for Judges, Tribunal Members and Court Administrators, Australasian Institute of Judicial Administration, disponible sur : https://ssrn.com/abstract=4162985; Statement on Algorithmic Transparency Accountability, Association for Computing Machinery (2017), disponible sur : https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf; voir également : https://www.tableau.com/data-insights/ai/algorithms.

13 OCDE (2019). Artificial Intelligence in Society, disponible sur : https://www.oecd.org/publications/artificial-intelligence-in-society-eedfee77-en.htm

14 Access Now (2018). Human rights in the age of artificial intelligence, disponible sur : https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf

15 GAO (2021). Artificial Intelligence, An Accountability Framework for Federal Agencies and Other Entities, disponible sur : https://www.gao.gov/products/gao-21-519sp.

Figure 1. Vagues de développement de l'IA



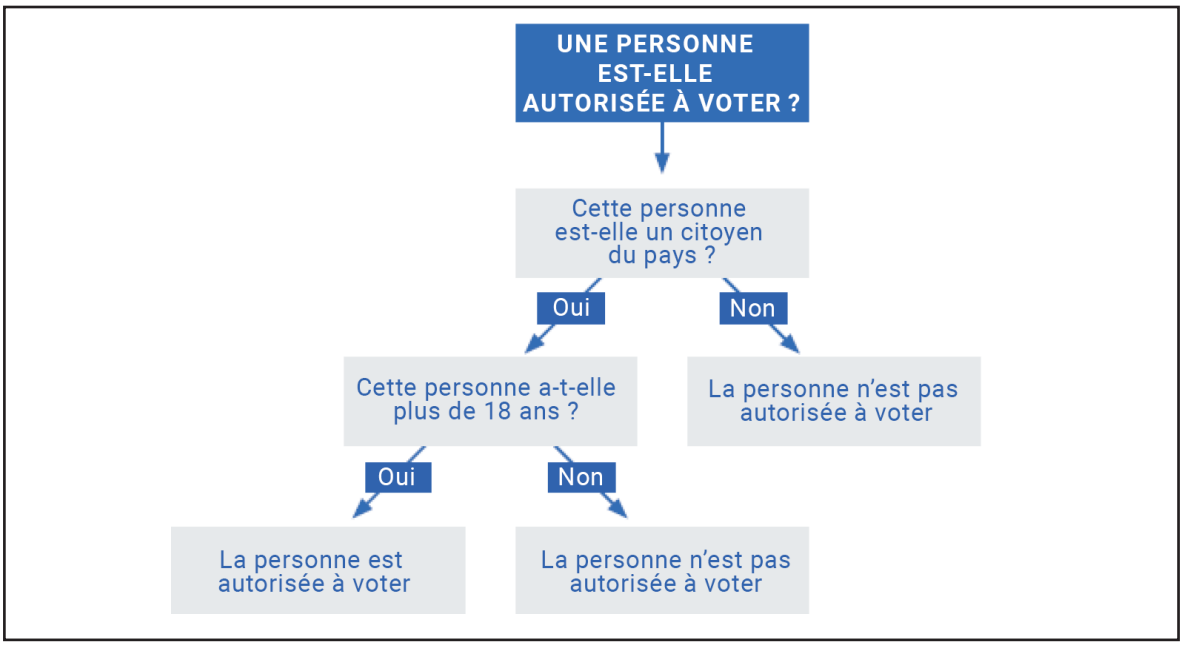
Source : Adaptation de GAO (2021). Artificial Intelligence, An Accountability Framework for Federal Agencies and Other Entities, disponible sur : https://www.gao.gov/products/gao-21-519sp

Systèmes experts et programmation traditionnelle

Un « système expert » est un système d'IA de « première génération » qui fournit des prévisions, des recommandations ou des conclusions basées sur l'entrée de données. Il s'agit d'une séquence d'étapes clairement programmées et de règles dites « si… alors », qu'un ordinateur peut appliquer pour produire un résultat. Ces systèmes sont généralement incapables de traiter des informations fraîches ou des défis inattendus.

Les choix possibles sont appelés « nœuds » dans un arbre de décision, qui est une représentation visuelle des règles du système expert. La figure 2 ci-dessous montre un exemple d'arbre de décision qui décide si une personne peut voter à une élection dans un pays où les seules conditions pour pouvoir voter sont que l'individu ait plus de 18 ans et qu'il soit citoyen d'un pays particulier. Étant donné que chaque branche n'a que deux nœuds, la figure 2 est un exemple d'arbre de décision « binaire ».

Figure 2. Exemple d'arbre de décision



Source : Bell F., Bennett Moses L., Legg M., Silove J., Zalnieriute M. (2022). AI Decision-Making and the Courts: A Guide for Judges, Tribunal Members and Court Administrators, Australasian Institute of Judicial Administration, disponible sur : https://ssrn.com/abstract=4162985

Les systèmes experts en IA de première génération sont largement utilisés dans les systèmes de planification et d'optimisation. Parmi les exemples, citons les logiciels de traitement fiscal, les systèmes de service à la clientèle et d'assistance technique et les systèmes de diagnostic médical. Il peut également s'agir d'une méthode d'alerte à la fraude dans laquelle un expert précise que si les informations administratives fournies comportent plus de cinq inexactitudes, le système doit émettre une alerte indiquant que ce dossier doit faire l'objet d'une enquête.

Initialement, la maîtrise d'un langage de programmation était nécessaire pour créer des règles compréhensibles par la machine. Un « système expert » permet à un expert du domaine (un avocat, par exemple) sans compétences en programmation d'élaborer des règles. Une variété de plateformes « sans code » sont désormais disponibles et facilitent la « programmation » d'un ordinateur pour suivre une certaine procédure ou tirer des conclusions basées sur un ensemble de règles. Parmi celles-ci, on trouve Datalex d'Austlii[16], Josef Legal[17], Checkbox[18], Neota Logic[19] et Realta Logic[20]. Ces plateformes permettent aux professionnels du droit de concevoir un ensemble de règles en utilisant, en fonction de la plateforme utilisée, des mots, des déclarations, des flèches, des menus glisser-déposer ou déroulants, ou d'autres processus similaires. En conséquence, même un avocat sans expérience en programmation peut coder un arbre de décision comme celui de la figure 2.[21]

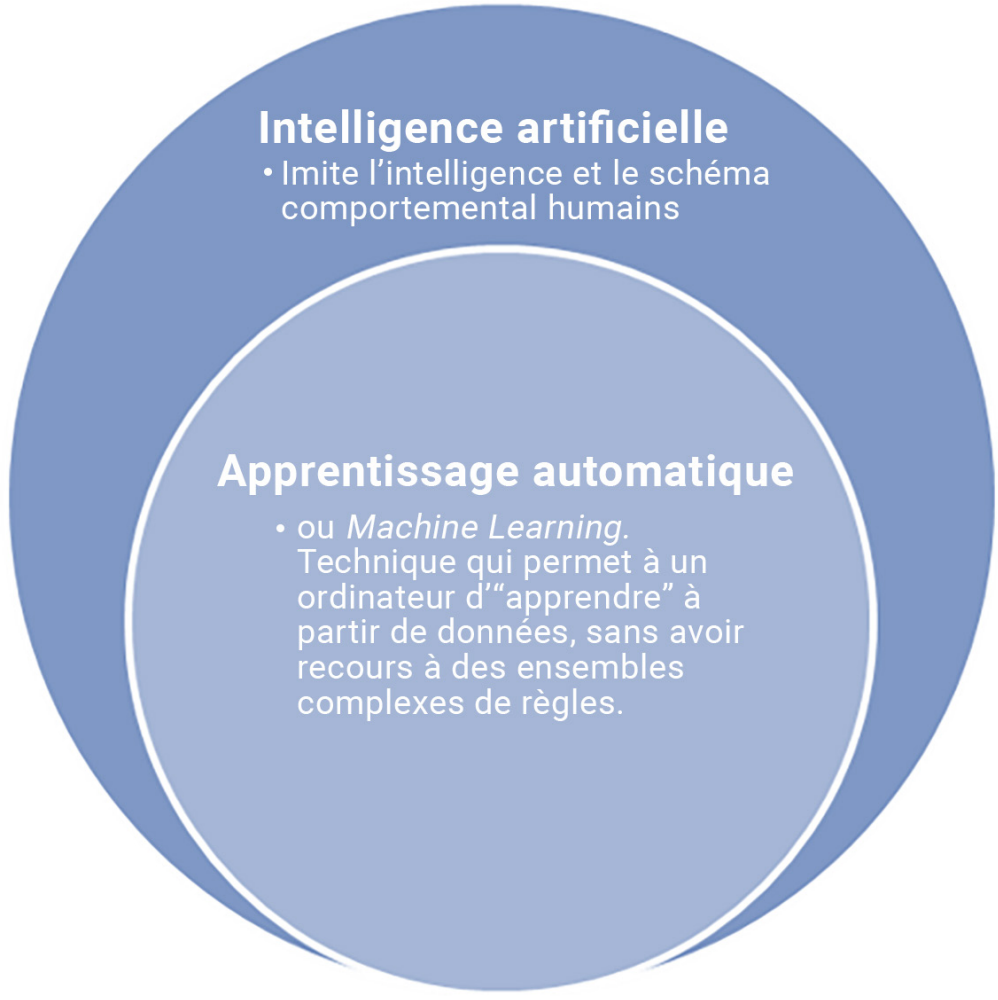
Qu'est-ce que l'apprentissage automatique ?

Les systèmes d'IA utilisent de plus en plus l'apprentissage automatique ou *machine learning* (ML), qui est un sous-ensemble de l'IA. Le ML est un ensemble de techniques qui permet aux machines d'apprendre automatiquement en utilisant des modèles et des déductions, plutôt que des instructions directes d'une personne.[22] Les techniques de ML demandent souvent aux machines d'arriver à un résultat en fournissant de nombreuses instances de résultats corrects. Cependant, elles peuvent également spécifier un ensemble de directives et laisser la machine les découvrir seule dans les données.[23]

Il existe de nombreuses applications de ML. Certains sont conçus pour un problème spécifique, comme la reconnaissance vocale ou d'image, tandis que d'autres peuvent être utilisés pour un plus large éventail de tâches.[24] Le ML a été intégré dans des produits pour s'attaquer à une variété de problèmes trop compliqués pour les systèmes d'IA de « première génération » ou la prise de décision humaine.

16 Voir : https://austlii.community/wiki/DataLex
17 Voir : https://joseflegal.com/
18 Voir : https://www.checkbox.ai/
19 Voir : https://neota.com/
20 Voir : https://www.realtalogic.com/
21 Bell F., Bennett Moses L., Legg M., Silove J., Zalnieriute M. (2022). AI Decision-Making and the Courts: A Guide for Judges Tribunal Members and Court Administrators, Australasian Institute of Judicial Administration, disponible sur : https://ssrn.com/abstract=4162985
22 OCDE (2019). Artificial Intelligence in Society, disponible sur : https://www.oecd.org/publications/artificial-intelligence-in-society-eedfee77-en.htm
23 *Ibid.* De nombreuses méthodes employées par les économistes, les scientifiques et les ingénieurs depuis des années peuvent être trouvées dans le ML. Il s'agit notamment de l'analyse en composantes principales, des arbres de décision, des réseaux neuronaux profonds et des régressions linéaires et logistiques. Voir : https://www.oecd-ilibrary.org/sites/8b303b6f-en/index.html?itemId=/content/component/8b303b6f-en.
24 OCDE (2019). Artificial Intelligence in Society, disponible sur : https://www.oecd.org/publications/artificial-intelligence-in-society-eedfee77-en.htm

Figure 3. La relation entre l'IA et le ML



Source : Auteurs

Le ML alimente les chatbots, le texte prédictif, les applications de traduction linguistique, les recommandations Netflix et l'organisation des flux de médias sociaux. Il permet également aux voitures et machines autonomes adaptées de diagnostiquer des conditions médicales à l'aide de l'analyse d'images.[25]

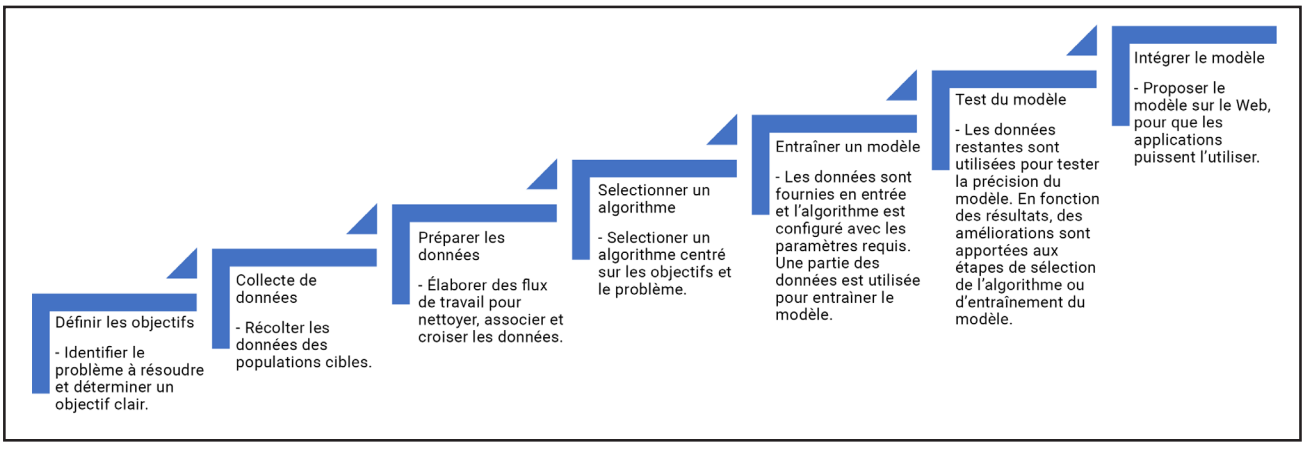
Les systèmes de ML « apprennent » à mesure qu'ils analysent les données. Le ML est distinct de l'apprentissage humain. Bien que le fait de voir peu de photographies d'un chat permette à un enfant moyen de comprendre le terme « chat » et d'identifier des images supplémentaires comme étant celles de chats, les systèmes de ML nécessitent un ensemble de données beaucoup plus important pour effectuer la même tâche de catégorisation. Le programme de ML s'appuie sur une base de données contenant des images de chats et de chiens. Chaque image est étiquetée « chat » ou « chien ». Si le programme

25 Brown S. (2021). Machine learning, explained, disponible sur : https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

de ML voit suffisamment d'images étiquetées, il commencera à différencier les caractéristiques de chaque animal (entraînement ou ajustement du ML). Une fois que le programme de ML aura appris, il sera en mesure de deviner à quelle classe appartient chaque image. Des expériences très similaires peuvent être menées avec du texte.[26] Un autre bon exemple de programme de ML est le processus d'attribution de pointages de crédit par les institutions financières, où les données utilisées pour former le système de ML sont déjà classées comme positives ou négatives, en fonction des antécédents de crédit du client.[27] Nous devons nous rappeler que l'efficacité des modèles de ML dépend de la quantité de données de formation disponibles, de la qualité de la formation et des données d'entrée, et du volume de puissance de calcul utilisée pour construire le modèle.[28]

La figure 4 ci-dessous donne un aperçu simplifié d'un processus de ML, comprenant les phases suivantes : (i) définition des objectifs ; (ii) collecte de données ; (iii) préparation des données ; (iv) sélection de l'algorithme ; (v) formation du modèle ; (vi) test du modèle ; et (vii) intégration du modèle.

Figure 4. Aperçu simpliste du processus de ML



Source : Auteurs

26 Medvedeva M., Vols M., Wieling M. (2020). Using machine learning to predict decisions of the European Court of Human Rights, Artif Intell Law, 28, 237–266, disponible sur : https://link.springer.com/article/10.1007/s10506-019-09255-y

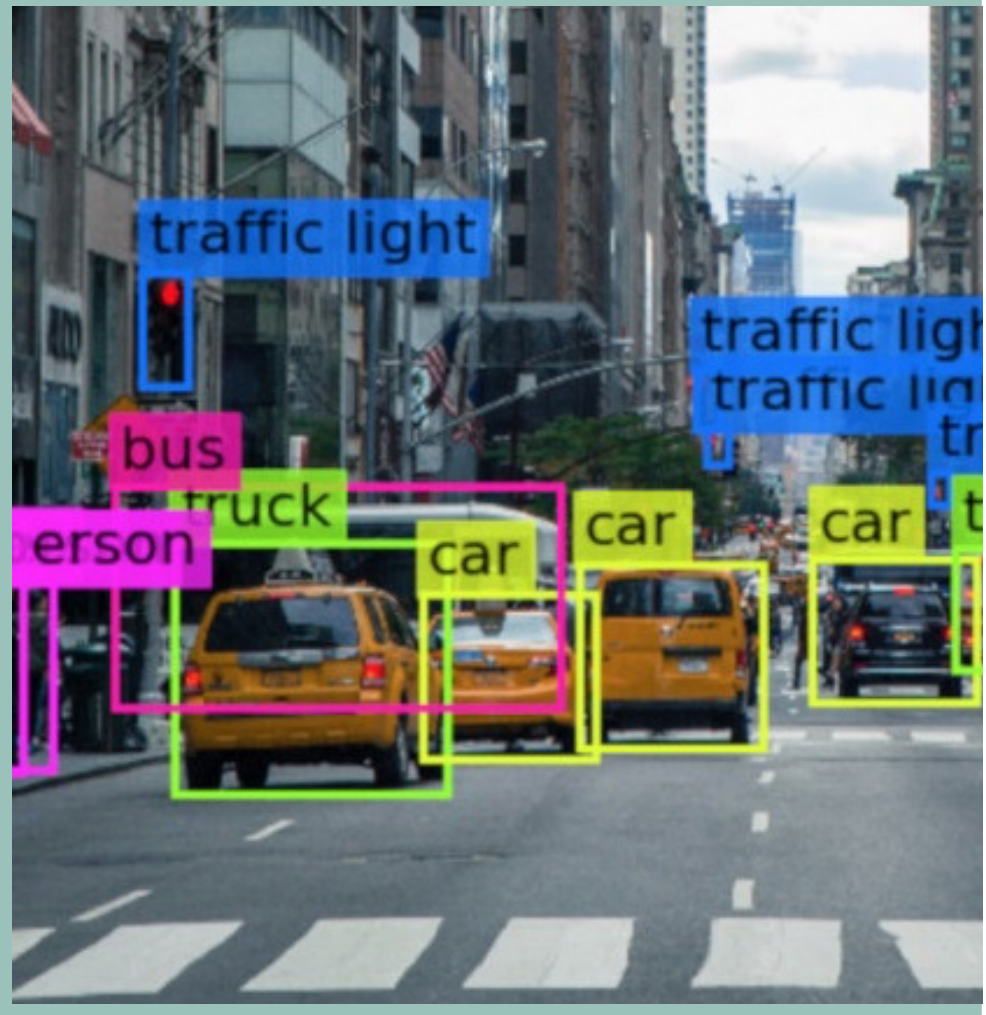
27 The Royal Society (2012). Machine Learning : The Power and Promise of Computers that Learn by Example, disponible sur : https://royalsociety.org/~/media/policy/projects/machine-learning/publications/machine-learning-report.pdf 16 ; Allens Linklaters (2018). AI Toolkit : Ethical, Safe, lawful ; Practical Guidance for AI Projects, disponible sur : https://lpscdn.linklaters.com/-/media/files/insights/thought-leadership/ai-toolkit/ethical-safe-lawful-toolkit-for-artificial-intelligence-projects-nov2018.ashx?rev=b82597fb-d88a-457d-a41a-a24ec1fc7253&extension=pdf) ; https://humanrights.gov.au/our-work/rights-and-freedoms/publications/human-rights-and-technology-final-report-2021.

28 Stankovich M., Behrens E., Burchell J. (2023). Toward Meaningful Transparency and Accountability of AI Algorithms in Public Service Delivery, disponible sur : https://www.dai.com/uploads/ai-in-public-service.pdf

Étiquetage des données ML

L'étiquetage des données dans l'apprentissage des données (ML) est le processus de reconnaissance des données brutes (images, fichiers texte, vidéos, etc.) auxquelles on ajoute une ou plusieurs étiquettes pertinentes et utiles pour offrir un contexte à un modèle de ML, afin qu'il en tire des leçons. Les étiquettes peuvent indiquer si une photographie contient un oiseau ou une automobile, si des mots ont été prononcés dans un enregistrement audio ou si une radiographie montre une tumeur. De nombreux cas d'application, notamment la vision par ordinateur, le traitement du langage naturel et la reconnaissance vocale, nécessitent un étiquetage des données[29].

Figure 5. Étiquetage des données en ML



Source : Energy (2021). The One, Two, Threes of Data Labeling for Computer Vision, disponible sur : https://medium.com/unpackai/the-one-two-threes-of-data-labeling-for-computer-vision-4c0b022cef4

29 Voir : https://aws.amazon.com/sagemaker/data-labeling/what-is-data-labeling/

Le processus de découverte dans les litiges peut servir d'excellent exemple de mise en évidence de la complexité de l'utilisation du ML dans le système judiciaire. Voir Figure 6, ci-dessous.

Figure 6. Découverte en contentieux : trois niveaux d'automatisation possibles

Niveau 1 - Aucune automatisation. À ce niveau, un juriste examine les documents juridiques suivant une série de paramètres déterminés.

Niveau 2 - Automatisation sans apprentissage automatique. Un système informatique utilise des critères fixes comme des intervalles de dates, des listes de phrases, des emplacements de fichiers, pour mener à bien la recherche de documents.

Niveau 3 – Apprentissage automatique. Un juriste identifie les documents qui font partie des résultats de recherche. Ce sont les données d'entraînement. Ensuite, pour déduire des critères de recherche sur la base de schémas issus des données d'entraînement, on peut recourir à un système d'apprentissage automatique plutôt que de faire appel à de la main d'œuvre humaine. Le modèle d'apprentissage automatique ainsi entraîné classera les documents restants suivant qu'ils sont ou non susceptibles d'être trouvés grâce à ces schémas.

Source : Auteurs

Activité :

Les participants à la formation discutent d'un scénario hypothétique concernant l'utilisation de preuves générées par l'IA dans les procédures judiciaires. Que feriez-vous si vous étiez dans une situation similaire ? Quelles questions juridiques clés prendriez-vous en considération ?

Dans un avenir pas si lointain, les preuves générées par l'IA jouent un rôle central dans une affaire judiciaire très médiatisée. Voici le déroulé de l'affaire :

Contexte : Une entreprise technologique de premier plan est accusée d'utiliser des algorithmes biaisés dans son processus d'embauche, ce qui entraîne une discrimination à l'encontre de certains groupes démographiques. L'affaire, qui a attiré l'attention du public, est surveillée de près quant à ses implications potentielles dans l'éthique de l'IA et la responsabilité des entreprises.

Preuves générées par l'IA :

1. **Rapport d'audit algorithmique :** Les plaignants ont employé une équipe d'éthiciens de l'IA et de scientifiques des données pour mener un audit complet des algorithmes d'embauche de l'entreprise. Ils présentent un rapport détaillé généré par les systèmes d'IA qui met en évidence les cas de biais et de discrimination dans le processus décisionnel de l'algorithme.
2. **Simulation générée par l'IA :** Pour démontrer le comportement de l'algorithme, les demandeurs introduisent une simulation générée par l'IA qui imite le processus d'embauche de l'entreprise. Cette simulation utilise des données historiques pour montrer comment l'algorithme a tendance à favoriser certains groupes démographiques par rapport à d'autres.
3. **Témoignage d'expert généré par l'IA :** La défense appelle un expert en éthique de l'IA qui utilise l'IA de traitement du langage naturel pour analyser les communications et les documents internes de l'entreprise. L'IA identifie des cas où les employés ont exprimé des préoccupations concernant le biais algorithmique, suggérant que l'entreprise était potentiellement au courant du problème.

Implications juridiques : L'introduction de preuves générées par l'IA présente plusieurs défis et considérations juridiques :

1. **Recevabilité :** Le tribunal doit déterminer l'admissibilité des éléments de preuve générés par l'IA, en évaluant leur fiabilité et leur pertinence par rapport à l'affaire.
2. **Témoignage d'expert :** Le tribunal se penche sur la question de savoir si l'IA peut être considérée comme un « témoin expert » et comment son témoignage doit être traité.
3. **Implications éthiques :** L'affaire soulève des questions éthiques sur la responsabilité des entreprises lors du déploiement de systèmes d'IA et les conséquences potentielles des biais algorithmiques.
4. **Impact sur le précédent :** L'issue de cette affaire pourrait créer un précédent sur la façon dont les preuves générées par l'IA sont traitées dans les futures procédures judiciaires, influençant le paysage juridique en matière d'éthique de l'IA.
5. **Surveillance humaine :** Malgré les preuves générées par l'IA, le jugement humain reste crucial pour interpréter les preuves, assurer l'équité et prendre des décisions juridiques.

Ce scénario hypothétique souligne le rôle évolutif de l'IA dans les procédures judiciaires, ainsi que la nécessité de cadres juridiques solides pour répondre aux complexités et aux préoccupations éthiques associées aux preuves générées par l'IA dans les salles d'audience.

2. Pourquoi les données sont-elles importantes ?

Les algorithmes d'IA nécessitent un accès aux données, car les machines ne peuvent pas « apprendre » à moins qu'elles ne disposent de grands ensembles de données à partir desquels discerner des modèles. La disponibilité des données est une exigence nécessaire au développement de l'IA, lui permettant d'effectuer certaines tâches antérieurement réalisées manuellement par l'homme.

Le processus de « datafication » fait référence à la prolifération des outils numériques utilisés pour intégrer, analyser et afficher des modèles de données. La datafication indique que de nombreux aspects de la vie sociale prennent la forme d'empreintes numériques. Les amitiés deviennent des « j'aime » sur Facebook, les mouvements à travers la ville laissent de vastes empreintes numériques dans les gadgets compatibles avec le GPS, et les recherches d'informations révèlent ce que les individus et les communautés apprécient ou désirent.[30]

Une fois que les appareils connectés à Internet commencent à communiquer entre eux, une quantité extraordinaire de nouvelles données est transmise sans le savoir et de manière pratiquement inaperçue par la plupart des utilisateurs. Par exemple, il existe des métadonnées (données sur les données), telles que les informations de routage contenues dans les en-têtes des e-mails ou des messages texte, ou les informations de géolocalisation dissimulées dans une photographie numérique. Les métadonnées, en tant qu'informations structurées, peuvent être plus facilement comparées et évaluées par des algorithmes et, par conséquent, donner fréquemment des informations exceptionnellement exactes sur les intérêts, les mouvements et les relations des individus.

Les plateformes numériques ont accès à de nombreuses informations sur ce que les gens font en ligne. Ces flux massifs de traces numériques, appelés données de masse, peuvent être utilisés conjointement avec des techniques de tri automatisé, telles que les algorithmes et l'IA, pour révéler des modèles importants et générer des informations analytiques sur les clients, les maladies et les activités criminelles. De nombreuses plateformes et entreprises numériques cherchent à verrouiller les clients dès le début, en devenant l'endroit où ils achètent des livres ou visionnent des films, par exemple. Ils veulent également construire des écosystèmes fermés, comme Netflix ou Amazon, où ils peuvent contrôler et extraire de la valeur de ces données.[31]

La qualité des données a un impact sur le résultat de l'IA, en termes de biais [pour le biais d'IA, se reporter au module 3]. Les données doivent idéalement être exemptes de biais, la propriété des données doit être clairement établie et les algorithmes doivent être suffisamment transparents pour indiquer la responsabilité des parties prenantes. Les obligations de toutes les parties prenantes du cycle de vie de l'IA doivent être définies pour prévenir les dommages et réparer ou compenser les dommages causés par les systèmes d'IA.

Lorsqu'ils statuent sur des affaires impliquant le déploiement de l'IA et son impact sur les droits humains, les opérateurs judiciaires doivent tenir compte des questions suivantes liées aux données et aux ensembles de données qui alimentent les systèmes d'IA (voir tableau 2).

30 Matteson A. (2018). The Concept of Datafication ; Definition & Examples, disponible sur : https://www.datasciencecentral.com/the-concept-of-datafication-definition-amp-examples/

31 Flyverbom M., Deibert R., Matten, D. (2019). The Governance of Digital Technology, Big Data, and the Internet : New Roles and Responsibilities for Business. Business & Society, 58(1), 3–19, disponible sur : https://doi.org/10.1177/0007650317727540

Tableau 2. Questions liées aux données et aux ensembles de données qui alimentent les systèmes d'IA

Questions	Points à souligner
Accès aux données et disponibilité	L'absence de systèmes nécessaires qui génèrent et maintiennent des données robustes, précises et pertinentes a rendu le développement d'applications d'IA difficile dans certains contextes.
Exactitude des données	L'accès à des données précises est crucial pour réussir le déploiement de l'IA et des ressources numériques. Une bonne pratique en matière de protection de l'exactitude des données est la pratique dite de « dégorgement algorithmique », qui oblige les développeurs de systèmes d'IA à supprimer toutes les données obtenues illégalement et utilisées pour former les systèmes d'IA.[32]
Qualité des données	L'un des principaux obstacles au déploiement efficace de l'IA dans le système judiciaire est l'accès à des données FAIR (soit, équitables : trouvables, accessibles, interopérables et réutilisables). Ce problème est exacerbé dans certains contextes, car les données ne sont pas toujours numérisées et sont difficilement accessibles. Questions clés à poser à cet égard : Quelle est la qualité des données sur lesquelles le système d'IA est formé ? Y a-t-il un risque de biais de données et d'amplification d'informations incorrectes, en utilisant l'IA ? Le problème est que les données qui alimentent les systèmes d'IA peuvent être inexactes, incomplètes, ou contenir des erreurs ou du matériel sans importance. Les données peuvent être biaisées. Souvent, les machines collectent des données déjà faussées, car provenant d'une réalité erratique et biaisée. Par exemple, les essais cliniques excluent souvent les femmes et les personnes de couleur, ce qui entraîne une représentation inadéquate des données. Cela peut avoir de graves conséquences si des algorithmes formés à l'aide de ces données sont utilisés pour analyser les images de la peau ou donner la priorité aux soins des patients. Par conséquent, il est crucial de veiller à ce que les algorithmes d'IA soient formés à l'aide de données représentatives, afin d'éviter de tels biais et d'assurer des résultats équitables pour tous.[33] Exemple : La plupart des systèmes d'IA utilisés en justice pénale sont des modèles statistiques, basés sur des données d'application de la loi ou de casiers judiciaires, qui représentent des biais structurels et des inégalités sociales. Ces données sont un enregistrement des crimes, des lieux et des groupes policiers, et ne constituent pas nécessairement un enregistrement de la survenance réelle du crime. Ces données utilisées dans les systèmes d'IA peuvent renforcer et réintégrer les schémas de discrimination dans les systèmes de justice ou d'application de la loi.[34] Les régulateurs des modèles d'IA doivent s'assurer que les données utilisées respectent les principes FAIR et sont collectées de manière éthique, avant de certifier que le modèle est adapté au marché. Cela peut s'accompagner d'une évaluation de la qualité organisationnelle dans les points de contrôle de pré-commercialisation. Ces conditions peuvent signaler à l'industrie que l'intégrité des données et la collecte éthique sont d'une importance primordiale pour mettre les solutions d'IA sur le marché et entraîner des changements structurels positifs dans le fonctionnement des entreprises.

32 La FTC a utilisé cette pratique pour obliger Everalbum, les créateurs de l'application Ever, aujourd'hui disparue, à supprimer les systèmes de reconnaissance faciale développés à partir du contenu obtenu auprès des utilisateurs de l'application. Voir aussi : Kay K. (2021). Why the FTC is forcing tech firms to kill their algorithms along with ill-gotten data, disponible sur : https://digiday.com/media/why-the-ftc-is-forcing-tech-firms-to-kill-their-algorithms-along-with-ill-gotten-data/

33 Siwicki B. (2021). How does bias affect healthcare AI, and what can be done about it ?, disponible sur : https://www.healthcareitnews.com/news/how-does-bias-affect-healthcare-ai-and-what-can-be-done-about-it

34 Fair Trials (2021). Regulating Artificial Intelligence for Use in Criminal Justice Systems in the EU Policy Paper, disponible sur : https://www.fairtrials.org/sites/default/files/Regulating%20Artificial%20Intelligence%20for%20Use%20in%20Criminal%20Justice%20Systems%20-%20Fair%20Trials.pdf

Questions	Points à souligner
Représentativité des données	Un ensemble de données est représentatif s'il reflète ou mesure avec précision la population ou le phénomène qu'il est censé enregistrer, par rapport à son application prévue.[35] Exemple : Une dépendance excessive aux techniques de collecte de données « automatisées » peut exclure des groupes extrêmement vulnérables et éroder la confiance dans la prise de décision automatisée. Les personnes sans accès numérique (c'est-à-dire celles qui n'ont pas de connectivité ou d'appareils) ou qui manquent de compétences numériques ne seront pas prises en compte dans les analyses de la population et de ses besoins. Les fractures numériques dans de nombreux pays du Sud ont conduit à une « invisibilité des données », ce qui peut avoir un impact sur les groupes historiquement marginalisés comme les femmes, les castes, les communautés tribales, les minorités religieuses et linguistiques et les travailleurs migrants. L'utilité et la validité des algorithmes d'IA développés sur des données facilement disponibles peuvent être limitées par des biais perpétués par l'invisibilité des données. Cela souligne les exigences de transparence et de responsabilité algorithmiques.
Propriété des données	Un problème clé dans le développement et le déploiement de l'IA est la propriété des données, c'est-à-dire qui possède, gère et collecte les données qui intègrent le système d'IA. Les questions importantes à considérer à cet égard sont la définition des objectifs (pourquoi avons-nous besoin du système d'IA), la détermination des données de formation à acquérir et la manière de catégoriser ces données. Par conséquent, des jugements humains sont constamment nécessaires, lors de la compilation d'ensembles de données et du développement d'algorithmes de prédiction.
Stockage des données/ minimisation des données	Le stockage à long terme des données personnelles comporte des dangers, car les données sont susceptibles d'être exploitées d'une manière qui n'était pas prévue au moment de leur collecte. Les données peuvent devenir obsolètes, non pertinentes ou contenir une mauvaise interprétation historique au fil du temps, ce qui peut entraîner des résultats biaisés ou incorrects du traitement des données à l'avenir.[36]
Protection des données/de la vie privée	Des lois *ad hoc* sur la protection des données abordent des questions telles que la confidentialité des données (un droit humain fondamental), la gestion et le partage des données et des mécanismes innovants de gouvernance des données, tels que les sandboxes de données et les fiducies de données. Les politiques et réglementations actuelles en matière de données entre les pays et les régions sont très parcellaires, avec des approches réglementaires mondiales, régionales et nationales divergentes. De nombreux pays et régions ont pris des mesures pour mettre à jour les règles sur l'utilisation des données personnelles. Le Règlement général sur la protection des données de l'UE[37] (RGPD) impose une longue liste d'exigences aux entreprises qui traitent des données à caractère personnel. Les infractions entraînent des amendes pouvant atteindre 4 % du chiffre d'affaires annuel mondial. Le RGPD permet un meilleur contrôle des données personnelles, donnant droit à la protection individuelle de l'anonymat, du pseudonyme et du droit à l'oubli. La portabilité des données donne aux individus le droit de demander que leurs données soient transférées à un autre responsable du traitement, et aux responsables du traitement d'utiliser des formats communs. Plus de 30 % des pays, principalement des pays en développement, n'ont pas de législation sur la gouvernance des données, et peu ont élaboré une loi complète sur la protection des données[38]. Parmi les cadres régionaux visant à établir des règles sur la confidentialité des données personnelles, on retrouve le Cadre de protection de la vie privée de l'APEC (2015), les Principes directeurs de l'OCDE sur la protection de la vie privée (2013) et la Convention 108+ du Conseil de l'Europe[39], qui a mis à jour les lignes directrices sur la protection des données. Dans les pays qui ne disposent pas d'un système de protection des données, il apparaît inutile que les tribunaux soient amenés à établir des lignes directrices pour l'utilisation des données, qui seraient conformes aux droits légaux.

35 AAAS. Artificial Intelligence and the Courts: Materials for Judges, disponible sur : https://www.aaas.org/ai2/projects/law/judicialpapers

36 Conseil des droits de l'homme des Nations Unies (2021). The right to privacy in the digital age, Report of the United Nations High Commissioner for Human Rights, disponible sur : https://www.ohchr.org/en/HRBodies/HRC/RegularSessions/Session48/Documents/A_HRC_48_31_AdvanceEditedVersion.docx

37 Guide complet sur la conformité au RGPD, disponible sur : https://gdpr.eu/

38 UNCTAD, Data Protection and Privacy Legislation Worldwide, disponible sur : https://unctad.org/page/data-protection-and-privacy-legislation-worldwide

39 Conseil de l'Europe, Modernisation de la Convention 108, disponible sur : https://www.coe.int/fr/web/data-protection/convention108/modernised

Questions	Points à souligner
Infrastructure de données	Les progrès actuels en matière d'IA et de données de masse sont alimentés par de meilleures connexions numériques, des quantités croissantes de données, des algorithmes sophistiqués et une puissance de traitement accrue. L'IA et les données de masse peuvent grandement améliorer la vie dans les pays en développement et aider à atteindre les objectifs de développement durable des Nations Unies. Les décideurs politiques doivent tendre à permettre, encourager et/ou accélérer les investissements dans la construction d'une infrastructure de données adéquate et abordable. Des investissements dans les logiciels, le matériel et la connectivité à large bande sont nécessaires pour un accès et une utilisation généralisés des données. Ceci est essentiel pour atteindre les personnes mal desservies. Il est crucial d'encourager la création de données FAIR et d'une infrastructure de données FAIR.[40]
Questions supplémentaires à poser	• Le système d'IA a-t-il fait l'objet d'audits de transparence algorithmique ou d'évaluations de l'impact sur la vie privée? • Des techniques d'amélioration de la confidentialité ont-elles été utilisées pour préserver la confidentialité des données ? • Quel est l'état de l'information et de la cybersécurité pour la confidentialité des données ?

3. Les systèmes d'IA en tant que « boîtes noires »

Le terme « boîte noire » désigne un système technologique intrinsèquement opaque, dont le fonctionnement interne ou la logique sous-jacente ne sont pas correctement compris, ou dont les résultats et les effets ne peuvent être expliqués.[41] De nombreux systèmes d'IA sont considérés comme des « boîtes noires », c'est-à-dire des systèmes très complexes dont les processus de prise de décision et de raisonnement ne sont pas facilement compris par les utilisateurs, voire par leurs développeurs. Cela peut rendre extrêmement difficile la détection de résultats défectueux, en particulier dans les systèmes d'IA qui découvrent des modèles dans les données sous-jacentes de manière non supervisée.

Les systèmes d'IA analysent les données de formation pour identifier des modèles complexes, puis apprennent ces modèles pour classer les nouvelles données dont ils peuvent être alimentés. De nombreux systèmes d'IA n'expliquent toutefois pas comment les données pourraient être interdépendantes et comment elles parviennent à une certaine décision ou prédisent un certain résultat. Ces systèmes peuvent être beaucoup trop complexes pour la compréhension humaine, même pour ceux qui les programment et les forment.[42] Ils évoluent et apprennent continuellement, et leur comportement est imprévisible. Ils peuvent être en mesure de déduire des faits et des corrélations à partir de variables indirectes, telles que l'historique des achats ou la géographie.

40 FAIR Principles, disponible sur : https://www.go-fair.org/fair-principles/
41 AAAS, Artificial Intelligence and the Courts: Materials for Judges, disponible sur : https://www.aaas.org/ai2/projects/law/judicialpapers
42 OECD, AI in Society, disponible sur : https://www.oecd-ilibrary.org/sites/969ff07f-en/index.html?itemId=/content/component/969ff07f-en

Approfondissement : Discrimination par procuration dans les systèmes d'IA

La discrimination par procuration dans les systèmes d'IA a lieu lorsqu'une caractéristique apparemment neutre est substituée par une caractéristique interdite.[43]

Par exemple, les institutions financières utilisent souvent des codes postaux et des limites de voisinage (géographie). Ces données peuvent identifier la race des demandeurs de prêt, car certains codes postaux peuvent être associés à des groupes sociaux à faible revenu, à des minorités ethniques ou raciales. De même, un système d'IA créé par une compagnie d'assurance peut augmenter les primes pour les candidats qui seraient membres d'un groupe Facebook dédié à l'amélioration de la disponibilité des tests génétiques prédictifs du cancer. Dans ces conditions, l'assureur se livre probablement à une discrimination génétique indirecte en utilisant des procurations, telles que la demande d'un certain type de test génétique et l'adhésion à un groupe Facebook spécifique, pour déduire le lien entre ces procurations et l'histoire génétique (une pratique controversée) et facturer des primes d'assurance plus élevées à ces personnes.[44] On trouve aussi l'exemple des procurations liées à l'âge, « vingt ans d'expérience professionnelle » indique que la personne est au moins âgée de la quarantaine.

Les droits à la vie privée et à la non-discrimination dans les systèmes de prise de décision automatisée nécessitent la minimisation, la limitation ou l'interdiction de certaines utilisations des données, ou la suppression des données (voir le tableau 2 ci-dessus). Cependant, un système d'IA peut faire une prédiction basée sur des données de procuration qui ont une ressemblance étroite avec les catégories restreintes de données. En outre, la seule façon de découvrir ces procurations est d'acquérir des informations sensibles ou privées, telles que la race. Si de telles données sont acquises, il devient crucial de garantir qu'elles soient utilisées exclusivement à des fins adéquates et légitimes.[45] Par exemple, même si les créateurs d'algorithmes ont fait un effort conscient pour prévenir les préjugés raciaux en excluant la race comme paramètre, l'algorithme produira néanmoins des résultats biaisés sur le plan racial s'il inclut des substituts typiques pour la race, tels que le revenu, l'éducation ou le code postal.

L'opacité des algorithmes d'IA et la difficulté à déterminer la responsabilité pour les décisions produites par les systèmes d'IA signifient que des préjudices aux droits humains peuvent survenir, et aucune responsabilité n'est fixée pour ces préjudices. Sans l'intégration des garanties éthiques et des droits humains dans la conception et le déploiement de l'IA, les risques liés à l'IA augmenteront. Cela aura un impact sur l'approfondissement des inégalités existantes intégrées dans les ensembles de données utilisés pour former les algorithmes. Par exemple, ces inégalités pourraient provenir d'un tel parti pris des développeurs. Cela affectera gravement et de manière disproportionnée les groupes défavorisés, mal desservis et marginalisés, et ceux qui sont soumis à des formes croisées de discrimination.

43 Downs J., Auchterlonie S. (2022). Proxy Problems—Solving for Discrimination in Algorithms, disponible sur : https://www.bhfs.com/insights/alerts-articles/2022/proxy-problems-solving-for-discrimination-in-algorithms

44 Iowa Law Review (2020). Proxy Discrimination in the Age of Artificial Intelligence and Big Data, disponible sur : https://ilr.law.uiowa.edu/print/volume-105-issue-3/proxy-discrimination-in-the-age-of-artificial-intelligence-and-big-data

45 O'Neil C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, New York: Crown.

Un autre problème est l'utilisation abusive des garanties de propriété intellectuelle. Les outils algorithmiques tombent souvent sous le bouclier des logiciels propriétaires et des revendications de secrets commerciaux pour protéger la technologie derrière les algorithmes d'un examen minutieux extérieur (voir People c. Chubbs, abordé dans le module 4 de la formation). Cette pratique pourrait entraver tout effort de défense visant à contester la fiabilité de la science sous-jacente à l'outil d'IA. Lorsque les systèmes d'IA sont utilisés dans les opérations pour le compte des parties prenantes du système judiciaire, il existe un besoin accentué de responsabilité, de transparence et d'explicabilité. Les garanties de propriété intellectuelle des données et du système algorithmique peuvent empêcher cette transparence et cette responsabilité. Les parties prenantes de la gouvernance de l'IA doivent trouver un équilibre entre la transparence dans le cadre de l'éthique de l'IA et la nécessité légitime de protéger les secrets commerciaux, lorsque les entreprises privées développent des outils d'IA.

Activité :

Secrets commerciaux, algorithmes et droits fondamentaux : l'étude de cas de l'algorithme EVAAS (Educational Value-Added Assessment System)

Les secrets commerciaux qui protègent les algorithmes affectent les droits fondamentaux. Lisez l'étude de cas ci-dessous et discutez de la façon dont un cas similaire serait jugé dans votre pays. Comment cette affaire serait-elle tranchée en vertu de vos lois nationales ?

Entre 2011 et 2015, la performance au travail des enseignants de Houston a été évaluée à l'aide d'un algorithme « piloté par les données », EVAAS. Le programme a permis au conseil de l'éducation d'automatiser les choix concernant l'octroi de primes aux enseignants, leur pénalisation pour mauvaise performance ou même leur licenciement. Les codes sources sont des secrets commerciaux appartenant à SAS, un fournisseur tiers. À ce titre, les enseignants n'ont pas pu contester les décisions ou recevoir une explication sur la manière dont EVAAS a pris ses décisions.

Un long litige civil s'en est suivi et, en 2017, un juge fédéral américain a conclu que les droits constitutionnels des enseignants avaient été violés par le déploiement de l'algorithme secret permettant d'évaluer la performance des employés sans explication appropriée. Le juge a dû trouver un équilibre entre le droit compréhensible du vendeur privé à préserver ses secrets commerciaux et le droit constitutionnel des enseignants à une procédure régulière, qui protège les citoyens américains contre les privations de vie, de liberté ou de biens fondamentalement injustes ou erronés.

La décision du tribunal a déclaré que les enseignants et la Houston Federation of Teachers devaient être en mesure de vérifier et de contester de manière indépendante les résultats d'évaluation produits par l'algorithme. Cependant, SAS a refusé de révéler le fonctionnement interne de son algorithme EVAAS. En conséquence, le système scolaire de Houston n'utilise plus l'algorithme EVAAS.

Source: Hung K-H., Liddicoat J. (2018). The future of workers' rights in the AI age, disponible /sur : https://policyoptions.irpp.org/magazines/december-2018/future-workers-rights-ai-age

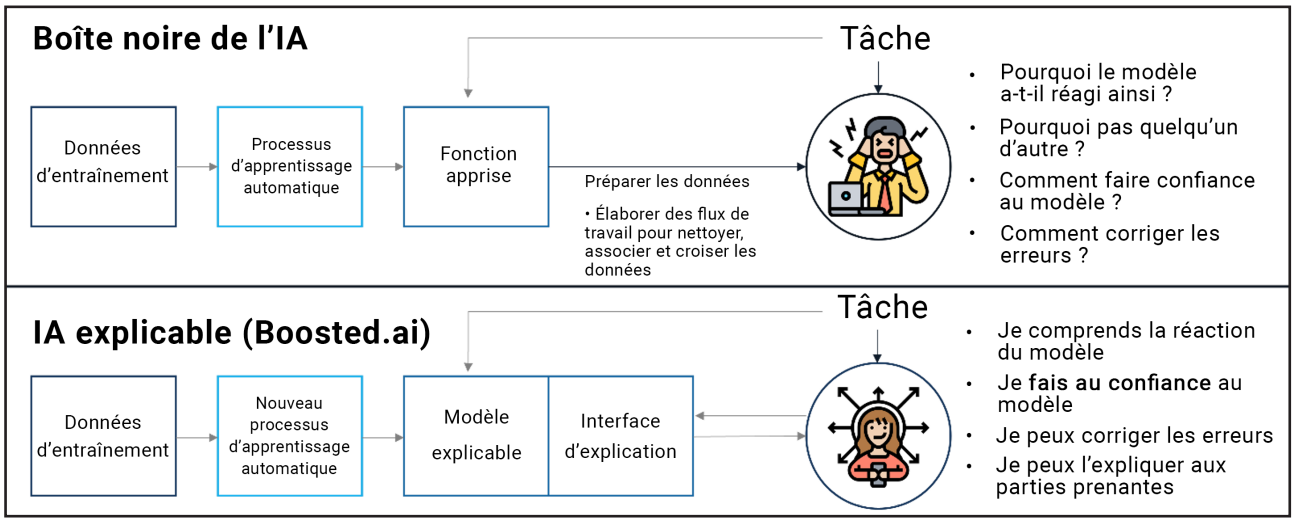
IA explicable (XAI)

La discussion autour des aspects de la boîte noire des systèmes d'IA est en constante évolution. Les progrès de la recherche en IA ont conduit au développement de modèles d'IA qui ne sont pas des boîtes noires.

L'IA explicable (XAI) fait référence à des systèmes, des algorithmes et des modèles capables d'expliquer la raison d'être de leurs décisions, de caractériser les forces et les faiblesses de leur processus décisionnel et de comprendre comment ils se comporteront à l'avenir.

Les chercheurs en XAI se concentrent sur la création de modèles d'IA qui peuvent être compris par la population, ainsi que sur la production d'explications sur les résultats utilisables du ML. Ce public doit avoir la possibilité d'analyser le modèle généré et de discerner sa signification, c'est-à-dire de comprendre la structure du système.

Figure 7. Boîte noire AI versus IA explicable



Source: https://boosted.ai/

Par exemple, Angelino et al (2018) ont développé un modèle de ML interprétable pour prévoir une nouvelle arrestation, qui ne comprend que quelques règles sur l'âge et les antécédents criminels d'un individu. Le modèle de ML complet prévoit qu'une personne sera de nouveau arrêtée dans les deux ans suivant son évaluation, si elle a déjà commis trois infractions ou plus, si elle est âgée de 18 à 20 ans et de sexe masculin ou si elle est âgée de 21 à 23 ans et a déjà commis deux ou trois infractions. Cet ensemble de lignes directrices est aussi précis que le modèle de boîte noire COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) largement utilisé (et propriétaire), en usage dans le comté de Broward, en Floride. Pour vous familiariser avec COMPAS, veuillez vous référer à la section sur les biais algorithmiques.

Étude de cas : Directives du National Institutes of Standards and Technology (NIST) sur l'explicabilité de l'IA

Le NIST américain a publié des directives sur l'explicabilité de l'IA qui pourraient faire partie des systèmes d'évaluation d'impact. Le projet de directives du NIST suggère quatre principes d'explicabilité pour les outils d'évaluation des systèmes de prise de décision automatisée (ADM) sensibles au public, axés sur les objectifs : (1) les systèmes offrent des preuves ou des raisons d'accompagnement pour toutes les résultats ; (2) les systèmes fournissent des explications compréhensibles pour les utilisateurs individuels ; (3) l'explication reflète correctement le processus du système pour générer le résultat ; et (4) le système ne fonctionne que dans les conditions pour lesquelles il a été conçu ou lorsqu'il atteint une confiance suffisante dans son résultat. Ces quatre principes façonnent les types d'explications nécessaires pour assurer la confiance dans les systèmes de prise de décision automatisée, telles que les explications pour le bénéfice de l'utilisateur, pour l'acceptation sociale, à des fins de réglementation et de conformité, pour le développement du système et pour le bénéfice du propriétaire.

Source : NIST (2020). Four Principles of Explainable Artificial Intelligence, disponible sur : https://www.nist.gov/system/files/documents/2020/08/17/NIST%20Explainable%20AI%20Draft%20NISTIR8312%20%281%29.pdf

4. Principe de l'humain dans la boucle

Conscients que de nombreux systèmes d'IA sont des boîtes noires et sont sujets à des préjugés, les opérateurs judiciaires vont commencer à se poser des questions sur la mesure dans laquelle les humains peuvent ou doivent dépendre de l'IA. Les humains doivent-ils superviser ou approuver certains résultats et décisions recommandés par l'IA avant leur mise en œuvre ? Qui est responsable des défauts ou du piratage des technologies basées sur l'IA ? Il y aura des différends sur l'incapacité des parties à comprendre ou à gérer pleinement certaines opérations alimentées par l'IA, ainsi que sur ce qui est juste dans l'ADM.

Pour l'efficacité et la sécurité des applications basées sur l'IA, les opérateurs judiciaires doivent s'assurer qu'il y a toujours un « humain dans la boucle », c'est-à-dire que l'IA ne remplace jamais complètement les humains, de sorte que des professionnels correctement formés valident les décisions de l'IA. L'IA ne vaut pas mieux que les données, le capital humain et l'expertise de l'équipe interdisciplinaire impliquée dans le développement de la solution d'IA. Un cadre adéquat d'IA et de gouvernance des données doit définir les responsabilités respectives de toutes les parties prenantes, y compris les parties prenantes judiciaires. Il doit mettre en place les conditions et les garanties nécessaires pour protéger les droits humains tout en œuvrant en faveur de l'intérêt collectif. Cela peut passer par une certification publique des systèmes d'IA qui garantirait la qualité des données et des algorithmes, afin d'éviter l'aggravation des inégalités existantes. La certification publique des applications d'IA renforcerait la

confiance du public et permettrait aux utilisateurs de donner leur consentement éclairé.[46]

Il est donc important de pouvoir mesurer le niveau de risque et l'impact des différents systèmes d'IA susceptibles d'être déployés dans le système judiciaire. À cet égard, il est important de déterminer l'exigence de surveillance humaine, en fonction du cas d'utilisation, de sa sensibilité, de la complexité et de l'opacité de l'algorithme et de l'impact potentiel sur les droits humains.[47] À titre d'exemple, un joueur d'échecs IA à faible risque pourrait ne nécessiter qu'une simple auto-évaluation, une éducation des utilisateurs et une supervision interne. Cependant, un chirurgien en IA à haut risque pourrait exiger des évaluations menées par des pairs, des dossiers publics, des interventions humaines importantes, une formation périodique et un examen externe.

Le modèle de cadre de gouvernance de l'intelligence artificielle, deuxième édition, développé par le gouvernement de Singapour (voir la figure 8 ci-dessous) décrit trois grandes approches de la supervision humaine des systèmes d'IA : (i) l'humain dans la boucle, (ii) l'humain hors boucle et (iii) l'humain en boucle. La mesure dans laquelle une supervision humaine est nécessaire dépend des objectifs du système d'IA et d'une évaluation des risques, comme l'illustrent les exemples ci-dessous.

- **Le terme « humain dans la boucle » (HITL)** désigne un processus dans lequel un système d'IA est étroitement surveillé par un humain, qui est responsable de prendre toutes les décisions finales. Ceci est particulièrement important dans des domaines comme les soins de santé, où l'IA peut fournir un soutien inestimable dans la formulation de recommandations pour le traitement du cancer, le traitement de la septicémie, la planification chirurgicale, etc. Bien que les outils d'IA puissent aider les prestataires de soins de santé à prendre des décisions éclairées rapidement et avec précision, la responsabilité ultime des soins aux patients incombe toujours à l'expert humain.
- **Le terme « humain hors boucle »** se rapporte à l'absence de supervision humaine dans les décisions prises par le système d'IA. Cela signifie que le système d'IA a un contrôle complet et qu'il n'y a aucune possibilité d'intervention humaine. Par exemple, un système de cybersécurité alimenté par l'IA qui peut détecter et corriger les vulnérabilités du système sans nécessiter d'intervention humaine. Mayhem, le système gagnant du Cyber Grand Challenge 2016 de la Defense Advanced Research Projects Agency (DARPA), est un système innovant qui recherche en permanence toute nouvelle vulnérabilité susceptible d'être exploitée par des pirates informatiques. Lorsque Mayhem détecte un nouveau bug, il génère automatiquement du code pour protéger le logiciel de cette vulnérabilité. Ce système est un expert en analyse prescriptive, ce qui signifie qu'il peut détecter et interagir avec des machines sans aucune intervention humaine. Cela contraste avec les systèmes traditionnels de détection d'intrusion qui s'appuient sur l'apport

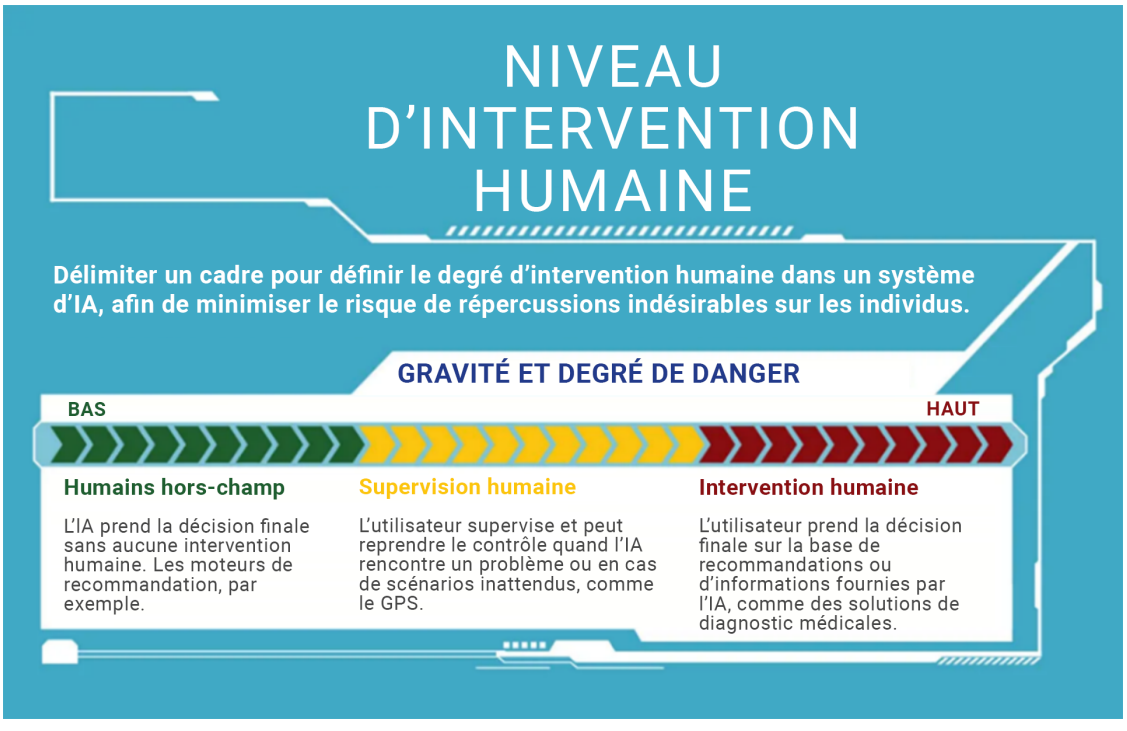
46 Stankovich M. (2021). Regulating AI and Big Data Deployment in Healthcare: Proposing Robust and Sustainable Solutions for Developing Countries' Governments, disponible sur : https://www.dai.com/uploads/regulating-ai-cda.pdf

47 Selon le groupe d'experts de haut niveau de la Commission européenne sur l'IA, « le HITL fait référence à la capacité d'intervention humaine dans chaque cycle de décision du système, ce qui, dans de nombreux cas, n'est ni possible ni souhaitable. HOTL fait référence à la capacité d'intervention humaine pendant le cycle de conception du système et de surveillance du fonctionnement du système. HIC fait référence à la capacité de superviser l'activité globale du système d'IA (y compris ses impacts économiques, sociétaux, juridiques et éthiques plus larges) et à la capacité de décider quand et comment utiliser le système dans une situation particulière », voir : https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1. Voir également le modèle de cadre de gouvernance de l'IA de Singapour, disponible sur: https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGModelAIGovFramework2.pdf

humain pour anticiper les cyberattaques.

- **Le terme « humain en boucle »** fait référence à l'implication des humains dans des rôles de supervision où ils ont la capacité de prendre le contrôle lorsque les modèles d'IA rencontrent des situations inattendues ou indésirables. Pour comprendre cela, prenons l'exemple de l'utilisation d'un système de navigation GPS. Le système GPS planifie l'itinéraire du point A au point B et offre diverses options en fonction de paramètres tels que la distance la plus courte, le temps le plus court ou l'évitement des routes à péage. Cependant, pendant la navigation, le conducteur peut toujours prendre le contrôle du GPS et modifier les paramètres de navigation en cas d'embouteillages inattendus.

Figure 8. Niveau d'implication humaine dans le déploiement de l'IA



Source : IMDA, Singapour

Il convient de noter que le principe HITL a ses limites, en raison du biais d'automatisation abordé dans le module 3, lorsque les humains sont plus prédisposés à de simples décisions prises par des algorithmes, en particulier dans les cas où il y a un effet de boîte noire et où les humains pourraient ne pas être en mesure de comprendre pourquoi une décision a été prise.

5. Pourquoi la cybersécurité est-elle importante dans le contexte de l'IA ?

La cybersécurité est la gestion des risques pour la confidentialité, l'intégrité ou la disponibilité des données et des systèmes. C'est une question fondamentale pour toute technologie. Les processus/algorithmes d'IA traitent de manière inhérente de grands ensembles de données et produisent fréquemment des résultats ayant des répercussions à la fois virtuelles et tangibles. En plus des menaces traditionnelles, des vulnérabilités propres à l'IA ont été identifiées, notamment :

- L'empoisonnement des données pendant la phase de formation.[48]
- Les attaques d'entrée qui manipulent des données pour altérer le résultat.[49]

Les cyberattaques continuent d'augmenter en fréquence, en sophistication et en dépenses. En 2022, les entreprises ont besoin en moyenne de 207 jours pour détecter un incident de sécurité et de 70 jours pour le contenir. Alors que les entreprises continuent de déployer rapidement la technologie sur l'ensemble de la chaîne de valeur, le risque d'interruption des activités joue un rôle central. À la maison, les appareils intégrés de l'Internet des objets (IdO) continuent de présenter des risques importants, et le travail à distance introduit un mélange compliqué de vulnérabilités. Les acteurs malveillants peuvent compromettre les systèmes d'IA pour atteindre divers objectifs, tels que causer des dommages, échapper à la détection ou dégrader la confiance dans un système.[50]

Par rapport aux systèmes traditionnels, les systèmes alimentés par l'IA présentent des caractéristiques uniques qui peuvent être vulnérables aux cyberattaques, de manière non traditionnelle. Par exemple, les attaquants peuvent compromettre un ensemble de données de formation de sorte que l'« apprentissage » résultant du système ne se déroule pas comme prévu. Ce type d'attaque s'appelle l'empoisonnement de données et tire parti du processus de développement unique de l'IA, qui consiste à utiliser des données de grande taille. Il est donc important de prévoir une protection supplémentaire des systèmes d'IA. L'augmentation des capacités d'apprentissage dans les technologies d'IA, telles que l'apprentissage profond et l'apprentissage par renforcement, a un impact significatif sur la cybersécurité et permet des actions criminelles plus efficaces.[51] Par conséquent, la protection des systèmes d'IA doit être soigneusement examinée et les vulnérabilités possibles identifiées, pour pouvoir mettre en place des mesures de sécurité solides pour (a) se prémunir des attaques, mais aussi (b) détecter les attaques dès que possible, afin d'atténuer les risques et les préjudices importants.

48 Poremba S. (2021). Data Poisoning: When Attackers Turn AI and ML Against You, disponible sur : https://securityintelligence.com/articles/data-poisoning-ai-and-machine-learning/

49 Comiter M. (2019). Attacking Artificial Intelligence: AI's Security Vulnerability and What Policymakers Can Do About It, disponible sur : https://www.belfercenter.org/publication/AttackingAI

50 *Ibid.*

51 Kaloudi N., Li J. (2020). The AI-Based Cyber Threat Landscape: A Survey. ACM Computing Surveys (CSUR), 53, 1–34, disponible sur : https://www.researchgate.net/publication/339081899_The_AI-Based_Cyber_Threat_Landscape_A_Survey.

Les cyberattaques sur les systèmes d'IA se produisent à trois phases différentes du développement de l'IA : 1) la préparation des données, 2) la formation au modèle et 3) le déploiement du modèle :[52]

- Pendant la préparation des données, les attaquants peuvent cibler des composants ou des bibliothèques de préparation de données communs, ou obtenir un accès non autorisé au pipeline de traitement des données, à des fins de falsification.
- Pendant la phase de formation, les attaquants peuvent ajouter, supprimer ou modifier des données de formation (empoisonnement des données). Ce faisant, les attaquants influencent le modèle résultant.
- Les attaquants qui ont accès aux modèles peuvent apporter des modifications aux poids et aux algorithmes, au stade du déploiement du modèle (altération du modèle).[53]

Réglementation de la cybersécurité

La réglementation en matière de cybersécurité consiste en des directives qui protègent les technologies de l'information et les systèmes informatiques, pour obliger les entités des secteurs privé et public à protéger leurs systèmes d'information et leurs données contre les cyberattaques telles que les virus, les vers, les chevaux de Troie, le phishing, les attaques par déni de service (DOS), les accès non autorisés (vol de propriété intellectuelle ou d'informations confidentielles) et les attaques de systèmes de contrôle.[54]

Dans cet esprit, il est extrêmement important que les opérateurs judiciaires prennent en compte les différentes lois et réglementations en matière de cybersécurité, et évaluent l'impact de l'IA sur ces réglementations. Par exemple, les réseaux intelligents utilisant des systèmes d'IA améliorent considérablement la gestion de la consommation et de la distribution d'énergie au profit des consommateurs, des fournisseurs d'électricité et des gestionnaires de réseau. Néanmoins, l'amélioration des opérations et des services expose l'ensemble du réseau énergétique à de nouvelles difficultés en matière de communication et de sécurité des systèmes d'information. Les vulnérabilités des réseaux de communication et des systèmes d'information pourraient être exploitées pour des raisons financières ou politiques, afin de couper l'alimentation de vastes zones ou de lancer des cyberattaques contre des unités productrices d'énergie. L'IA peut servir dans des campagnes de désinformation et d'anti-information susceptibles d'être utilisées pour les coupures d'Internet et la restriction de l'accès à l'information.[55] L'encadré ci-dessous décrit les dangers associés aux exemples contradictoires utilisés par les modèles de ML.

52 Gartner (2020). Artificial Intelligence Under Attack: How to Identify and Mitigate Threats to Machine Learning, disponible sur : https://www.gartner.com/en/documents/3989271; Wolff J. (2020). How to Improve Cybersecurity for Artificial Intelligence, disponible sur : https://www.brookings.edu/articles/how-to-improve-cybersecurity-for-artificial-intelligence/

53 *Ibid.*

54 EU Cybersecurity Act (2019), disponible sur : https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32019R0881&qid=1694014957942. Voir aussi : https://web.archive.org/web/20100613183200/http://www.privacyrights.org/ar/ChronDataBreaches.htm

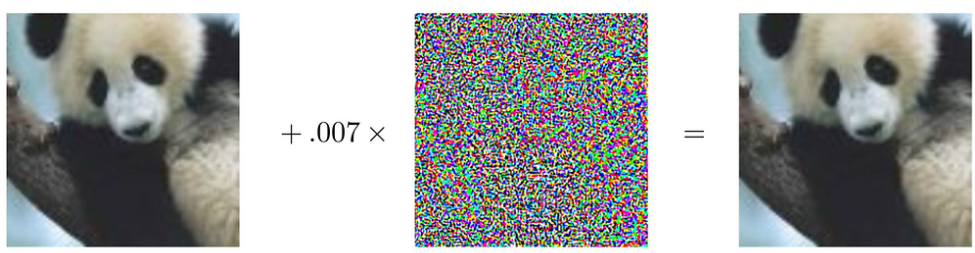
55 EU Cybersecurity Act (2019), disponible sur : https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32019R0881&qid=1694014957942. Voir aussi : https://web.archive.org/web/20100613183200/http://www.privacyrights.org/ar/ChronDataBreaches.htm.

Approfondissement : Les dangers associés aux exemples contradictoires utilisés par les modèles de ML

Les exemples contradictoires sont des entrées utilisées par des modèles de ML qui sont délibérément générés par un attaquant pour faire en sorte que le modèle se trompe tout en présentant un niveau de confiance élevé. Étant donné que de nombreux modèles de ML, y compris les réseaux neuronaux de pointe [56], sont sensibles aux instances contradictoires, cela peut constituer une menace sérieuse pour la sécurité et la robustesse de l'IA.

Les exemples peuvent être imperceptibles. L'image d'un panda ci-dessous a subi une petite perturbation indétectable, ou « insertion d'entrée contradictoire ». Elle est destinée à tromper l'algorithme de classification d'image. Cela a permis à l'ordinateur d'avoir un niveau de confiance de 99,3 % pour classer le panda comme un gibbon.

Des exemples contradictoires peuvent être produits en imprimant une image sur du papier ordinaire et en la prenant en photo à l'aide d'un smartphone avec une résolution ordinaire. Un autocollant antagoniste sur un panneau « stop » pourrait tromper une voiture autonome, en lui faisant croire qu'il s'agit d'un panneau « céder le passage » ou de tout autre signe.[57]



Source : OCDE, AI in society, disponible sur : https://www.oecd-ilibrary.org/science-and-technology/artificial-intelligence-in-society_eedfee77

Les faiblesses de ces systèmes d'IA face à des exemples contradictoires ont des effets néfastes sur la sécurité des systèmes d'IA. L'adoption de systèmes critiques tels que ceux utilisés dans le transport autonome, l'imagerie médicale, la sécurité et la surveillance pourrait sérieusement souffrir de l'existence de cas où des perturbations subtiles mais ciblées induisent les modèles dans des erreurs de calcul flagrantes et des décisions incorrectes.

56 Les réseaux neuronaux sont un type de technique de ML qui permet aux ordinateurs d'apprendre à effectuer des tâches en analysant des exemples d'entraînement. En règle générale, ces exemples sont pré-étiquetés. Par exemple, un système de reconnaissance d'objets peut recevoir des milliers d'images étiquetées d'objets tels que des voitures, des maisons et des tasses à café. Grâce à l'analyse, il peut identifier des motifs dans les images qui correspondent à des étiquettes spécifiques. Un réseau neuronal est conçu pour ressembler vaguement à la structure du cerveau humain, avec des milliers ou des millions de nœuds de traitement interconnectés. Ces nœuds sont généralement organisés en couches et les données les traversent dans une seule direction, ce qui les rend « feed-forward ». Chaque nœud reçoit des données des nœuds de la couche située en dessous et envoie des données aux nœuds de la couche située au-dessus. Définition donnée dans Hardesty L. (2017). Explained neural networks, disponible sur : https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414

57 Goodfellow I. J., Shlens J., Szegedy (2015). Explaining and harnessing adversarial examples. International Conference on Learning Representation, disponible sur : https://arxiv.org/pdf/1412.6572.pdf ; Kurakin A., Goodfellow I., Bengio S. (2017). Adversarial examples in the physical world. Atelier de l'ICLR, disponible sur : https://arxiv.org/abs/1607.02533

6. Activités

Ces activités de groupe visent à encourager les participants à la formation à discuter et à débattre de diverses questions pertinentes liées à l'IA et à ses éléments constitutifs, ainsi qu'aux risques associés au déploiement de l'IA dans le système judiciaire.

Activité 1 - Temps de discussion

Veuillez discuter de ces questions avec les autres participants à la formation :

- Comment un défendeur peut-il légitimement contester la logique d'un algorithme si le code source et (le cas échéant) les données de formation ou les ensembles de données nécessaires pour reproduire les résultats ne sont pas mis à sa disposition ?
- Quelles informations doivent être fournies au défendeur pour contester la logique d'un algorithme ?
- Est-il suffisant pour lui d'avoir simplement accès aux entrées et sorties générées par l'algorithme ?
- Le défendeur doit-il recevoir des informations sur la marge d'erreur du ou des algorithmes utilisés ?

Activité 2 - Temps de discussion

Veuillez discuter de ces questions avec les autres participants à la formation :

- Comment les tribunaux peuvent-ils faire respecter une procédure régulière si l'algorithme déploie l'apprentissage automatique et que personne, pas même le développeur, ne comprend complètement l'« analyse » du ML ?
- Comment les tribunaux peuvent-ils évaluer la précision des algorithmes, en particulier lorsqu'ils prévoient le comportement humain futur ?

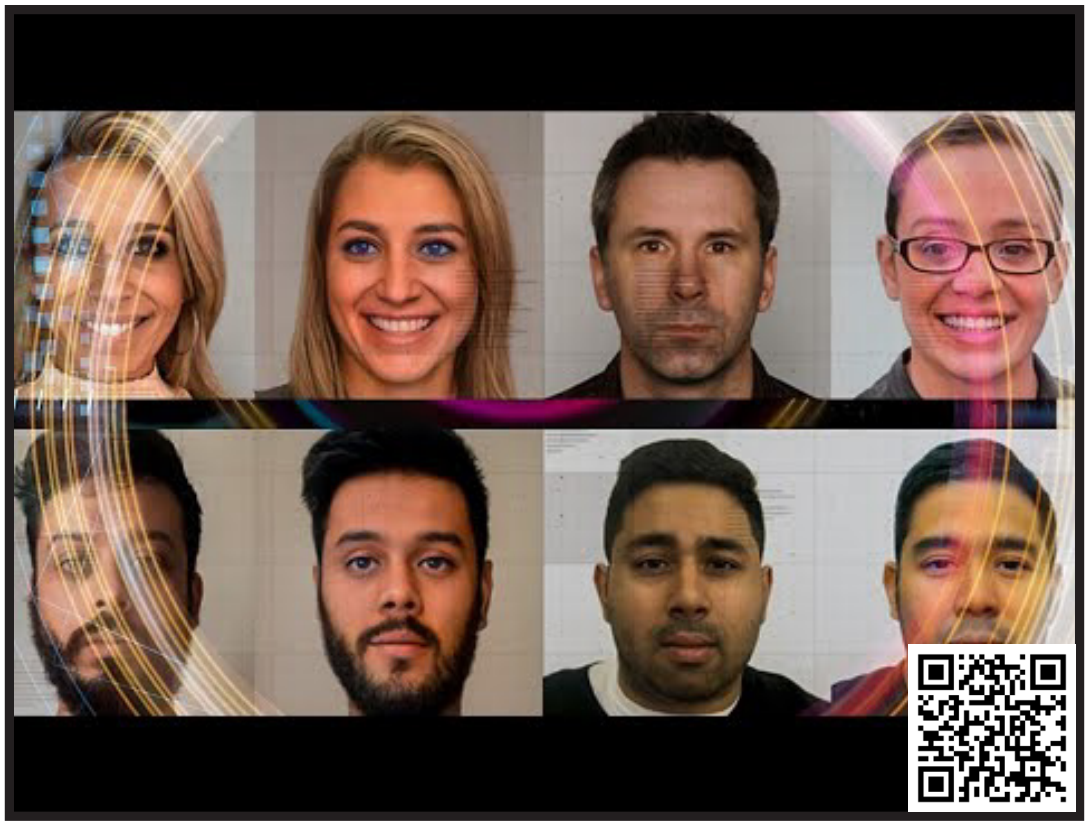
Activité 3 - Temps de discussion

Veuillez discuter de ces questions avec les autres participants à la formation :

- Que se passe-t-il si les algorithmes ont été formés avec des ensembles de données antérieurs à la dernière jurisprudence ?
- Quel est le régime d'admissibilité des preuves recueillies à l'aide d'algorithmes, en particulier par les enquêteurs de police ?
- Cette collecte peut-elle être considérée comme irrégulière ou injuste ?
- Les données ont-elles été collectées conformément aux lois sur la protection des données et, dans le cas contraire, comment l'algorithme doit-il être traité ?

Activité 4 - Temps de discussion

Les participants à la formation regardent la vidéo et discutent des différents impacts sociétaux des préjugés liés à l'IA.



Source : BBC, https://youtu.be/b4UyT85H3Hg

Activité 5 - Les participants à la formation discutent des questions suivantes liées à l'application de l'IA dans les opérations judiciaires.

Souvent, les modèles d'IA ne peuvent pas fournir de justifications compréhensibles par l'homme à leurs décisions ou recommandations. De nombreux algorithmes d'IA « apprennent par eux-mêmes », on parle alors de ML auto-apprenant (lire et se référer au principe de l'humain dans la boucle, module 4). Discutez avec d'autres participants à la formation pour tenter de répondre aux questions suivantes :

- Comment votre capacité à comprendre ou à sonder les résultats d'un modèle d'IA affecte-t-elle sa valeur probante dans les procédures judiciaires ?
- Quelles responsabilités juridiques et sociales devrions-nous donner aux algorithmes protégés par une «impartialité» dérivée de données statistiques?
- Qui est responsable lorsque l'IA se trompe ?

Il y a beaucoup de débat quant à savoir qui, parmi les différents acteurs tout au long du cycle de vie de la conception, du développement et du déploiement de l'IA et des systèmes autonomes, doit être tenu responsable de tout dommage qui pourrait être causé. Un écosystème d'IA complexe et la multiplicité des acteurs rendent difficile la détermination du responsable des dommages causés au(x) demandeur(s), car ces dommages peuvent résulter d'une série de causes imbriquées par plusieurs acteurs.

- L'autonomie et les capacités d'auto-apprentissage modifieraient-elles la chaîne de responsabilité du producteur ou du développeur, en tant que « machine pilotée

par l'IA ou autrement automatisée qui, après avoir pris en compte certaines données, a évolué au fil du temps grâce à ses capacités d'auto-apprentissage rendues possibles par des techniques de ML et/ou d'apprentissage profond, a pris une décision autonome et a causé un préjudice à la vie, à la santé ou aux biens d'un être humain » ?

- Comment les capacités des systèmes de ML non supervisés affecteront-elles les questions de responsabilité ? Par exemple, se pose le défi de la dépendance à l'égard des données externes – lorsque ces données sont fournies par des sources externes, prouver à la fois la défectuosité et un lien de causalité avec le préjudice ou le dommage subi pourrait s'avérer très difficile.
- Est-ce que « l'insertion d'une couche de code impénétrable, non intuitif et statistiquement modifié entre un décideur humain et les conséquences de cette décision, fait que l'IA perturbe notre compréhension typique de la responsabilité des choix qui ont mal tourné » ? Ou le producteur/programmeur devrait-il prévoir la perte ou les dommages potentiels, même lorsqu'il peut être difficile d'anticiper, en particulier dans des circonstances inhabituelles, les actions d'un système autonome ? Ces questions deviendront de plus en plus sensibles à mesure que de plus en plus de décisions autonomes seront prises par les systèmes d'IA.
- Quels niveaux d'incertitude dans les résultats du modèle de ML les tribunaux accepteront-ils et à quelles conditions ? Comment les différents niveaux de certitude du modèle de ML se rapportent-ils à diverses normes de preuve ? (c.-à-d., quand le degré de corrélation X, Y ou Z [moins X %, Y % ou Z % d'incertitude] est-il égal à une norme légale de preuve (telle que « claire et convaincante », « prépondérance de la preuve » ou « au-delà de tout doute raisonnable ») ?
- Cette question sera-t-elle laissée à la discrétion des tribunaux ou des juges ? Ou des normes nationales ou régionales seront-elles élaborées et mises en œuvre ? Ces exigences doivent-elles être rigides ou flexibles ?
- La plupart des processus de ML sont itératifs et auto-apprenants, car ils ajustent les formules et la précision en traitant de nouvelles données. Pensez-vous que de telles applications de ML, étant donné qu'elles changent constamment, pourraient devoir faire l'objet d'une nouvelle enquête sur une base continue, si elles sont utilisées comme preuves ?

Source: AAAS, Artificial Intelligence and the Courts: Materials for Judges, disponible sur : https://www.aaas.org/ai2/projects/law/judicialpapers.

7. Ressources

1. AAAS, Artificial Intelligence and the Courts: Materials for Judges, disponible sur : https://www.aaas.org/ai2/projects/law/judicialpapers
2. AccessNow (2018). Toronto Declaration on Protecting the Rights to Equality and Non-Discrimination in Machine Learning Systems, disponible sur : https://www.accessnow.org/press-release/the-toronto-declaration-protecting-the-rights-to-equality-and-non-discrimination-in-machine-learning-systems/
3. Amnesty International (2017). Artificial Intelligence for Good, disponible sur : https://www.amnesty.org/en/latest/news/2017/06/artificial-intelligence-for-good
4. Amnesty International (2021). Xenophobic Machines: Discrimination through unregulated use of algorithms in the Dutch childcare benefits scandal, disponible sur : https://www.amnesty.org/en/documents/eur35/4686/2021/en/#:~:text=Xenophobic%20machines%3A%20Discrimination%20through%20unregulated%20use%20of%20algorithms,their%20processes%20in%20the%20hope%20of%20detecting%20fraud
5. Bell F., Bennett Moses L., Legg M., Silove J., Zalnieriute M. (2022). AI Decision-Making and the Courts: A Guide for Judges, Tribunal Members and Court Administrators, Australasian Institute of Judicial Administration, disponible sur : https://ssrn.com/abstract=4162985
6. Buolamwini J., Gebru T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR, 81, 77–91, disponible sur : https://proceedings.mlr.press/v81/buolamwini18a.html
7. Burgess M. (2023). The Security Hole at the Heart of ChatGPT and Bing. disponible sur : https://www.wired.co.uk/article/chatgpt-prompt-injection-attack-security
8. Burrell J. (2015). How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms disponible sur : https://ssrn.com/abstract=2660674 or http://dx.doi.org/10.2139/ssrn.2660674
9. Conn A. (2017). Artificial Intelligence: The Challenge to Keep It Safe., disponible sur : https://futureoflife.org/ai/safety-principle/European Union Agency for Fundamental Rights (2019), disponible sur : https://fra.europa.eu/sites/default/files/fra_uploads/fra-2019-data-quality-and-ai_en.pdf
10. IEEE (2019). Ethically Aligned Design. A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. First Edition, disponible sur : https://standards.ieee.org/wp-content/uploads/import/documents/other/ead1e.pdf
11. International Commissioner's Office, Explaining decision made with AI, disponible sur : https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/
12. McGregor L., Murray D., Ng V. (2019). International Human Rights Law as a Framework for Algorithmic Accountability, International & Comparative Law Quarterly, 68(2), 309–343, disponible sur : www.cambridge.org/core/journals/international-and-comparative-law-quarterly/article/international-human-rights-law-as-a-framework-for-algorithmic-accountability/1D6D0A456B36BA7512A6AFF17F16E9B6
13. National Science and Technology Council: Committee on Technology (2016). Preparing for the Future of Artificial Intelligence. Washington, D.C.: Executive Office of the President, 2016. https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf

14. Noble S. U. (2018). Algorithms of Oppression: How Search Engines Reinforce Racism, New York University Press.

15. Obermeyer Z., Powers B., Vogeli C., Mullainathan S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations, Science, 366(6464), 447–453, disponible sur : https://www.science.org/doi/10.1126/science.aax2342

16. OECD (2022). Framework for the Classification of AI systems, disponible sur : https://www.oecd.org/publications/oecd-framework-for-the-classification-of-ai-systems-cb6d9eca-en.htm.

17. O'Neil C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, New York: Crown.

18. Stanford (2022). Artificial Intelligence Index Report, disponible sur : 2022-AI-Index-Report_Master.pdf (stanford.edu)

19. The Alan Turing Institute, Human Rights, Democracy, and the Rule of Law Assurance Framework for AI Systems: A proposal prepared for the Council of Europe's Ad hoc Committee on Artificial Intelligence, disponible sur : https://www.turing.ac.uk/news/publications/ai-human-rights-democracy-and-rule-law-primer-prepared-council-europe

20. The Royal Society (2012). Machine Learning: The Power and Promise of Computers that Learn by Example, disponible sur : https://royalsociety.org/~/media/policy/projects/machine-learning/publications/machine-learning-report.pdf 16

21. UNESCO (2021). Recommendation on the Ethics of Artificial Intelligence, disponible sur : https://unesdoc.unesco.org/ark:/48223/pf0000380455

22. Ward J. (2019). 10 Things Judges Should Know About AI, Judicature, 103(1), disponible sur : https://judicature.duke.edu/articles/10-things-judges-should-know-about-ai.

23. Weinberger D. (2017). Our Machines Now Have Knowledge We'll Never Understand, disponible sur : https://www.wired.com/story/our-machines-now-have-knowledge-well-never-understand/

24. Wong A. (2020). The Laws and Regulation of AI and Autonomous Systems. In: Strous L., Johnson R., Grier D. A., Swade D. (eds) Unimagined Futures – ICT Opportunities and Challenges, IFIP Advances in Information and Communication Technology(), 555, disponible sur : https://link.springer.com/chapter/10.1007/978-3-030-64246-4_4

25. Wong A. (2021). Ethics and Regulation of Artificial Intelligence. In: Mercier-Laurent E., Kayalica M.Ö., Owoc M.L. (eds) Artificial Intelligence for Knowledge Management, AI4KM, IFIP Advances in Information and Communication Technology, 614, disponible sur : https://www.researchgate.net/publication/352477342_Ethics_and_Regulation_of_Artificial_Intelligence

26. Wong A. (2023). Generative AI: The Global debate and controversies on use of copyrighted content as training data, disponible sur : https://unctad.org/news/cstd-dialogue-anthony-wong

Module 2

Adoption de l'IA par le pouvoir judiciaire

Le module 2 traite de l'adoption de l'IA dans le système judiciaire. Il présente les différentes applications de l'IA dans le système judiciaire, telles que la découverte électronique et l'examen des documents, l'utilisation de l'IA générative pour aider à la rédaction de documents, l'analyse prédictive, les outils d'évaluation des risques, le règlement des différends, la reconnaissance linguistique, la gestion numérique des dossiers et des affaires. Le module met ensuite en évidence des études de cas sur le déploiement de l'IA dans le système judiciaire, en discutant de certaines des opportunités et défis rencontrés par les systèmes judiciaires du monde entier dans l'utilisation de l'IA.

Qu'allez-vous apprendre ?

Après avoir terminé ce module, les participants seront en mesure de :

- Comprendre les différentes applications de l'IA dans le système judiciaire.
- Comprendre les défis et les opportunités liés au déploiement des systèmes d'IA dans le système judiciaire, à travers les études de cas présentées dans le module.

Les avocats, les cabinets d'avocats, les tribunaux et les agences gouvernementales utilisent l'IA à des fins différentes. Par exemple, les avocats utilisent l'IA pour la recherche juridique et pour trouver des précédents pertinents, afin de renforcer leurs arguments. Les cabinets d'avocats, quant à eux, l'utilisent pour prévoir les résultats des affaires, évaluer les chances de réussite et conseiller les clients en matière de procédures judiciaires. L'IA a également été utilisée par les avocats pour prévoir comment des juges particuliers se prononceraient sur divers sujets. De même, les entités gouvernementales utilisent l'IA pour évaluer la probabilité de succès dans la poursuite de mesures particulières contre les particuliers et les entreprises, comme dans les affaires fiscales.

À Buenos Aires, en Argentine, les procureurs des impôts utilisent des systèmes d'IA pour rédiger des décisions de justice.[58] Le tribunal Internet de Hangzhou (Chine) a mis en place un système d'analyse des preuves qui utilise des technologies de pointe telles que la blockchain, l'IA, les données de masse et le nuage informatique. Ce système analyse et compare tous les éléments de preuve présentés par les deux parties, les transformant en une liste de preuves et de pièces pertinentes. Les informations sont ensuite triées et classées avant d'être présentées visuellement au juge humain pour examen.[59] Au Mexique, les tribunaux peuvent utiliser l'IA pour donner des conseils sur la détermination du droit à une forme de sécurité sociale ou non. Un programme nommé Expertius fonde ses calculs sur des informations sur les réclamations passées, les résultats des réclamations, les procès-verbaux d'audience et les jugements définitifs.[60]

On trouve un autre exemple dans le système judiciaire colombien, qui explore les moyens de réduire la charge de travail des juges humains. La Cour constitutionnelle colombienne développe actuellement un système d'IA nommé PretorIA, pour aider à la sélection des tuteurs légaux. PretorIA ne remplace pas les humains dans ce processus, mais rationalise plutôt la tâche en analysant les peines de tutelle et en fournissant des informations plus précises aux personnes chargées d'identifier les personnes susceptibles d'être désignées comme tutrices.[61]

La pression pour une justice efficace dans un contexte de contraintes budgétaires

Comme pour les autres services aux consommateurs, les tribunaux sont censés fournir des services judiciaires modernes, numériques et réactifs, tout en réduisant la durée des affaires, dans un contexte de contraintes budgétaires croissantes. Les systèmes judiciaires basés sur l'IA promettent d'améliorer la qualité des services tout en réduisant les dépenses liées aux opérations judiciaires.[62]

58 Dejusticia (2021). Conoce nuestra Investigación sobre PretorIA, la tecnología que incorpora la Inteligencia Artificial a la Corte Constitucional, disponible sur : https://www.dejusticia.org/conoce-nuestra-investigacion-sobre-pretoria-la-tecnologia-que-incorpora-la-inteligencia-artificial-a-la-corte-constitucional/

59 Xuan H. (2021). Accès en un clic aux résultats de l'analyse des preuves. Hangzhou Internet Court Launches Intelligent Evidence Analysis System, China Courts Network, disponible sur : https://www.chinacourt.org/article/detail/2019/12/id/4747683.shtml

60 Goretty C., Martinez B. (2012). La inteligencia artificial y su aplicación al campo del Derecho, Alegatos, 82, 827–846, disponible sur : http://alegatos.azc.uam.mx/index.php/ra/article/viewFile/205/184

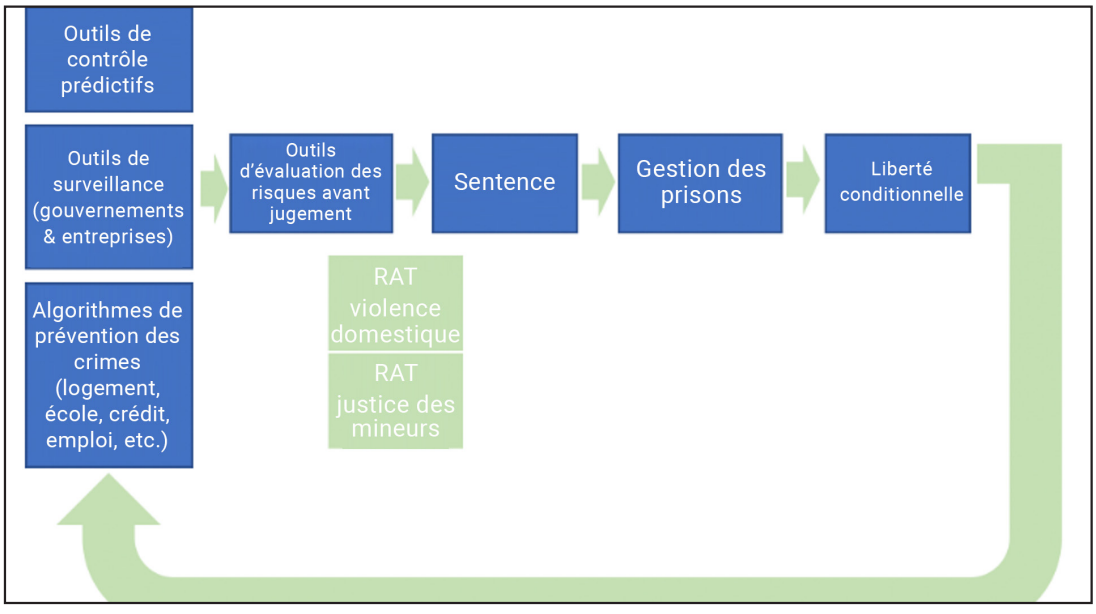
61 https://www.dejusticia.org/conoce-nuestra-investigacion-sobre-pretoria-la-tecnologia-que-incorpora-la-inteligencia-artificial-a-la-corte-constitucional/

62 Wu J. (2019). AI Goes to Court: The Growing Landscape of AI for Access to Justice, disponible sur : https://medium.com/legal-design-and-innovation/ai-goes-to-court-the-growing-landscape-of-ai-for-access-to-justice-3f58aca4306f

Lorsqu'ils sont déployés avec des garanties en matière de droits humains et d'éthique, les systèmes d'IA peuvent rendre les procédures juridiques plus accessibles à un groupe plus large de personnes, dans plusieurs langues et à moindre coût. Par exemple, les estimations montrent que l'utilisation du ML dans la découverte électronique, par la présentation de documents dans des groupes conceptuels, peut augmenter la vitesse d'examen de 15 % à 20 %. Il s'agit d'une économie importante.[63]

D'autre part, le développement et le déploiement de l'IA dans les opérations judiciaires peuvent avoir un impact sur les droits fondamentaux. Les technologies d'IA contiennent des biais intégrés (abordés dans le module 3), et ce sont souvent des boîtes noires (abordées dans le module 1). Par conséquent, l'état de droit et la préservation des droits humains doivent continuer d'être au premier plan de l'administration de la justice.[64]

Figure 9. Un cycle simpliste d'utilisation algorithmique en justice pénale



Source : EPIC, AI in the criminal justice system, disponible sur : https://epic.org/issues/ai/ai-in-the-criminal-justice-system/

La numérisation des documents de la Cour est une première étape essentielle vers l'utilisation de l'IA

La numérisation des documents judiciaires a permis aux tribunaux et à d'autres opérateurs judiciaires de compter sur l'assistance de l'IA pour les tâches administratives. Les algorithmes d'IA sont de plus en plus utilisés dans le contexte des systèmes de justice civile et pénale pour soutenir la prise de décision humaine.[65] Les systèmes d'IA sont testés pour identifier les modèles de prise de décision judiciaire complexe et prédire les résultats des décisions. À mesure qu'ils rassemblent et

63 Deloitte, Artificial intelligence and machine learning in e-discovery and beyond: Driving efficiencies in e-discovery using AI, disponible sur : https://www2.deloitte.com/ch/en/pages/forensics/articles/AI-and-machine-learning-in-E-discovery.html,

64 Sur l'impact de l'IA sur les droits humains lorsqu'elle est appliquée dans les systèmes judiciaires, voir également UNESCO (2021). Manuel de formation mondial pour les acteurs du judiciaire : normes juridiques internationales relatives à la liberté d'expression, l'accès à l'information et la sécurité des journalistes, Module 5, p. 164, disponible sur : https://unesdoc.unesco.org/ark:/48223/pf0000378755

65 Parlement européen (2019). A governance framework for algorithmic accountability and transparency, disponible sur : https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2019)624262

analysent de vastes réserves d'informations, identifient des modèles, prédisent des approches optimales, détectent des anomalies, classent les problèmes et rédigent des documents, la promesse est que les systèmes judiciaires deviendront plus efficaces et seront en mesure de prioriser le temps et les ressources, afin de rendre une justice opportune.

Dans le système de justice pénale, des modèles d'IA ont été déployés pour surveiller et reconnaître les défendeurs, appuyer les décisions de condamnation et de mise en liberté sous caution et soutenir l'évaluation des preuves.[66] La figure 9 donne un aperçu simplifié de l'utilisation algorithmique de l'IA dans le système de justice pénale.

Dans le système de justice civile, l'IA a été déployée dans les litiges familiaux, de logement, d'endettement, d'emploi et de consommation.[67] Les tribunaux civils recueillent de plus en plus de données sur l'administration, les plaidoiries, le comportement des plaideurs et les décisions. Cela offre la possibilité d'automatiser certaines fonctions judiciaires, telles que la gestion des dossiers, la planification des audiences et des procès, et la gestion des fonctions de jury, afin d'optimiser l'efficacité.[68] Par exemple, l'IA est utilisée pour rédiger des modèles de jugement pour les juges, faire des prédictions ou des recommandations de condamnation pour la mise en liberté sous caution, la détermination de la peine et les calculs financiers. Elle est également utilisée pour évaluer l'issue des affaires sur la base des activités passées des procureurs et des juges. Un outil d'IA peut fournir des informations à un juge qui tiennent compte d'une grande quantité de jurisprudences et peut réduire le temps de recherche dans la préparation des décisions.

À l'aide d'un algorithme d'IA créé par des chercheurs de l'Université catholique de Louvain (UCL), de l'Université de Sheffield et de l'Université de Pennsylvanie, les décisions judiciaires de la Cour européenne des droits de l'homme ont été anticipées avec une précision de 79 %.[69] Le Dr Nikolaos Aletras, qui a dirigé l'étude à UCL Computer Science, a expliqué que l'équipe « ne voyait pas l'IA remplacer les juges ou les avocats, mais pensait qu'elle pourrait être utile pour identifier rapidement les tendances dans les affaires qui mènent à certains résultats. Cela pourrait également être un outil précieux pour mettre en évidence les cas les plus susceptibles de constituer des violations de la Convention européenne des droits de l'homme ».[70]

Le défi des systèmes d'IA perçus comme plus objectifs que les humains

Cependant, compte tenu du nombre élevé d'affaires et du manque de ressources adéquates qui affligent la plupart des systèmes judiciaires, il existe un risque que les juges utilisent de manière inappropriée les systèmes de soutien basés sur l'IA pour « déléguer » des décisions à des systèmes technologiques qui n'ont pas été conçus à cette fin, mais qui sont perçus comme plus objectifs qu'ils ne le sont. Afin de ne pas compromettre le droit à un procès équitable, il convient

66 Završnik A. (2020). Justice pénale, systèmes d'intelligence artificielle et droits humains. Forum ERA. 20, 567-583, disponible sur : https://doi.org/10.1007/s12027-020-00602-0.
67 Cabral J. E, Chavan A., Clarke T. M., Greacen J., Hough B. R., Rexer L., Ribadeneyra L., Zorza R. (2012). Using Technology to enhance access to justice, disponible sur : http://jolt.law.harvard.edu/articles/pdf/v26/26HarvJLTech241.pdf
68 Martin A. (2010). Automated Debt-Collection Lawsuits Engulf Courts, disponible sur : https://www.nytimes.com/2010/07/13/business/13collection.html
69 UCL (2016). AI predicts outcomes of human rights trials, disponible sur : https://www.ucl.ac.uk/news/2016/oct/ai-predicts-outcomes-human-rights-trials
70 *Ibid.*

de prendre grand soin d'évaluer ce dont ces dispositifs sont capables et dans quelles conditions ils peuvent être déployés. Cela est particulièrement vrai lorsque de tels systèmes sont utilisés pour rendre des décisions en matière de libération conditionnelle. Dans un système de justice axé sur les algorithmes, les juges ne devraient pas être de simples applicateurs d'algorithmes, mais aussi leurs évaluateurs critiques. Le tableau ci-dessous décrit les principales implications positives et négatives de l'utilisation de l'ADM et de l'IA dans le système judiciaire.

Tableau 3. Implications positives et négatives de l'utilisation de l'IA dans le système judiciaire

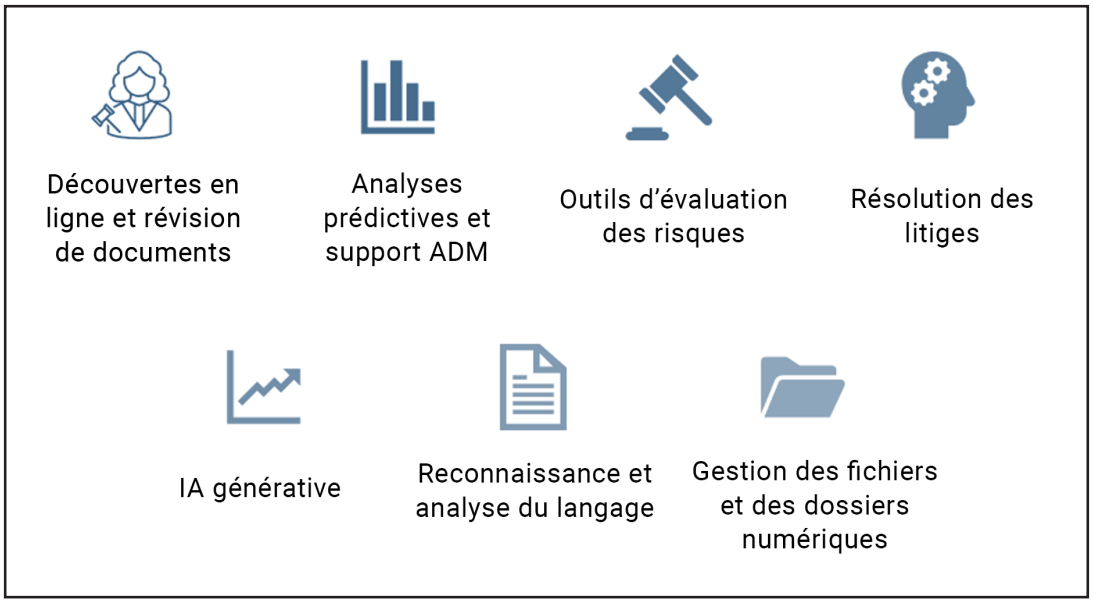
	Implications positives	Incidences négatives
Excellence judiciaire	Donne aux juges une analyse rapide de l'éventail des cas et des facteurs. Accélère la recherche et la rédaction. L'optimisation des processus, la réduction des coûts, l'agilité accrue, les gains de productivité, l'élimination du travail mécanique et répétitif augmentent la sécurité juridique.	Intégration de préjugés raciaux, de genre/sexe et d'autres types de préjugés. Réduction de la discrétion judiciaire et de l'élément humain dans la prise de décision. Complication de l'utilisation. Menace pour l'indépendance judiciaire, partialité automatisée. Le profilage des juges peut affecter le droit fondamental à la protection des données personnelles, créer des pressions et affecter l'indépendance judiciaire.
Confidentialité et sécurité	Protocoles de sécurité automatiques et nettoyage des données, ce qui permet une plus grande précision des résultats de l'IA.	Piratage, violations de données.
Propriété des données	Les données agrégées par les systèmes d'IA peuvent être utilisées pour identifier les tendances, les lacunes dans les services et l'innovation.[71]	Selon les propriétaires du système, les partenaires du secteur privé pourraient avoir accès aux données personnelles. Les données agrégées peuvent être utilisées pour cibler et discriminer des individus ou des groupes. Une réglementation limitée de la propriété des données limite la protection des droits et des réparations pour les personnes concernées par les systèmes d'IA.
État de droit	Empêche un intérêt puissant de s'emparer du système judiciaire.	Risque d'empiètement sur les droits fondamentaux, comme abordé dans le module 4. Menaces pour la démocratie telles que la désinformation, l'anti-information, les canulars, la propagande, la falsification profonde, les opérations d'influence ou la manipulation de l'opinion publique, principalement dans les processus électoraux.
Accès à la justice	Peut identifier des modèles de préjugés à l'encontre des groupes vulnérables, dans la prise de décision et les services. Peut rendre les délais des tribunaux plus rapides et plus prévisibles.	Pas uniformément à la disposition des parties pour analyser les données ou soutenir leur cas, en raison de problèmes d'infrastructure et d'accès (électricité, Internet, matériel). Le manque de formation des opérateurs judiciaires et des assistants pourrait avoir un impact sur les résultats positifs que l'IA pourrait apporter.

Adapté de UNDP (2021) Emerging Technologies and Judicial Integrity Toolkit for Judges.

Source : https://www.undp.org/asia-pacific/emerging-technologies-and-judicial-integrity

71 IBM (2021). Data aggregation involves gathering a significant amount of information from a database and presenting it in a more manageable and inclusive format, disponible sur : https://www.ibm.com/docs/en/tnpm/1.4.2?topic=data-aggregation

Figure 10. Principales applications de l'IA dans le système judiciaire



Source : Auteurs.

Découverte électronique et examen des documents

Les outils d'IA sont utilisés dans le système judiciaire pour identifier, trier et examiner (i) les règles juridiques, les avoirs juridiques et les constatations factuelles, (ii) les arguments expliquant les conclusions et les explications des motifs, et (iii) les considérations juridiques spécifiques et les éléments de preuve.

La découverte électronique est l'identification, la collecte et la production d'informations stockées électroniquement (ESI) en réponse à une demande de divulgation, dans une procédure judiciaire ou une enquête. Les ESI peuvent se composer de courriels, de documents, de présentations, de bases de données, de fichiers audio et vidéo et de sites web.[72]

Activité

Réfléchissez à la façon dont l'IA peut changer le processus de découverte et discutez-en avec d'autres participants à la formation.

Questions à prendre en compte : Quelles seront les normes d'admissibilité des déclarations ou d'autres preuves, ou des idées générées par l'IA et/ou invoquées (ou rejetées) par les humains ? Comment évaluerons-nous leur crédibilité ou leur authenticité ?

72 https://cdslegal.com/knowledge/the-basics-what-is-e-discovery/

La découverte électronique repose sur le regroupement, un exemple de ML non supervisé, où des éléments « similaires » (par exemple, des documents) sont regroupés afin que les utilisateurs puissent reconnaître leurs caractéristiques similaires et en apprendre davantage sur la composition de l'ensemble de données. Les utilisateurs n'ont aucun contrôle sur la ou les dimensions selon lesquelles la « similitude » est définie et n'ont pas à étiqueter d'exemples d'éléments dans chaque groupe pour former le système. Cependant, le concepteur du système doit spécifier les caractéristiques selon lesquelles la similitude des éléments doit être mesurée et le nombre de groupes.[73] Par exemple, si le système de ML est chargé d'identifier des informations sur le tennis et le baseball dans les fichiers, l'algorithme regroupera également des fichiers contenant des informations sur toutes sortes de sports.[74] De même, une recherche de « petite enveloppe brune » ou de « graisse » regroupera des informations sur tout ce qui concerne la corruption.[75]

La recherche conceptuelle est une autre méthode de ML non supervisée utilisée dans la découverte électronique, où l'ordinateur apprend le contexte dans lequel les mots sont utilisés et modélise les relations entre les mots. Les utilisateurs peuvent ensuite effectuer une recherche par signification et non par termes individuels. Il est probable qu'un document contenant des mots tels que « avocat », « contrat » ou « litige civil » soit un document juridique. L'utilisation de l'un de ces mots peut conduire à la conclusion que le sujet du document est juridique.[76]

L'**examen assité par la technologie (TAR)** ou codage prédictif est une technique de ML supervisée dans laquelle les ordinateurs apprennent à distinguer les documents pertinents des documents non pertinents, en fonction du codage effectué par les examinateurs humains, puis classent les documents non étiquetés sans assistance.[77] Par exemple, CLAUDETTE (Automated CLAUse DETectEr) est un projet de recherche interdisciplinaire hébergé au Département de droit de l'Institut universitaire européen et une plateforme d'analyse et d'annotation automatisée des documents juridiques et de détection d'anomalies.[78]

En outre, les outils d'IA peuvent être déployés pour anonymiser les informations personnelles, confidentielles ou privilégiées incluses dans les enregistrements électroniques. Cela peut faciliter le respect des réglementations en matière de protection des données.[79]

Activité :

Comment fonctionne CLAUDETTE ?

L'objectif de CLAUDETTE est de responsabiliser les consommateurs et la société civile, en développant à terme des outils côté utilisateur qui permettent à chacun d'évaluer facilement l'équité des contrats de consommation et des réglementations en matière de confidentialité, avant d'utiliser les plateformes Internet. La technologie est actuellement au stade expérimental en laboratoire, et les participants à la formation peuvent y accéder ici : http://claudette.eui.eu/demo

73 EDRM (2021). The Use of Artificial Intelligence in eDiscovery, disponible sur : https://edrm.net/download/152621/75 IBM (2021). L'agrégation de données implique la collecte d'une quantité importante d'informations à partir d'une base de données et leur présentation dans un format plus facile à gérer

74 Deloitte. Artificial intelligence and machine learning in e-discovery and beyondnDriving efficiencies in e-discovery using AI, disponible sur : https://www2.deloitte.com/ch/en/pages/forensics/articles/AI-and-machine-learning-in-E-discovery.html

75 *Ibid.*

76 EDRM (2021). The Use of Artificial Intelligence in eDiscovery, disponible sur : https://edrm.net/download/152621/

77 *Ibid.*

78 EUI. CLAUDETTE, disponible sur : http://claudette.eui.eu/about/index.html

79 EDRM (2021). The Use of Artificial Intelligence in eDiscovery, disponible sur : https://edrm.net/download/152621/

Les participants à la formation regardent la vidéo (http://claudette.eui.eu/claudette.mp4) et se demandent si des plateformes similaires existent dans leurs juridictions respectives. Quels sont les avantages et les inconvénients du TAR ?

Analyse prédictive et assistance ADM

Fréquemment, les systèmes d'IA sont utilisés comme outils de prévision. Ils analysent de grandes quantités de données, dont des données historiques, pour évaluer les risques et prédire les tendances futures, à l'aide d'algorithmes. Les données de formation peuvent contenir des casiers judiciaires, des dossiers d'arrestation, des statistiques sur la criminalité, des dossiers d'interventions policières dans certains quartiers, des publications sur les réseaux sociaux, des données de communication et des dossiers de voyage. Les systèmes prédictifs peuvent aider les juges à mieux connaître les tendances de la jurisprudence et à anticiper l'évolution d'une éventuelle décision dans le contexte de la jurisprudence.[80]

L'analyse prédictive est la catégorie générale d'outils et de modèles statistiques, par exemple les systèmes de ML, qui utilisent et analysent des données historiques pour créer des prédictions sur l'avenir, afin de guider la prise de décision. Ces prédictions peuvent être à faible risque (recommandation de film), à risque moyen (acceptation de demande de prêt) ou à risque élevé (prédiction du défendeur le plus susceptible d'adopter un comportement particulier).[81]

Le développement d'applications d'IA qui prévoient comment un tribunal statuera sur une réclamation, une affaire ou un règlement augmente rapidement dans le secteur judiciaire. Par exemple, les technologies d'IA sont déjà utilisées pour dresser le profil des personnes, identifier les lieux susceptibles d'être des sites d'activité criminelle ou signaler les futurs récidivistes.[82] Ces pratiques sont très controversées, comme expliqué dans les modules 3 et 4.

Par exemple, le système EXPERTIUS, au Mexique, conseille les juges et les greffiers pour savoir si un demandeur est éligible à une pension. Le programme se compose de trois modules : (1) il offre aux juges et aux greffiers la possibilité de comprendre le processus (le module tutoriel), (2) il permet aux utilisateurs de fournir des preuves à l'appui de leur cas et d'attribuer des « poids » à chaque pièce justificative (le module inférentiel), et (3), il permet aux utilisateurs de calculer le montant de la pension auquel ils ont droit en fonction de critères socio-économiques spécifiés (le module financier).[83]

80 Committee of Experts on Internet Intermediaries (MSI-NET) (2018). Algorithms and Human Rights: Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications, Étude du Conseil de l'Europe, DGI/2017/12, disponible sur : https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5

81 AAAS, Artificial Intelligence and the Courts: Materials for Judges, disponible sur : https://www.aaas.org/ai2/projects/law/judicialpapers

82 RAND (2013). Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operation, disponible sur : www.rand.org/content/dam/rand/pubs/research_reports/RR200/RR233/RAND_RR233.pdf.

83 Bell F., Bennett Moses L., Legg M., Silove J., Zalnieriute M. (2022). AI Decision-Making and the Courts: A Guide for Judges, Tribunal Members and Court Administrators, Australasian Institute of Judicial Administration, disponible sur : https://ssrn.com/abstract=4162985

Étude de cas : le cas du système australien de scission

Un groupe d'experts et d'avocats en IA a développé le système Split-Up, utilisé dans les tribunaux australiens du droit de la famille. Le système Split-Up utilise un raisonnement basé sur des règles en conjonction avec des réseaux neuronaux, pour anticiper les résultats des litiges fonciers en matière de divorce et d'autres questions de droit de la famille.

Le système de scission est utilisé par les juges pour soutenir leur prise de décision en les aidant à identifier les biens matrimoniaux qui devraient être inclus dans un règlement. Le système aide le juge à déterminer le pourcentage du pool commun que chaque partie devrait recevoir en fonction de facteurs tels que les contributions, les sources de revenus et les besoins futurs. Le système analyse 94 éléments clés à l'aide de techniques statistiques basées sur l'architecture des réseaux neuronaux. Le juge peut alors proposer une ordonnance de propriété finale basée sur cette analyse. Le système vise également à fournir des justifications claires pour ses décisions.

Un défi en termes de biais, lors de l'utilisation de systèmes tels que Split-Up, est que les données utilisées dans ce contexte (les litiges de divorce sont généralement marqués par des déséquilibres entre les sexes et les données historiques peuvent présenter un schéma de discrimination) pourraient être lues comme une vérité fondamentale par les machines. Les opérateurs judiciaires doivent être informés de ces défis et risques liés aux systèmes d'IA tels que Split-Up.

Source : Zeleznikow J., Stranieri A. (1995). The split-up system: integrating neural networks and rule-based reasoning in the legal domain, ICAIL '95: Proceedings of the 5th international conference on Artificial intelligence and law, 185–194, disponible sur : https://dl.acm.org/doi/10.1145/222092.222235

Outils d'évaluation des risques (prédiction des risques, modèles de risques et notation sociale)

De plus en plus, on recourt à des outils d'évaluation des risques basés sur les données pour anticiper la probabilité d'un comportement criminel futur. Dans plusieurs pays, ces technologies sont utilisées pour faciliter la prise de décision dans le système de justice pénale, notamment les jugements concernant la détermination de la peine, la libération sous caution et les limitations post-sentence pour les personnes jugées susceptibles de commettre d'autres crimes. Ces outils exploitent les données historiques pour évaluer la probabilité qu'une personne présente un risque « élevé », « moyen » ou « faible » de manquer ses dates de comparution ou de se faire arrêter à nouveau. L'algorithme prend en compte des facteurs tels que le casier judiciaire et l'âge au moment de l'arrestation, et génère une note que les juges utilisent pour décider du maintien en prison ou de la libération.[84]

Pour évaluer le risque de récidive d'une personne et identifier les domaines d'intervention, des outils d'évaluation des risques sont utilisés à différentes phases du processus judiciaire. Par exemple :

i) Avant le procès, pour guider les choix sur la libération en attente de résolution ou d'incarcération.

84 Wykstra S. (2018). Bail reform, which could save millions of unconvicted people from jail, explained, disponible sur : https://www.vox.com/future-perfect/2018/10/17/17955306/bail-reform-criminal-justice-inequality

ii) Par les services de probation et de libération conditionnelle, pour déterminer le niveau approprié de surveillance, qui peut inclure la surveillance électronique et la détention à domicile.

iii) Dans le cadre des plans de réinsertion et de supervision, les gestionnaires de cas et les prestataires de traitement déploient des évaluations des risques pour identifier les besoins des clients et les mettre en relation avec les bons services.[85]

Selon leurs partisans, les outils d'évaluation des risques rendent le système de justice pénale plus équitable.[86] Ils soutiennent que l'IA pourrait remplacer l'intuition et les préjugés des juges, en particulier les préjugés raciaux, par le biais d'une note d'évaluation des risques qui semble plus « objective ».[87]

Cependant, dans la pratique, de nombreuses études ont montré que ces outils pourraient intégrer et amplifier les préjugés envers les populations marginalisées et vulnérables. Plusieurs droits humains peuvent être affectés par l'utilisation de l'IA dans le système de justice pénale, notamment les droits à l'égalité et à la non-discrimination, à l'égalité devant la loi, à la sécurité et à la liberté personnelles, à la vie privée, à une audience équitable et publique, à l'équité procédurale et à la présomption d'innocence (voir la figure 11 qui donne un aperçu de l'impact des outils d'évaluation des risques de la justice pénale sur les droits humains ; pour des exemples spécifiques, se référer au module 4 de ce manuel de formation).[88] Pour illustrer ces points, certains outils d'évaluation des risques s'appuient sur les données des appels à la police, qui peuvent s'avérer un indicateur peu fiable des tendances réelles de la criminalité (par rapport aux dossiers d'arrestation). Ces données sont souvent déformées par des préjugés raciaux, comme dans le cas tristement célèbre d'Amy Cooper, qui a appelé la police au sujet d'un ornithologue noir qui lui avait simplement demandé de maintenir son chien en laisse, à Central Park.[89] Il est essentiel de comprendre que ce n'est pas parce qu'un appel est passé pour signaler un crime que cela signifie nécessairement qu'un crime a réellement eu lieu. Cependant, de tels appels peuvent être utilisés comme points de données dans les systèmes d'évaluation des risques, pour justifier l'envoi de policiers dans un quartier particulier ou même le ciblage d'un individu spécifique, créant ainsi une boucle de rétroaction où les technologies basées sur les données légitiment la discrimination policière.[90]

Dans l'affaire Ewert c. Canada, la Cour suprême du Canada a souligné que les outils d'évaluation des risques créés et vérifiés à l'aide de données provenant de groupes majoritaires peuvent ne pas être précis pour prédire les mêmes caractéristiques dans les groupes minoritaires.[91]

85 Committee of Experts on Internet Intermediaries (MSI-NET) (2018). Algorithms and Human Rights: Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications, Étude du Conseil de l'Europe, DGI/2017/12, disponible sur :https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5

86 Hao K., Stray J. (2019). Pouvez-vous rendre l'IA plus équitable qu'un juge ? Play our courtroom algorithm game, disponible sur : https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm/.

87 Wykstra S. (2018). Bail reform, which could save millions of unconvicted people from jail, explained, disponible sur : https://www.vox.com/future-perfect/2018/10/17/17955306/bail-reform-criminal-justice-inequality

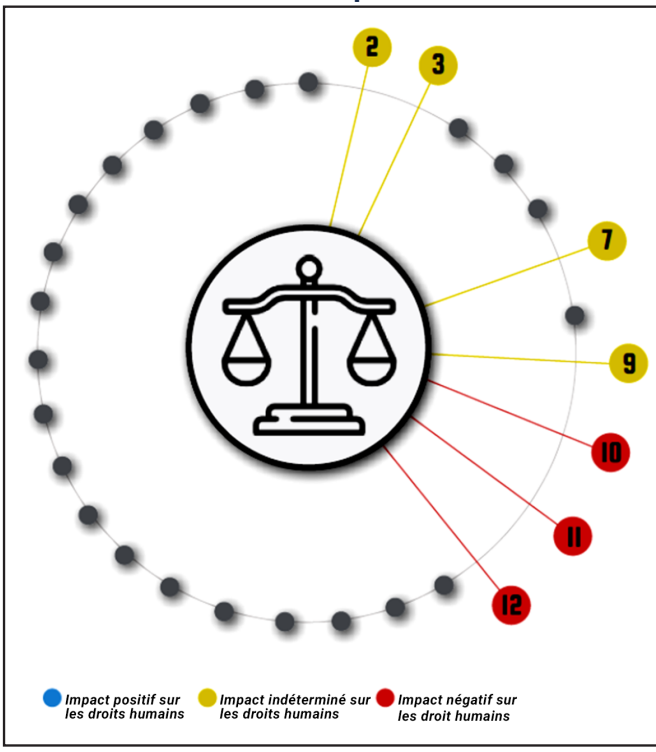
88 Committee of Experts on Internet Intermediaries (MSI-NET) (2018). Algorithms and Human Rights: Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications, Étude du Conseil de l'Europe, DGI/2017/12, disponible sur :https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5

89 Nir S. M. (2020). How 2 Lives Collided in Central Park, Rattling the Nation, disponible sur : https://www.nytimes.com/2020/06/14/nyregion/central-park-amy-cooper-christian-racism.html

90 Heaven W. D. (2020). Predictive policing algorithms are racist. They need to be dismantled, disponible sur : https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/

91 Disponible sur : https://www.scc-csc.ca/case-dossier/cb/37233-eng.pdf

Figure 11. Outils d'évaluation des risques en matière de justice pénale Impact sur les droits humains



2. Absence de discrimination et 7. Droit à l'égalité devant la loi. Ces systèmes peuvent reproduire et perpétuer des biais dans les données de formation, mais d'autres preuves suggèrent qu'ils peuvent néanmoins réduire les disparités raciales (entre autres) en matière de libération sous caution et de détermination de la peine.

3. Droit à la vie, à la liberté et à la sécurité personnelle. Les individus à faible risque peuvent bénéficier d'une plus grande liberté provisoire et de peines plus courtes, et la communauté peut bénéficier d'un taux de criminalité plus faible.

9. Absence d'arrestation, de détention et d'exil arbitraires. Des algorithmes complexes peuvent classer à tort certaines personnes comme « à haut risque », ce qui rend possible une détention arbitraire avant le procès ou après la condamnation.

10. Droit à un procès public équitable et 11. Droit à la présomption d'innocence. La nature exclusive de ces outils, leur complexité inhérente et le caractère impénétrable des résultats produits rendent difficile toute contestation devant les tribunaux.

12. Droit à la vie privée. Les systèmes automatisés d'évaluation des risques reposent sur la collecte, le stockage et l'analyse de grandes quantités de données personnelles, ce qui soulève d'importantes préoccupations en matière de confidentialité.

Source : Raso F., Hilligoss H., Krishnamurthy V., Bavitz C., Kim L. (2018). Artificial Intelligence & Human Rights: Opportunities & Risks, disponible sur : https://cyber.harvard.edu/publication/2018/artificial-intelligence-human-rights

Résolution de litige

Les systèmes d'IA peuvent être utilisés pour prévoir comment une affaire sera tranchée, offrant ainsi aux plaignants une meilleure compréhension de leurs options, ou générant une proposition de règlement. Dans cette approche, la prédiction des décisions judiciaires pourrait faciliter l'accès à la justice. De tels systèmes peuvent être intégrés à des plateformes judiciaires en ligne, où les individus explorent leurs alternatives juridiques, ou saisissent et échangent des informations relatives aux affaires. Le système d'IA aiderait les demandeurs à prendre de meilleures décisions de dépôt et aiderait les tribunaux à accélérer la prise de décision en complétant ou en remplaçant les conclusions des juges.[92]

De nombreuses plateformes de règlement des litiges en ligne (RLL) n'utilisent pas l'IA, mais servent plutôt de plateforme pour la coordination et la simplification du travail des demandeurs. Cependant, les plateformes de RLL telles que Rechtwijzer, aux Pays-Bas[93], MyLaw BC, au Canada[94], et celle utilisée par le Tribunal de résolution civile de la Colombie-Britannique (CRT), au Canada[95], utilisent des systèmes d'IA pour déterminer quelles parties peuvent utiliser la plateforme pour résoudre un différend, ainsi que pour automatiser la prise de décision et le règlement ou la recommandation de résultat.

Par exemple, au CRT de la Colombie-Britannique, la procédure de règlement des litiges commence par le recourt à Solution Explorer, un système expert d'IA, qui utilise une structure de questions-réponses pour fournir aux utilisateurs

92 Wu J. (2019). AI Goes to Court: The Growing Landscape of AI for Access to Justice, disponible sur : https://medium.com/legal-design-and-innovation/ai-goes-to-court-the-growing-landscape-of-ai-for-access-to-justice-3f58aca4306f
93 Voir : https://rechtwijzer.nl/
94 Voir : https://family.legalaid.bc.ca/retiring-mylawbc
95 Voir : https://civilresolutionbc.ca/

des informations juridiques linguistiques et des ressources d'auto-assistance gratuites, afin de résoudre leur problème sans avoir besoin de soumettre une réclamation au CRT. Des avocats de toute la Colombie-Britannique ont contribué à la production de contenu juridique pour Solution Explorer. Les ingénieurs des connaissances ont rendu visite à des avocats et les ont interrogés sur les problèmes les plus fréquents observés dans leurs domaines de pratique ainsi que sur les faits juridiques qu'ils estiment que le public devrait connaître. L'équipe du CRT a ensuite organisé ces données en cartes mentales détaillées, en veillant à ce que le langage et le contenu soient clairs et destinés aux élèves de 6e.[96]

Activité:

L'exemple de Solution Explorer, du CRT de Colombie-Britannique

Les participants à la formation regardent la vidéo ci-dessous et se demandent si des solutions similaires utilisant des systèmes experts en IA existent dans leurs juridictions.

Source : https://youtu.be/ueVUETHy8gc

Étude de cas : Bot du jury

Chaque année, la Cour supérieure du comté de Los Angeles traite environ 1,2 million d'infractions au code de la route. Il y a plusieurs années, en raison d'une crise financière de l'État qui a entraîné la fermeture des palais de justice et la réduction du personnel, les gens devaient attendre jusqu'à 2,5 heures pour voir un greffier concernant leur problème de circulation.[97] Aujourd'hui, un assistant en ligne pour la Cour supérieure de Los Angeles aide les gens avec leurs contraventions. Le bot du jury utilise des services de traduction ML et la compréhension du langage naturel. Il assiste plus de 5 000 citoyens chaque semaine et parle cinq langues.

Source : The Superior Court of California, County of Los Angeles, disponible sur : https://ww2.lacourt.org/traffic/ui/trafficOS.aspx?s=1&language=2

96 Salter S. (2018). What is the Solution Explorer?, disponible sur : https://www.cbabc.org/BarTalk/Articles/2018/April/Features/What-is-the-Solution-Explorer

97 SRLN (2023). News: Gina - LA›s Online Traffic Avatar Radically Changes Customer Experience (Los Angeles 2016), disponible sur : https://www.srln.org/node/1186/gina-las-online-traffic-avatar-radically-changes-customer-experience-news-2016

En Australie, l'État de Victoria pilote des plateformes de RLL dans le cadre de son projet pilote VCAT, pour les petites créances.[98] Ces projets pilotes utilisent des plateformes telles que Modria, Modron et Matterhorn by Court Innovations. On ne sait pas dans quelle mesure l'IA est incluse dans ces systèmes, mais il semble qu'il s'agisse principalement de plateformes d'enregistrement des faits et des préférences, d'interaction entre les parties et de rédaction/signature d'accords (sans qu'aucun algorithme ou outil d'IA ne décide ou n'élabore de stratégie pour les parties). Si les projets pilotes aboutissent et se transforment en initiatives au long cours, les futures itérations pourront inclure des recommandations ou des aides à la décision supplémentaires alimentées par l'IA.[99]

IA générative

Le domaine de l'IA générative connaît actuellement une ère de progrès sans précédent. Ces algorithmes d'apprentissage automatique ont été conçus pour créer du nouveau contenu, notamment de l'audio, du code, des images, du texte, des simulations et des vidéos. Récemment, des chatbots tels que ChatGPT, Bard et Copilot ont été développés. Ils utilisent des modèles linguistiques de grande taille (LLM) pour remplir diverses fonctions, telles que la collecte de recherches, la compilation de dossiers juridiques, l'automatisation des tâches de bureau répétitives et la recherche en ligne. Cette technologie innovante a le potentiel d'augmenter considérablement l'efficacité et la productivité en simplifiant des processus et des décisions spécifiques, tels que la rationalisation du traitement des notes ou l'aide à l'enseignement de la pensée critique pour les formateurs.[100]

Point de discussion : Les participants à la formation regardent la vidéo et discutent de la façon dont l'IA générative a influencé leur vie. Ont-ils essayé de l'utiliser dans les processus de prise de décision ? Quelles sont les principales opportunités et défis liés à l'IA générative ?

Source : https://www.youtube.com/watch?v=hfIUstzHs9A

98 Legaltech News (2020). A Future ODR Roadmap for Courts Post-COVID-19, disponible sur : https://www.law.com/legaltechnews/2020/06/23/a-future-odr-roadmap-for-courts-post-covid-19/
99 Legaltech News (2020). A Future ODR Roadmap for Courts Post-COVID-19.
100 Routley N. (2023). What is generative AI? An AI explains, disponible sur : https://www.weforum.org/agenda/2023/02/generative-ai-explain-algorithms-work/.

Les systèmes d'IA peuvent générer du texte, y compris des arguments juridiques ou des recherches, en prédisant le texte approprié selon une entrée donnée, à l'aide de modèles tirés de vastes ensembles de données. Cela fait de l'IA générative un puissant outil dans plusieurs domaines, notamment le secteur juridique. Alors que certains outils d'IA générative fonctionnent dans un univers fermé d'informations, d'autres sont ouverts et ont un accès plus large aux données, par exemple via des plugins Web ou des connexions Internet.[101]

De nombreux gouvernements du monde entier ont commencé à réduire l'utilisation de modèles linguistiques de grande taille (LLM).[102] Le projet de loi de l'UE sur l'IA contient également des règles pour l'IA à usage général, ou les systèmes d'IA qui peuvent être déployés pour une variété de tâches avec différents niveaux de risque. Parmi ces technologies, on retrouve ChatGPT et d'autres systèmes d'IA générative LLM. Dans un autre exemple, en raison de problèmes de protection des données et de la vie privée, le régulateur italien de la protection des données a temporairement interdit ChatGPT.[103]

Les LLM telles que ChatGPT collectent des quantités massives de données sur Internet, y compris des informations personnelles. Le gouvernement canadien a adopté une approche proactive pour réglementer l'utilisation de l'IA générative, en publiant un projet de code de pratique, désormais ouvert aux commentaires du public. Le code sera promulgué dans le cadre de la loi du pays sur l'intelligence artificielle et les données.[104]

Pendant ce temps, le G7 a lancé le processus d'IA d'Hiroshima, pour coordonner les discussions sur les risques liés à l'IA générative.[105] En juillet 2023, le président américain Joe Biden a annoncé que les grandes entreprises d'IA s'engageaient volontairement à donner la priorité à la sûreté, à la sécurité et à la confiance.[106] Le 13 juillet 2023, la Chine a mis en œuvre des mesures temporaires pour réglementer l'industrie de l'IA générative. Les nouvelles règles exigent que les fournisseurs de services subissent des évaluations de sécurité et soumettent des algorithmes pour examen.[107] En outre, l'Autorité municipale de la santé de Beijing a proposé 41 nouvelles règles qui interdisent strictement l'utilisation de l'IA dans diverses activités de soins de santé en ligne, y compris la génération automatique d'ordonnances médicales.[108]

101 Perkins Coie (2023). Use of Generative AI in Litigation Requires Care and Oversight, disponible sur : https://www.perkinscoie.com/en/news-insights/use-of-generative-ai-in-litigation-requires-care-and-oversight.html.

102 Définition de LLM par Tech Target : « Un modèle de grand langage (LLM) est un type d'algorithme d'intelligence artificielle (IA) qui utilise des techniques d'apprentissage en profondeur et des ensembles de données massivement volumineux pour comprendre, résumer, générer et prédire de nouveaux contenus. Le terme IA générative est également étroitement lié aux LLM, qui sont en fait un type d'IA générative spécialement conçu pour aider à générer du contenu textuel»,voir:https://www.techtarget.com/whatis/definition/large-language-model-LLM

103 McCallum S. (2023). ChatGPT banned in Italy over privacy concerns, disponible sur : https://www.bbc.com/news/technology-65139406

104 Canadian Guardrails for Generative AI – Code of Practice (2023), disponible sur : https://ised-isde.canada.ca/site/ised/en/consultation-development-canadian-code-practice-generative-artificial-intelligence-systems/canadian-guardrails-generative-ai-code-practice

105 La Maison blanche (2023). G7 Hiroshima Leaders' Communiqué, disponible sur : https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/20/g7-hiroshima-leaders-communique/

106 Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI, disponible sur : https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/

107 Reuters (2023). China says generative AI rules to apply only to products for the public, disponible sur : https://www.reuters.com/technology/china-issues-temporary-rules-generative-ai-services-2023-07-13/

108 Beijing to limit use of generative AI in online healthcare activities, including medical diagnosis, amid growing interest in ChatGPT-like services, disponible sur : https://www.scmp.com/tech/policy/article/3231828/beijing-limit-use-generative-ai-online-healthcare-activities-including-medical-diagnosis-amid

Par ailleurs, la Commission fédérale du commerce (FTC) des États-Unis a lancé une enquête sur OpenAI, au sujet d'allégations de violations de la loi sur la protection des consommateurs. La demande d'enquête civile de la FTC a soulevé des préoccupations selon lesquelles ChatGPT, un modèle de langage développé par OpenAI, pourrait produire des déclarations fausses ou désobligeantes sur des personnes réelles. L'agence a également demandé des informations à la suite d'une violation de la confidentialité des données, au cours de laquelle des données d'utilisateurs privées ont été exposées dans les résultats de ChatGPT.[109]

L'exemple de ChatGPT

ChatGPT (Generative Pre-trained Transformer) est un chatbot qui tire parti du traitement avancé du langage naturel (TAL) et de l'apprentissage par renforcement pour participer à des discussions réalistes avec les gens. Il peut générer des articles, des contes, de la poésie et même du code informatique. Il peut également répondre à des questions, engager des discussions et, dans certains cas, fournir des réponses détaillées à des interrogations et à des demandes de renseignements extrêmement précises. ChatGPT a été lancé en novembre 2022 et a acquis plus d'un million d'utilisateurs en une semaine.[110]

Le pouvoir judiciaire n'a pas échappé aux controverses liées à l'utilisation de l'IA générative. Ce fut notamment le cas en Colombie, en janvier 2023, après qu'un juge a révélé qu'il avait utilisé ChatGPT pour l'aider à déterminer si l'assurance d'un enfant autiste devait couvrir toutes les dépenses liées à son traitement médical.[111] Dix jours plus tard, toujours en Colombie, un magistrat a rendu une ordonnance judiciaire utilisant ChatGPT pour l'aider à décider comment mener un procès dans le métavers. En outre, fin mars 2023, un juge péruvien et un magistrat mexicain ont affirmé avoir utilisé le ChatGPT d'OpenAI pour motiver une décision de deuxième instance et illustrer leurs arguments, lors d'une audience.[112]

Suite à l'affaire Mata c. Avianca Airlines., Inc[113], où un avocat a soumis des citations falsifiées et des affaires créées par ChatGPT à un tribunal américain, les directives d'utilisation responsable sont devenues encore plus essentielles. Le juge fédéral Brantley Starr (District Nord du Texas) a mis en place une nouvelle règle qui exige une certification plus explicite et précise. Celle-ci garantit que tout texte généré par l'IA générative sera soumis à une vérification humaine de l'exactitude en recoupant des sources juridiques faisant autorité, avant d'être présenté à la Cour.[114] Son ordonnance exigeait ce qui suit :

> *« Tous les avocats et demandeurs comparaissant devant la Cour doivent, avec leur acte de comparution, ajouter au dossier un certificat attestant soit qu'aucune partie de tout dépôt n'est rédigée par une intelligence artificielle générative (telle que ChatGPT, Harvey.AI ou Google Bard), soit que toute langue rédigée par une intelligence artificielle générative est vérifiée pour exactitude, à l'aide de rapports imprimés ou de bases de données juridiques traditionnelles, par un être humain.*

109 Reuters (2023). US FTC opens investigation into OpenAI over misleading statements, disponible sur : https://www.reuters.com/technology/us-ftc-opens-investigation-into-openai-washington-post-2023-07-13/
110 https://chat.openai.com/
111 Gutiérrez J. D. (2023). ChatGPT in Colombian Courts: Why we need to have a conversation about the digital literacy of the Judiciary, disponible sur : https://verfassungsblog.de/colombian-chatgpt/
112 Gutiérrez J. D. (2023). Judges and Magistrates in Peru and Mexico Have ChatGPT Fever, disponible sur : https://techpolicy.press/judges-and-magistrates-in-peru-and-mexico-have-chatgpt-fever/
113 Mata v. Avianca, Inc., 1:22-cv-01461, disponible sur : https://www.courtlistener.com/docket/63107798/mata-v-avianca-inc/
114 Hunton Andrews Kurth (2023). Will Mandatory Generative AI Use Certifications Become The Norm In Legal Filings?, disponible sur : https://www.huntonak.com/en/insights/will-mandatory-generative-ai-use-cerfifications-become-the-norm-in-legal-filings.html. Voir aussi : https://law.mit.edu/ai

Ces plateformes sont incroyablement puissantes et comportent de nombreuses utilisations en droit : divorces de forme, demandes de découverte, suggestions d'erreurs dans les documents, questions anticipées lors des plaidoiries. Mais le document d'information juridique n'en fait pas partie. Voici pourquoi. Ces plateformes, dans leur état actuel, sont sujettes aux hallucinations et aux préjugés. Sur les hallucinations, elles recourent à des citations et références. Un autre problème se pose : la fiabilité ou le biais. Alors que les avocats jurent de mettre de côté leurs préjugés, biais et croyances personnels pour respecter fidèlement la loi et représenter leurs clients, l'intelligence artificielle générative est le produit d'une programmation conçue par des humains qui n'ont pas eu à prêter serment. En tant que tels, ces systèmes n'ont aucune loyauté envers leurs clients, l'état de droit ou les lois et la Constitution des États-Unis (ou, comme indiqué ci-dessus, la vérité). Libérés de tout sens du devoir, de l'honneur ou de la justice, ces programmes agissent selon le code informatique plutôt que la conviction, et reposent sur la programmation plutôt que sur les principes. Toute partie estimant qu'une plateforme dispose de l'exactitude et de la fiabilité requises pour les informations juridiques peut demander un report et expliquer pourquoi. En conséquence, la Cour rejettera tout dépôt d'une partie qui omet d'adjoindre au dossier un certificat attestant qu'elle a lu les exigences spécifiques du juge de la Cour et comprend qu'elle sera tenue responsable, en vertu de la règle 11, du contenu de tout dépôt qu'elle signe et soumet à la Cour, que l'intelligence artificielle générative ait rédigé ou non une partie de ce dépôt ».

Source : United States District Court, Northern District of Texas, Mandatory Certification Regarding Generative Artificial Intelligence

Voici trois principaux risques de l'IA générative en ce qui concerne les systèmes judiciaires :

- **Objectif/champ d'application de la dérive**. Un système d'IA conçu et déployé dans le but « A » ne doit pas être utilisé aveuglément pour une fonction alternative. Par exemple, un outil de TAL principalement destiné à la traduction des ordonnances judiciaires ne doit pas être utilisé arbitrairement pour aider les requêtes de cas ou aider les juges dans la prise de décision, sans divulguer son utilisation à ces fins supplémentaires. Dans certains cas, les objectifs supplémentaires peuvent être valides, dans d'autres non. Même lorsque des fonctions supplémentaires peuvent être jugées légales et valides, il peut être nécessaire de former l'algorithme de base sur des données pertinentes supplémentaires pour assurer l'exactitude et la fiabilité. Fondamentalement, une expansion aveugle de la dérive de l'objectif exacerbe généralement les risques potentiels d'un système d'IA à usage général et doit être dissuadée ou au moins réglementée.
- **Hallucinations et désinformation/anti-information.** Il est important de garder à l'esprit que les modèles d'IA générative sont formés sur de grandes quantités de données, ce qui entraîne des réponses très réalistes et pertinentes. Cependant, il convient de noter que les outils utilisant de tels modèles peuvent produire des

résultats plausibles, mais pas entièrement précis en raison de la nature de leur conception, qui vise à générer des résultats ressemblants mais potentiellement différents des informations sources. L'IA à usage général, en particulier les LLM, démontre de plus en plus le potentiel d '« halluciner », c'est-à-dire de donner des résultats inexacts d'une manière convaincante à la manière humaine, ce qui les rend crédibles et augmente le risque qu'ils soient considérés comme exacts (une forme de biais d'automatisation). Ceci est particulièrement dangereux dans le système judiciaire. Au cours des derniers mois, nous avons connu différents cas où des juges se sont appuyés sur ChatGPT pour donner des informations sur la jurisprudence existante concernant les questions juridiques. Cela a été signalé en Colombie, dans une affaire d'assurance, et même en Inde (juge de la Haute Cour du Pendjab et de l'Haryana). La production hallucinée peut s'avérer extrêmement problématique, en particulier pour l'arbitrage.

- **Traitement intellectuel de l'information.** Les LLM doivent à nouveau être pris en compte, au regard des préoccupations concernant les droits de propriété intellectuelle traditionnels des créateurs d'œuvres originales.

Activité :

Les participants à la formation lisent le texte ci-dessous sur les implications du droit d'auteur quant à l'utilisation de l'IA générative, et se demandent si les doctrines de « l'utilisation équitable » ou des « exceptions au droit d'auteur autorisées » pourraient être appliquées dans le contexte de l'IA générative ?

Avec la montée de l'IA générative, les poursuites semblent devenir un phénomène quotidien. En novembre 2022, Microsoft, GitHub et OpenAI ont fait face à un recours collectif alléguant que le système Copilot, appartenant à GitHub, qui a été formé sur des milliards de lignes de code public, viole la loi sur le droit d'auteur en régurgitant des extraits de code sous licence sans attribution.[115] En retour, les entreprises ont fait valoir devant un tribunal fédéral de San Francisco que le procès actuel concernant leur utilisation de code open source pour former leurs systèmes d'IA n'était pas soutenable. Les sociétés ont affirmé que la plainte manque de spécificité dans ses allégations. En outre, ils ont fait valoir que le système Copilot de GitHub, qui fournit des suggestions de code aux programmeurs, utilise le code source d'une manière conforme aux principes d'utilisation équitable.[116]

Il existe également une action en justice contre Midjourney et Stability AI, les sociétés responsables d'outils artistiques d'IA largement utilisés. L'affaire prétend que ces entreprises ont violé les droits de millions d'artistes en utilisant des images copiées du Web pour former leurs outils.[117]

115 Vincent J. (2022). The lawsuit that could rewrite the rules of AI copyright, disponible sur : https://www.theverge.com/2022/11/8/23446821/microsoft-openai-github-copilot-class-action-lawsuit-ai-copyright-violation-training-data

116 IT world Canada (2023). Microsoft, GitHub, and OpenAI ask court to dismiss AI copyright lawsuit, disponible sur : https://www.itworldcanada.com/post/microsoft-github-and-openai-ask-court-to-dismiss-ai-copyright-lawsuit

117 Vincent J. (2023). AI art tools Stable Diffusion and Midjourney targeted with copyright lawsuit, disponible sur : https://www.theverge.com/2023/1/16/23557098/generative-ai-art-copyright-legal-lawsuit-stable-diffusion-midjourney-deviantart

En outre, Getty Images a intenté une action en justice contre Stability AI pour avoir prétendument utilisé des millions d'images de leur site sans autorisation afin de former Stable Diffusion, une IA capable de générer de l'art.[118]

La principale préoccupation de l'IA générative est sa tendance à reproduire des images, du texte et d'autres types de contenu, y compris ceux protégés par le droit d'auteur, à partir de ses données de formation. Ce problème a été mis en évidence dans un incident récent où un outil d'IA utilisé par CNET pour écrire des articles explicatifs s'est avéré avoir plagié des articles écrits par des humains, qui faisaient probablement partie de son ensemble de données de formation.[119] De plus, une étude universitaire de décembre a révélé que les modèles d'IA capables de générer des images, tels que DALL-E 2 et Stable Diffusion, peuvent reproduire certains éléments d'images à partir de leurs données de formation.[120]

Certaines plateformes qui hébergent des images ont interdit l'utilisation de contenu généré par l'IA, en raison de répercussions juridiques potentielles. Les professionnels du droit ont également averti sur le fait que l'utilisation d'outils d'IA générative peut exposer les entreprises à des risques, si elles intègrent par inadvertance du contenu protégé par le droit d'auteur produit par ces outils, dans leurs produits en vente.

Des entreprises telles que Stability AI et OpenAI, les créateurs de ChatGPT, ont fait valoir qu'elles étaient protégées par la doctrine du « fair use » (utilisation équitable), même si leurs systèmes étaient formés à l'aide de contenu sous licence. Ce principe juridique, reconnu aux États-Unis, permet une utilisation limitée du matériel protégé par le droit d'auteur sans obtenir l'autorisation du titulaire des droits. Les défenseurs de l'utilisation équitable citent souvent l'exemple d'Authors Guild c. Google, où la Cour d'appel des États-Unis pour le deuxième circuit à New York a déterminé que la numérisation manuelle par Google de millions de livres protégés par le droit d'auteur pour développer sa plateforme de recherche de livres constituait une utilisation équitable, même sans licence. Cependant, le concept d'utilisation équitable est fréquemment débattu et modifié, et il reste largement non testé dans le domaine de l'IA générative.[121]

La question de savoir si les œuvres produites par l'IA peuvent être protégées en invoquant le « fair use » dépend de leur caractère transformateur. Cela signifie que les œuvres doivent utiliser des matériaux protégés par le droit d'auteur d'une manière qui diffère considérablement des originaux. Des affaires juridiques antérieures, telles que la décision Google c. Oracle de la Cour suprême des États-Unis en 2021, indiquent que la création de nouvelles œuvres à partir de données collectées peut être transformatrice. Le tribunal a estimé que l'utilisation par Google de certaines parties du code Java SE pour développer son système d'exploitation Android était considérée comme une utilisation équitable.[122]

Source : Tech Crunch (2023). The current legal cases against generative AI are just the beginning, disponible sur : https://techcrunch.com/2023/01/27/the-current-legal-cases-against-generative-ai-are-just-the-beginning/

118 Brittain B. (2023). Getty Images lawsuit says Stability AI misused photos to train AI, disponible sur : https://www.reuters.com/legal/getty-images-lawsuit-says-stability-ai-misused-photos-train-ai-2023-02-06/

119 Futurism (2023). CNET's AI Journalist Appears to Have Committed Extensive Plagiarism, disponible sur : https://futurism.com/cnet-ai-plagiarism

120 Somepalli G., Singla V., Goldblum M., Geiping J., Goldstein J. (2022). Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models, University of Maryland, disponible sur : https://arxiv.org/pdf/2212.03860.pdf

121 Authors Guild v. Google, Inc., No. 13-4829 (2d Cir. 2015), disponible sur : https://law.justia.com/cases/federal/appellate-courts/ca2/13-4829/13-4829-2015-10-16.html

122 Setty R. (2023). First AI Art Generator Lawsuits Threaten Future of Emerging Tech, disponible sur : https://news.bloomberglaw.com/ip-law/first-ai-art-generator-lawsuits-threaten-future-of-emerging-tech

L'utilisation de techniques d'IA peut réduire le besoin de traduction humaine. Ces outils peuvent identifier rapidement les documents contenant du texte en langue étrangère et fournir une liste des langues qu'ils contiennent, ce qui permet une planification plus approfondie. Plusieurs technologies d'IA peuvent également traduire du texte d'une langue à une autre.

Le traitement automatique du langage naturel (TAL)

Le TAL est une technique de ML qui analyse de grandes quantités de données textuelles ou vocales humaines (transcrites ou acoustiques) pour des propriétés spécifiques, telles que le sens, le contenu, l'intention, l'attitude et le contexte.[123]

L'analyse linguistique est utilisée depuis longtemps dans le domaine juridique et en criminologie. Par exemple, la classification de texte a été utilisée en linguistique médico-légale. Alors que dans le passé, l'analyse était effectuée manuellement, aujourd'hui, les méthodes de ML sont utilisées pour identifier le genre, l'âge, les traits de personnalité et même l'identité d'un auteur, ou pour la transcription en direct.[124] Par exemple, le TAL peut aider les opérateurs judiciaires à identifier et à lier les références à la même personne ou organisation, tout au long d'un ensemble de contrats juridiques. Il peut également être utilisé pour analyser un ensemble d'affaires judiciaires, afin d'identifier des sujets ou des problèmes juridiques récurrents, ou pour extraire les noms des parties impliquées, les dates et les lieux mentionnés dans un avis judiciaire. De plus, les systèmes de TAL peuvent être utilisés pour expurger automatiquement les informations sensibles des documents judiciaires, tels que les numéros de sécurité sociale et les adresses personnelles, afin de protéger la vie privée des individus.

Il convient de noter que les modèles de TAL sont toujours sujets aux erreurs de traduction et que celles-ci peuvent avoir de graves conséquences sur les droits fondamentaux des individus, lorsque ces modèles sont déployés dans les opérations judiciaires.

123 Firth-Butterfield K., Silverman K. (2022). Artificial Intelligence and the Courts: Materials for Judges. Artificial Intelligence – Foundational Issues and Glossary, American Association for the Advancement of Science, disponible sur : https://doi.org/10.1126/aaas.adf0782
124 Medvedeva M., Vols M., Wieling M. (2020). Using machine learning to predict decisions of the European Court of Human Rights, Artif Intell Law, 28, 237–266, disponible sur : https://link.springer.com/article/10.1007/s10506-019-09255-y

Étude de cas

SUVAS, en Inde

Le logiciel Vidhik Anuvaad (SUVAS), un programme d'IA qui traduit les décisions et les ordonnances en neuf langues locales différentes, a été introduit par la Cour suprême en novembre 2019. SUVAS visait à permettre aux personnes qui ne parlent pas anglais d'obtenir plus facilement des jugements et des ordonnances et à les aider à mieux comprendre les procédures judiciaires.

Source : Press Trust of India, Software developed to translate SC judgments in 9 vernacular languages: Law Minister RS Prasad, disponible sur : https://www.business-standard.com/article/pti-stories/software-developed-to-translate-sc-judgments-in-9-vernacular-languages-law-minister-rs-prasad-119121200851_1.html.

En février 2023, Technology Enabled Resolution (TERES), une start-up technologique basée à Bangalore, en Inde, a commencé à utiliser l'IA pour entamer la transcription en direct des audiences de la Cour suprême.

Source : Menthe (2023). Bangalore techies bring AI to Supreme Court for the first time, disponible sur : https://www.livemint.com/news/india/supreme-court-uses-ai-based-transcript-for-the-first-time-here-s-how-it-works-11677403522929.html.

L'inde a réussi à créer ses propres modèles de rôle NER (reconnaissance d'entités nommées) et rhétorique formés sur le texte juridique indien. Le modèle de NER en particulier démontre 91 % de précision.

Source : https://github.com/OpenNyAI/Opennyai

Gestion numérique des dossiers et des affaires

L'IA pourrait également faciliter la gestion des fichiers numériques, ce qui, à son tour, rendrait les opérateurs judiciaires plus efficaces en leur permettant de se concentrer sur des questions plus substantielles.

Intelligent Trial 1.0, une IA de gestion intelligente des tribunaux en Chine

Par exemple, la Haute Cour du Hebei, en Chine, a développé Intelligent Trial 1.0, une IA intelligente de gestion des tribunaux. Elle parcourt et numérise automatiquement les dépôts, classe les documents dans des fichiers électroniques, associe les parties à des affaires existantes, identifie les lois, les affaires et les documents juridiques pertinents à prendre en compte, génère tous les documents de procédure judiciaires nécessaires tels que les avis et les sceaux, et distribue les affaires aux juges afin qu'elles puissent être correctement orientées. La technologie coordonne de nombreuses tâches d'IA dans un flux de travail qui peut minimiser la charge de travail du personnel judiciaire et des juges.

Outil pour l'anonymisation des documents juridiques, Argentine

Pour accélérer le processus judiciaire et réduire la marge d'erreur, Cambá Cooperative, une coopérative de travail de banque de logiciels, a créé un système d'IA évolutif pour anonymiser les documents juridiques en espagnol. Le système d'IA vise à anonymiser les données personnelles des documents publics, à réduire le temps et les erreurs et à protéger le droit à la vie privée. Le tribunal pénal n ° 10 de Buenos Aires, en Argentine, a mis en œuvre cet outil d'IA dans ses décisions.[125]

Approfondissement : l'IA comme preuve dans les procédures judiciaires

La nature complexe des algorithmes de ML et leur opacité posent des défis à l'utilisation des systèmes d'IA comme preuves dans les procédures judiciaires. Les tribunaux doivent établir une méthode fiable pour vérifier l'exactitude des résultats de l'IA, ce qui peut impliquer des témoignages d'experts ou des moyens techniques tels que des filigranes intégrés aux images. Déterminer un expert qualifié pour témoigner sur la précision des applications d'IA est également une question cruciale, avec des options allant des ingénieurs logiciels et des ingénieurs de conception aux ingénieurs des données et aux PDG d'entreprise.[126]

Les juges ont du mal à déterminer la précision des outils de diagnostic alimentés par l'IA. Bien que l'IA de diagnostic médical puisse être comparée aux diagnostics des médecins, la manière dont les algorithmes conçus pour prédire le comportement futur, tels que les outils d'évaluation criminelle, peuvent être évalués scientifiquement ou de manière probante n'est pas claire. Dans le contexte criminel, il peut être difficile de déterminer la causalité avec des algorithmes prédictifs, car ils tiennent également compte des facteurs sociaux qui peuvent influencer le comportement. L'évaluation de l'exactitude, des taux d'erreur et la réalisation de tests et d'examens par les pairs sont des tâches cruciales mais ardues, dans ce domaine. Lorsqu'une personne a été incarcérée ou condamnée, il devient difficile de prédire la part d'influence de son emprisonnement sur son comportement futur. Les effets de l'emprisonnement, y compris le soutien des êtres chers à l'extérieur, peuvent avoir un impact significatif sur le comportement futur d'une personne, ce qui rend extrêmement difficile l'évaluation précise de l'exactitude de la prédiction du ML.

Les parties au litige chercheront également à contester la pertinence et la précision du système de ML, en demandant l'accès à l'algorithme sous-jacent, aux données sur lesquelles il a été formé, validé et testé, ainsi qu'à ce qui se passe et est pondéré à l'intérieur de toute boîte noire d'apprentissage automatique. Ainsi, les tribunaux pourraient être confrontés à des contestations juridictionnelles à plusieurs niveaux, chaque fois que des preuves générées par l'IA sont proposées. Lorsque les résultats d'IA sont admis, les adversaires chercheront à contre-interroger les ingénieurs logiciels responsables de sa conception. En outre, parce que chaque application d'IA est différente, c'est-à-dire qu'elle :

- a des objectifs de résultats différents ;
- s'appuie sur différents algorithmes ;
- utilise différentes méthodologies d'apprentissage automatique ;
- forme, teste et valide à l'aide de différentes données.

Ces questions ne sont généralement pas résolues par l'application de la jurisprudence de la même manière, par exemple, que l'analyse de l'ADN est maintenant généralement acceptée par les tribunaux. Il faut s'attendre à ce que chaque demande soit jugée et ce, dans chaque contexte pour lequel elle est présentée comme preuve.

125 Voir : https://www.empatia.la/en/proyecto/ia2/ ; voir aussi : Selvood I., Uribe P. (2022). Open Justice is Moving Forward in the Americas, disponible sur : https://www.opengovpartnership.org/stories/open-justice-is-moving-forward-in-the-americas/

126 Baker J. E., Hobart L. N., Mittelstead M. G. (2021). AI for Judges. A Framework. Center for Security and Emerging Technology, disponible sur : https://www.armfor.uscourts.gov/ConfHandout/2022ConfHandout/Baker2021DecCenterForSecurityAndEmergingTechnology1.pdf

Les progrès rapides des technologies de l'IA et du TAL offrent de nouvelles possibilités pour moderniser le secteur judiciaire en Afrique. Par exemple, des entreprises comme Juta[127], en Afrique du Sud, tirent parti de ces innovations pour développer des solutions de pointe qui aident les cabinets d'avocats et d'autres organisations juridiques à mener des recherches juridiques approfondies et à découvrir des ressources précieuses pour leurs affaires.[128] En capitalisant sur le vaste répertoire de documents juridiques de Juta et en utilisant des techniques analytiques avancées, les systèmes judiciaires africains peuvent améliorer leur efficacité et leur efficience.

Données de l'affaire

Un domaine potentiel où la technologie de l'IA pourrait être intégrée dans les systèmes judiciaires africains est la numérisation des données des affaires judiciaires. En capturant des informations détaillées sur divers aspects du processus juridique - y compris les jugements, les décisions, les antécédents, les parties impliquées, etc. - cela permettrait aux algorithmes d'apprentissage profond d'identifier les modèles et les idées à partir de ces données. L'organisation et le stockage appropriés des données recueillies sur les cas dans de grandes bases de données aideront à jeter les bases permettant aux Africains de tirer parti de leur valeur, pour une multitude d'applications et de fonctionnalités exploitables. La précision du système d'enregistrement doit également être vérifiable au moyen d'une preuve physique prioritaire sur les enregistrements numériques. Ce n'est pas une tâche facile et cela nécessite une grande coordination dans le système judiciaire. Des analyses aussi simples que le suivi des progrès des différents tribunaux par rapport aux prédécesseurs de chaque juge au cours des années précédentes pourraient déterminer quel juge attribuer à certains types de procès, en fonction de la productivité sur des périodes cibles moyennes, par des études de corrélation à taux de réussite élevé. Une autre sphère serait l'examen des déclarations juridiques au sein d'une affaire issues d'une source libre ou d'ensembles de données générés par le gouvernement.

Gestion de la découverte et de la récupération d'informations

Pour améliorer l'efficacité de la phase de découverte dans les procédures judiciaires et faciliter un partage plus efficace de la documentation pertinente entre les parties prenantes, la mise en œuvre d'archives numériques est cruciale.[129] En mettant en place une plateforme en ligne pour stocker les fichiers et les preuves essentiels, les systèmes judiciaires peuvent tirer parti des mécanismes de recherche de pointe pour localiser rapidement et avec précision les informations sensibles. Une telle approche rationalise non seulement la gestion de l'information, mais permet également aux avocats d'élaborer un argumentaire plus solide et étayé par des faits

127 Juta and Company is a leading provider of quality legal, regulatory, business and academic content across Africa ; voir : https://juta.co.za
128 Jutastat Evolve is a cognitive analytical research solution for fast, accurate discovery, data insights and analytics ; voir https://jutastatevolve.co.za/
129 Kufakwababa C. Z. (2021). Artificial intelligence tools in legal work automation: The use and perception of tools for document discovery and privilege classification processes in Southern African legal firms, Doctoral dissertation, Stellenbosch: Stellenbosch University.

fiables provenant de sources accessibles et interconnectées.

Recours aux procédures judiciaires multimodales

La modernisation de l'environnement des salles d'audience grâce à des rassemblements de médias polyvalents pourrait grandement bénéficier aux opérations judiciaires dans toute l'Afrique. L'intégration d'une gamme d'entrées sensorielles, notamment des enregistrements audio et vidéo, offre plusieurs avantages. Les progrès technologiques dans la vision par ordinateur et l'écoute des machines peuvent considérablement améliorer la façon dont les transcripteurs convertissent les mots parlés en texte, diminuant l'erreur humaine tout en augmentant la vitesse et la précision. Ces transcriptions numériques deviennent des outils d'analyse d'enquête post-procès et, lorsqu'elles sont combinées à des capacités de modélisation prédictive, ouvrent la voie à une aide à la décision plus sophistiquée pendant les audiences actives. En outre, l'indexation des actifs multimédias pour un accès facile simplifie à la fois la référence judiciaire et l'examen public, contribuant à une fiabilité accrue dans le cadre juridique. Les décideurs pourraient envisager de mettre en œuvre des projets pilotes en utilisant des méthodes de tenue de dossiers intelligentes et en observant des résultats prometteurs avant une adoption plus large des formats multimédias ouverts. Ensuite, les changements systémiques peuvent être adaptés aux priorités nationales spécifiques.

Améliorer les outils linguistiques pour le pouvoir judiciaire

L'utilisation de technologies avancées de traitement du langage naturel, telles que la traduction automatique[130] et la classification des documents, offre aux autorités judiciaires africaines une excellente occasion de s'attaquer aux obstacles linguistiques. Le manque de soutien à la langue locale entrave l'engagement du public et la diffusion d'informations vitales concernant les procédures judiciaires. L'adoption de solutions modernes d'IA pour les traductions garantit un accès équitable aux ressources juridiques parmi diverses populations de langues maternelles différentes. Dans le même temps, la classification du contenu en langue locale permet au système judiciaire d'accepter et d'analyser les dépôts multiculturels, réduisant ainsi les divisions géographiques entre les interprètes, les demandeurs et le personnel des tribunaux. Par exemple, des rapports universitaires publiés par des associations professionnelles soulignent que l'élimination de la discrimination linguistique et la promotion de la parité dans le domaine juridique pourraient atténuer des problèmes similaires liés à la jurisprudence autour de l'Afrique.[131] Compte tenu de l'intérêt croissant porté au développement de lexiques régionaux et de techniques inférentielles, davantage de nations peuvent capitaliser sur des présentations constitutives sur mesure. Par la suite, les gouvernements

130 Adelani D., Alabi J., Fan A., Kreutzer J., Shen X., Reid M., Ruiter D., Klakow D., Nabende P., Chang E., Gwadabe T. (2022). A Few Thousand Translations Go a Long Way! Leveraging Pre-trained Models for African News Translation, In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 3053–3070.

131 Docrat Z., (2022). A Review of Linguistic Qualifications and Training for Legal Professionals and Judicial Officers: A Call for Linguistic Equality in South Africa's Legal Profession, International Journal for the Semiotics of Law-Revue internationale de Sémiotique juridique, 35(5), 1711–1731

feraient preuve d'un engagement tangible en faveur d'une intégration constructive des zones reculées.

Dépôts légaux ouverts

Les référentiels ouverts contenant des recueils complets de décisions judiciaires continentales constituent des ressources précieuses pour les praticiens du droit et les universitaires. Des aspects tels que la facilité de navigation amplifient l'importance de ces bases de données dans la promotion de délibérations éclairées. Alors que certains pays africains ont fait des progrès significatifs dans la numérisation de leurs décisions de la haute cour, les tribunaux inférieurs restent comparativement sous-représentés. Bien que disposant d'organisations technologiques juridiques dédiées, une telle distribution déséquilibrée mérite notre attention. Par conséquent, l'amélioration de l'infrastructure des technologies de l'information pour les institutions judiciaires devient nécessaire, afin d'assurer une couverture uniforme de tous les tribunaux, en favorisant une accessibilité équilibrée et des chances égales d'avancement, par le biais des informations juridiques.

Connexion avec l'IA locale

Les institutions universitaires africaines et les centres de recherche privés axés sur l'intelligence artificielle devraient être recherchés par le pouvoir judiciaire pour renforcer les collaborations conjointes qui maximisent les avantages découlant de ces partenariats. L'encouragement de ces interactions aide à naviguer dans des cadres réglementaires internationaux complexes, grâce à une expertise et une expérience partagées. En outre, la participation à des initiatives prolifiques de la communauté de l'IA locale, telles que Deep Learning Indaba, Data Science Africa, Masakhane Research Organisation, Data Science Network, qui comprend de nombreux chercheurs répartis dans divers pays, pourrait considérablement améliorer la connexion des systèmes judiciaires aux esprits innovants de la région. Ainsi, la collaboration à l'échelle du continent représente un potentiel de transformation couvrant les compétences techniques au sein des tribunaux et l'inclusion sociétale générale.

Cette section donne un aperçu général de certains cas de déploiement de l'IA dans le système judiciaire au Brésil, à Singapour, en Argentine, en Colombie, en Inde, au Royaume-Uni et aux États-Unis. Il convient de noter que cela ne constitue pas une approbation de ces cas d'utilisation de l'IA dans certains systèmes judiciaires nationaux, et que les opérateurs judiciaires doivent être conscients de tous les risques (préjugés, boîtes noires, cybersécurité et atteinte aux droits humains) qui pourraient survenir avec l'utilisation de systèmes d'IA dans les opérations judiciaires.

VICTOR, Brésil

La Cour suprême brésilienne (STF) utilise le système VICTOR AI, qui a été développé en collaboration avec l'Université de Brasilia (UnB). La technologie de l'IA analyse l'énorme volume d'appels portés devant la Haute Cour et automatise le processus d'examen en identifiant les cas avec *repercussão geral* (répercussion générale), une exigence pour le traitement d'un appel devant le STF.

Ce n'est qu'en 2018 que plus de cinquante mille appels ont été déposés devant cette Cour, qui a le potentiel de statuer sur environ cent vingt mille affaires par an. La première étape de l'analyse de tous les appels qui parviennent au STF consiste à déterminer s'ils comportent des répercussions générales. Avant VICTOR, cette analyse était effectuée par des fonctionnaires de la cour, sur la base des précédents contraignants des juges, et il fallait compter environ quarante minutes par affaire.

En ce qui concerne sa conception logicielle, VICTOR intègre diverses technologies de pointe et une vaste base de données de documents judiciaires. L'ensemble de données utilisé pour former VICTOR contient plus de 100 000 poursuites et près de trois millions de dossiers, extraits sur une période de deux ans (2017-2019).

Son problème initial était de faire face à la variété de formats des documents judiciaires de tous les tribunaux brésiliens (étatiques, fédéraux, du travail, militaires, de la justice électorale) qui arrivent au STF, tels que des volumes PDF non structurés contenant des documents non indexés.[132]

Système de transcription judiciaire intelligent de Singapour

Le système de transcription judiciaire intelligent (iCTS) a été mis en œuvre dans les tribunaux de Singapour, en partenariat avec A*STAR's Institute for Infocomm Research. L'iCTS a le potentiel d'accroître l'efficacité du tribunal, en transcrivant les audiences du tribunal en temps réel, en supprimant la nécessité d'embaucher un transcripteur humain et en permettant aux juges et aux parties d'examiner les témoignages oraux

132 Salomao L. F., Braga R. (2020). The role of the Judiciary in the realization of the UN 2030 Agenda, disponible sur : https://www.conjur.com.br/2021-jul-09/salomao-braga-judiciario-agenda-2030-onu. Voir aussi : https://portal.fgv.br/en/news/artificial-intelligence-Judiciary-and-its-role-implementing-un-agenda-2030 ; https://sifocc.org/app/uploads/2020/06/Victor-Beauty-or-the-Beast.pdf

devant le tribunal, immédiatement. Il le fait en utilisant des réseaux neuronaux formés avec des modèles de langage et des termes spécifiques au domaine (tels que la terminologie juridique).[133]

Il convient de noter que les systèmes de reconnaissance vocale ont la « réputation » de ne pas bien fonctionner lorsqu'ils sont exposés à certains accents, ce qui finit par être discriminatoire dans certaines circonstances. Les opérateurs judiciaires doivent être conscients de ces lacunes.

Prometea, Argentine

Le système Prometea utilise des approches d'IA pour générer automatiquement des avis judiciaires. En 2017, le bureau du procureur de la ville autonome de Buenos Aires, en Argentine, a commencé à développer Prometea. L'outil a permis au Bureau du Procureur d'améliorer considérablement l'efficacité de ses processus : une réduction de 90 minutes à une minute (- 99 %) pour la résolution d'un processus d'appel d'offres, et de 167 jours à 38 jours (- 77 %) pour la préparation du procès.[134]

Prometea se distingue par trois caractéristiques principales :

- Il offre une interface intuitive et conviviale qui permet la reconnaissance du langage naturel et de « parler » à la machine. Sur un seul écran, l'utilisateur a accès à toutes ses ressources liées au travail.
- Il fonctionne comme un système expert multifonctionnel, avec la capacité d'automatiser le traitement des documents et de fournir une assistance intelligente.
- Il utilise des approches de ML et de regroupement supervisées, basées sur l'étiquetage manuel et la formation sur les ensembles de données générés par la machine.[135]

Les fonctionnalités de Prometea peuvent être divisées en quatre catégories :

- Assistance intelligente : Prometea aide les décideurs et les utilisateurs à atteindre un résultat en utilisant sa voix ou un chatbot. Le système automatise les tâches associées au contrôle des délais des appels judiciaires déposés, analyse les documents pertinents accompagnant le dossier et, grâce à un système basé sur la requête comprenant seulement cinq questions, les juges peuvent élaborer un avis juridique pour statuer sur un appel.
- Automatisation : Le concept d'automatisation a différentes subtilités basées sur de nombreuses circonstances. Il existe principalement deux grands groupes :
    - Automatisation complète : Les algorithmes associent automatiquement les données et les informations aux documents. Le document est généré sans interaction de la part d'une personne.

133 Lee J. (2020). Legal Tech-ing Our Way to Justice, disponible sur : https://lawtech.asia/legal-tech-ing-our-way-to-justice/. Voir aussi : https://www.a-star.edu.sg/docs/librariesprovider10/default-document-library/fw-new-infosheets/smart-nation-digital-economy/intelligent-court-transcription-system.pdf?sfvrsn=72a5a971_3

134 UNESCO Chair on Knowledge Societies and Digital Government (2020). PROMETEA: Transforming the administration of justice with artificial intelligence tools, disponible sur : https://unescochair.cs.uns.edu.ar/en/2020/06/prometea-transforming-the-administration-of-justice-with-artificial-intelligence-tools/. Voir aussi : Corvalan J. G., Le Fevre Cervini E. M. (2020). Prometea experience. Using AI to optimize public institutions, disponible sur : https://ceridap.eu/prometea-experience-using-ai-to-optimize-public-institutions; https://www.ibanet.org/article/14AF564F-080C-4CA2-8DDB-7FA909E5C1F4

135 Corvalan J. G., Le Fevre Cervini E. M. (2020). Prometea experience. Using AI to optimize public institutions, disponible sur : https://ceridap.eu/prometea-experience-using-ai-to-optimize-public-institutions

- Automatisation avec intervention humaine réduite : Dans de nombreux cas, l'interaction humaine avec un système automatisé est nécessaire pour compléter ou améliorer la génération d'un document.

- Classification et détection intelligentes : La détection repose sur la lecture et l'analyse d'un volume massif d'informations, dans lequel Prometea peut identifier des documents en fonction de différentes combinaisons de critères, quelle que soit la diversité linguistique des documents. Ensuite, le système segmente les données en fonction de modèles partagés (mots-clés) tout au long des documents.
- Prédiction : C'est la fonction la plus complexe proposée par Prometea. Une prédiction est faite en fonction des réponses passées. Lorsque Prometea trouve une correspondance entre le présent document et un document précédent, il prend note de la réponse fournie dans des situations précédentes et suggère la même résolution, car les conditions sont similaires. Ce travail découle de la lecture et de la reconnaissance de modèles de décision judiciaire précédents, accessibles sur le Web et provenant d'instances antérieures. Une fois que Prometea a identifié la solution, il permet à l'utilisateur de compléter l'avis juridique en fonction de quelques questions, puis d'afficher un aperçu modifiable en ligne du document final. La première ébauche du document est générée automatiquement par le système d'IA.[136]

Compte tenu des préoccupations persistantes concernant la justification des décisions de Prometea et leurs implications pour une procédure régulière, la société civile a appelé à une surveillance soutenue de l'exécution du programme. D'autres questions à prendre en compte sont le niveau de responsabilité des acteurs concernés (développeurs et juges) et les biais potentiels dans les données et la conception de la formation.[137]

PretorIA, Colombie

Début 2019, la Cour constitutionnelle colombienne a annoncé un projet pilote de mise en œuvre de Prometea pour remédier à l'inefficacité et aux retards. Chaque jour, le tribunal reçoit plus de 2 000 ordonnances de protection de tous les tribunaux du pays. Seuls neuf juges et moins de 200 membres du personnel travaillent pour la Cour constitutionnelle. Cependant, les universitaires et les membres de la société civile ont soulevé de nombreuses préoccupations concernant les effets potentiels de Prometea, ainsi que son fonctionnement et le processus de prise de décision, considérés comme opaques. Le projet pilote Prometea a été mis en pause. Le plus grand défi concernait la confidentialité et la protection des données liées au partage d'informations sensibles avec des tiers, tels que les développeurs de logiciels. Il est crucial que l'identité

136 *Ibid.*
137 OCDE, AI use cases in LAC governments, disponible sur : https://www.oecd-ilibrary.org/sites/08955f48-en/index.html?itemId=/content/component/08955f48-en

des victimes et leurs informations ou données personnelles soient protégées dans les cas d'implication de mineurs ou d'infractions sexuelles, entre autres circonstances. L'accès à ces informations ou données par toute personne autre que le tribunal et les parties impliquées dans le traitement des affaires constituait une violation de la confidentialité. Compte tenu de la faiblesse du système à cet égard, il était particulièrement préoccupant qu'une fuite potentielle d'informations personnelles vers les médias ou d'autres parties intéressées puisse se produire, avec des résultats potentiellement désastreux pour la protection de la vie privée des personnes engagées dans les cas traités par le système d'IA.[138]

À la suite de multiples débats, la Cour constitutionnelle a modifié le projet en mettant en œuvre une technologie plus claire et transparente. C'est pour cette raison que PretorIA, sorti au milieu de l'année 2020, utilise la technologie de modélisation thématique plutôt que les réseaux neuronaux. La nouvelle version peut être complètement expliquée, interprétée et suivie.[139]

SUPACE, Inde

Le pouvoir judiciaire indien a un grand nombre d'affaires en instance. Selon les données de la Grille nationale de données judiciaires, environ 38 millions d'affaires sont en suspens dans divers tribunaux de district et de taluka en Inde, et plus de cent mille affaires sont en instance depuis plus de trois décennies.[140]

La Cour suprême de l'Inde a mis en place un système d'IA, le Portail de la Cour suprême pour l'assistance en matière d'efficacité des tribunaux (SUPACE), qui aidera à l'administration et à la prestation de la justice, en cataloguant un grand nombre de décisions judiciaires antérieures pour un meilleur traitement du matériel de l'affaire, qu'il s'agisse de comprendre la matrice factuelle d'instances spécifiques ou de mener des recherches dynamiques sur les précédents. SUPACE ne sera pas utilisé dans la prise de décision. Le rôle de l'IA se limitera à la collecte et à l'analyse de données.[141]

L'outil d'IA SUPACE est déployé à titre expérimental avec des juges traitant des affaires pénales devant les tribunaux de grande instance de Bombay et de Delhi.

La Cour suprême de l'Inde explore l'utilisation d'une application mobile qui traduira les décisions de la cour en neuf langues. En outre, l'Inde utilise l'IA pour résoudre des accusations mineures telles que des infractions routières.[142]

138 Guitierrez O. L. C., Castañeda J. D., Saavedra Rionda V. P. (2019). Enthusiasm and complexity: Learning from the "Prometea" pilot in Colombia's judicial system, disponible sur : https://giswatch.org/node/6166
139 *Ibid.*
140 Shanthi S. (2021). Behind SUPACE: The AI Portal Of The Supreme Court of India, disponible sur : https://analyticsindiamag.com/behind-supace-the-ai-portal-of-the-supreme-court-of-india/
141 *Ibid.*
142 *Ibid.*

HART (Harm Assessment Risk Tool), Royaume-Uni

L'outil d'évaluation des risques (« HART ») est utilisé par la Durham Constabulary, au Royaume-Uni. En utilisant plus de trente caractéristiques qui décrivent les antécédents criminels et socio-économiques d'une personne, HART utilise un algorithme de ML pour déterminer la probabilité de récidive d'un suspect. La police locale utilise les évaluations des risques effectuées par HART pour décider d'inculper une personne ou de l'orienter vers un programme de réadaptation. HART ne décide pas si une personne est coupable ou innocente, mais son évaluation peut déclencher une série d'actions qui conduisent à ce qu'une personne soit privée de sa liberté ou reconnue coupable d'un crime. Les accusations devraient sans aucun doute être déterminées par le bien-fondé de chaque cas individuel, et il est difficile de voir comment les jugements sur la participation aux programmes de réadaptation pourraient être décidés autrement qu'en analysant soigneusement la situation unique de chaque personne. Il devrait toujours y avoir un humain dans la boucle, qui supervise le résultat d'un système de prise de décision automatisé prenant des décisions à fort impact et sensibles aux faits.[143]

HART est enclin à surcriminaliser, car il est intentionnellement destiné à sous-estimer les personnes admissibles au programme de réadaptation. Cette méthode va à l'encontre de l'idée selon laquelle toute ambiguïté dans une affaire pénale devrait profiter au défendeur (*in dubio pro reo*). Contrairement à ce que fait HART, une approche de la prise de décision en matière de justice pénale fondée sur le respect des droits humains devrait favoriser le défendeur.[144]

143 Oswald M., Grace J., Urwin S., Barnes G. C. (2018). Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental' proportionality, Information & Communications Technology Law, 27 (2), 223–250, disponible sur : https://doi.org/10.1080/13600834.2018.1458455

144 *Ibid.*

3. Activités

Les activités de groupe suivantes visent à encourager les participants à la formation à discuter des diverses implications liées à l'utilisation de l'IA dans le système judiciaire.

Activité 1

Veuillez discuter des questions suivantes avec les autres participants à la formation :

- Qui devrait être responsable des décisions automatisées et comment la responsabilité devrait-elle être répartie au sein de la chaîne d'acteurs, lorsque l'IA aide à la décision finale ?
- Qu'est-ce qu'un procès équitable, si l'ADM aide aux décisions ?
- L'accusé est-il privé du droit à une procédure régulière, lorsque des systèmes d'IA sont déployés à un certain stade de la procédure pénale ?

Activité 2

Veuillez cliquer sur le lien suivant : https://bja.ojp.gov/program/psrac/basics/what-is-risk-assessment#illustration. L'illustration démontre comment les notes de risque sont calculées dans l'évaluation des risques. À titre d'explication, cet exemple hypothétique ne couvre que cinq domaines de prédicteurs, notamment la démographie, les antécédents criminels, l'éducation/l'emploi, le soutien familial/social et la cognition antisociale, et un seul indicateur pour chaque domaine. Chaque indicateur s'est vu attribuer une note allant de 0 à 2 ; plus la note est élevée, plus on est susceptible de récidiver (par exemple, parce que les jeunes sont plus susceptibles de récidiver que les personnes âgées, les valeurs de l'indicateur « âge au prononcé de la peine » diminuent à mesure que l'âge augmente).[145]

Saisissez certaines caractéristiques pour mieux comprendre le fonctionnement de l'outil d'évaluation des risques. Discutez avec les autres participants à la formation de ses avantages et de ses inconvénients.

Activité 3

Les participants à la formation lisent le scénario hypothétique : « Naviguer dans les risques : les juges utilisent l'IA générative » et discutent des principaux défis liés au déploiement de l'IA générative par les tribunaux.

Description du scénario :

Dans un avenir où l'IA générative a fait des progrès significatifs, les juges ont commencé à expérimenter son utilisation dans la salle d'audience. Cependant, ils ont rapidement rencontré plusieurs défis et risques associés à son adoption. Ce scénario met en évidence les risques et les pièges potentiels de l'utilisation de l'IA générative dans un contexte judiciaire.

145 https://bja.ojp.gov/program/psrac/basics/what-is-risk-assessment

Éléments du scénario :

1. **Génération automatisée de documents juridiques :**
   - Les juges commencent à utiliser l'IA générative pour automatiser la rédaction de documents juridiques, tels que les jugements et les opinions.
   - Le système d'IA, bien qu'efficace, génère parfois des arguments et des conclusions juridiques biaisés ou inexacts.

2. **Dépendance excessive à l'aide de l'IA :**
   - Les juges s'appuient de plus en plus sur l'analyse juridique générée par l'IA, réduisant progressivement leur propre pensée critique et leurs compétences décisionnelles.
   - On craint de plus en plus que les juges ne deviennent des utilisateurs passifs de l'IA, diminuant ainsi leur rôle dans l'interprétation et l'application de la loi.

3. **Préjugés éthiques et juridiques :**
   - Les modèles d'IA utilisés par les juges héritent des biais présents dans leurs données de formation. Cela conduit à des décisions qui favorisent de manière disproportionnée certains groupes ou perpétuent les préjugés existants dans le système juridique.
   - Les juristes et les militants soulèvent des préoccupations en matière d'équité et de discrimination.

4. **Transparence et responsabilité :**
   - Les modèles d'IA générative peuvent être complexes et difficiles à interpréter. Les juges ont du mal à expliquer les décisions générées par l'IA aux demandeurs, aux avocats et au public.
   - Des questions se posent sur la responsabilité des décisions générées par l'IA, en particulier dans les cas où elles entraînent des conséquences négatives.

5. **Protection des données et sécurité :**
   - L'utilisation de l'IA générative dans les procédures judiciaires implique le traitement de grandes quantités de données juridiques sensibles. Des préoccupations émergent concernant les violations de données et la sécurité des informations confidentielles.
   - Les tribunaux doivent investir massivement dans la cybersécurité pour se protéger contre les menaces potentielles.

6. **Confiance et perception du public :**
   - À mesure de la hausse du recourt à l'IA générative dans le processus juridique, la confiance du public dans le système judiciaire s'érode.
   - Les citoyens et les demandeurs expriment leur scepticisme quant à l'équité et à l'impartialité des décisions assistées par l'IA.

7. **Défis et précédents juridiques :**
    - Des contestations juridiques se posent quant à l'admissibilité des preuves générées par l'IA et à la question de savoir si l'IA peut être considérée comme une source fiable d'analyse juridique.
    - Les tribunaux sont confrontés à la tâche d'établir des précédents juridiques pour régir l'utilisation de l'IA dans leurs décisions.

Résultat du scénario :

Alors que les juges sont confrontés aux risques et aux défis associés à l'utilisation de l'IA générative dans la salle d'audience, ils doivent soigneusement équilibrer les avantages potentiels de l'efficacité et de la précision avec la nécessité de préserver la transparence, l'équité et le jugement humain dans le système juridique. Le scénario souligne l'importance de lignes directrices complètes, de mécanismes de surveillance et d'une formation continue pour atténuer ces risques et veiller à ce que l'IA améliore, plutôt que de saper, les principes de justice.

4. Ressources

1. AAAS, Artificial Intelligence and the Courts: Materials for Judges, disponible sur : https://www.aaas.org/ai2/projects/law/judicialpapers
2. Abu Elyounes D. (2019). Contextual Fairness: A Legal and Policy Analysis of Algorithmic Fairness, Journal of Law, Technology and Policy, disponible sur : https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3478296
3. Ada Lovelace Institute, AI Now Institute and Open Government Partnership (2021). Algorithmic Accountability for the Public Sector, disponible sur : https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/
4. Bhuiyan J. (2021). LAPD ended predictive policing programs amid public outcry. A new effort shares many of their flaws, disponible sur : https://www.theguardian.com/us-news/2021/nov/07/lapd-predictive-policing-surveillance-reform
5. Brittain B. (2023). Getty Images lawsuit says Stability AI misused photos to train AI, disponible sur : https://www.reuters.com/legal/getty-images-lawsuit-says-stability-ai-misused-photos-train-ai-2023-02-06/ Conseil de l'Europe, disponible sur : https://www.coe.int/en/web/cepej
6. Council of Europe (2021). CEPEJ Action plan 2022 – 2025: "Digitalisation for a better justice", disponible sur : https://rm.coe.int/cepej-2021-12-en-cepej-action-plan-2022-2025-digitalisation-justice/1680a4cf2c
7. Futurism (2023). CNET's AI Journalist Appears to Have Committed Extensive Plagiarism, disponible sur : https://futurism.com/cnet-ai-plagiarism
8. Hind M. (2019). Explaining Explainable AI by Michael Hind, The ACM Magazine for Students, 25(3), disponible sur : https://doi.org/10.1145/3313096
9. Jauhar A., Misra M., Sengupta A., Chakrabarti P. P., Ghosh S., Ghosh K., (2021). Responsible Artificial Intelligence for the Indian Justice System, disponible sur : https://vidhilegalpolicy.in/wp-content/uploads/2021/04/Responsible-AI-in-the-Indian-Justice-System-A-Strategy-Paper.pdf
10. IT world Canada (2023). Microsoft, GitHub, and OpenAI ask court to dismiss AI copyright lawsuit, disponible sur : https://www.itworldcanada.com/post/microsoft-github-and-openai-ask-court-to-dismiss-ai-copyright-lawsuit
11. Somepalli G., Singla V., Goldblum M., Geiping J., Goldstein J. (2022). Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models, University of Maryland, disponible sur : https://arxiv.org/pdf/2212.03860.pdf
12. The Surveillance and Policing of Looted Land (2021). Automating banishment, disponible sur : https://automatingbanishment.org/section/2-architecture-of-data-driven-policing/
13. UNESCO MOOC on AI and the Rule of Law, disponible sur : https://www.unesco.org/en/articles/unesco-global-mooc-ai-and-rule-law-engaged-thousands-judicial-operators
14. UNESCO (2021). Global Toolkit for Judicial Actors International legal standards on freedom of expression, access to information and safety of journalists, disponible sur : https://unesdoc.unesco.org/ark:/48223/pf0000378755

15. Vincent J. (2022). The lawsuit that could rewrite the rules of AI copyright, disponible sur : https://www.theverge.com/2022/11/8/23446821/microsoft-openai-github-copilot-class-action-lawsuit-ai-copyright-violation-training-data

16. Vincent J. (2023). AI art tools Stable Diffusion and Midjourney targeted with copyright lawsuit, disponible sur : https://www.theverge.com/2023/1/16/23557098/generative-ai-art-copyright-legal-lawsuit-stable-diffusion-midjourney-deviantart

17. Završnik A. (2019). Algorithmic justice: Algorithms and big data in criminal justice settings, European Journal of Criminology, 18, 623–642, disponible sur : https://doi.org/10.1177/1477370819876762

Module 3

Défis juridiques et éthiques du déploiement de l'IA

Le module 3 traite des risques juridiques et éthiques associés aux systèmes d'IA, et des défis de la transparence algorithmique et de la responsabilité dans le système judiciaire. Il se poursuit avec un aperçu des questions juridiques les plus importantes liées à l'identification biométrique et à la technologie de reconnaissance faciale. Ce module développe également les principaux défis liés à l'IA et à l'éthique, sur la base de la Recommandation de l'UNESCO 2021 sur l'éthique de l'intelligence artificielle.

Qu'allez-vous apprendre ?

Après avoir terminé ce module, les participants seront en mesure de :

- Comprendre et expliquer les principaux défis liés à la transparence algorithmique et à la responsabilité dans le système judiciaire, ainsi que la jurisprudence pertinente.
- Comprendre les questions juridiques les plus importantes liées à l'identification biométrique, à la technologie de reconnaissance faciale et aux deepfakes.
- Disposer d'une solide compréhension des principaux défis liés à l'IA et à l'éthique, sur la base de la Recommandation de l'UNESCO sur l'éthique de l'intelligence artificielle (2021).

1. Qu'est-ce que l'éthique de l'IA ?

La Recommandation de l'UNESCO sur l'éthique de l'IA aborde ce sujet comme une réflexion normative systématique, basée sur un cadre holistique, complet, multiculturel et évolutif de valeurs, de principes et d'actions interdépendants qui peuvent guider les sociétés dans la gestion responsable des impacts connus et inconnus des technologies de l'IA sur les êtres humains, les sociétés, l'environnement et les écosystèmes, et leur offre une base pour accepter ou rejeter les technologies de l'IA.

L'UNESCO considère l'éthique comme une base dynamique pour l'évaluation normative et l'orientation des technologies de l'IA, se référant à la dignité humaine, au bien-être et à la prévention des dommages comme une boussole, et enracinée dans l'éthique de la science et de la technologie.

Dans la pratique, l'IA éthique implique de considérer les implications éthiques des systèmes d'IA et de s'assurer que leur conception et leur mise en œuvre s'alignent sur des valeurs et des normes sociétales plus larges.

Expérience de pensée

Essayons une expérience de pensée : Vous êtes à l'arrêt de tramway et remarquez soudainement un wagon qui accélère vers cinq personnes inconscientes de son approche. Vous voyez également une deuxième voie occupée par une seule personne. Que feriez-vous ? Choisiriez-vous de détourner le wagon vers la deuxième voie pour sauver les cinq personnes au prix d'une vie ?

Depuis de nombreuses années, le dilemme du tramway est un raisonnement éthique reconnu abordé en cours de philosophie. Cependant, l'émergence de voitures autonomes expérimentales a amené ce problème théorique à la réalité. En conséquence, nous sommes maintenant confrontés au défi de déterminer la programmation appropriée pour les systèmes d'IA, dans des situations critiques de vie ou de mort.

Source : Utrecht University, Unboxing the black box of AI, disponible sur : https://www.uu.nl/en/organisation/in-depth/unboxing-the-black-box-of-ai

De nombreuses initiatives d'autorégulation se sont penchées sur les risques éthiques de l'IA. Les gouvernements, les organisations internationales, le secteur privé, les organisations de la société civile ont tous produit des règles et des principes éthiques non contraignants pour guider le développement et l'utilisation de l'IA. Ce chapitre donne un aperçu des principaux cadres éthiques de l'IA, en se concentrant sur la Recommandation de l'UNESCO sur l'éthique de l'intelligence artificielle (2021). Il convient de noter que la Recommandation de l'UNESCO ainsi que d'autres cadres éthiques sur l'IA n'ont pas les effets contraignants du droit.

Le tableau 4 ci-dessous donne un aperçu des principes clés de l'éthique de l'IA.

Tableau 4. Principes clés de l'éthique de l'IA

Principes	Explication
Équité et parti pris	Les systèmes d'IA doivent être conçus pour assurer l'équité et éviter les préjugés qui peuvent conduire à des résultats discriminatoires. Il est essentiel de s'attaquer aux biais dans les données de formation, les algorithmes et les processus de prise de décision pour prévenir le traitement injuste ou la marginalisation de certains individus ou groupes.
Transparence et explicabilité	Les systèmes d'IA doivent être transparents et permettre aux utilisateurs de comprendre comment ils fonctionnent et comment les décisions sont prises. L'explicabilité est importante pour assurer la responsabilité, permettre l'audit et renforcer la confiance dans les technologies d'IA.
Confidentialité et protection des données	Les systèmes d'IA s'appuient souvent sur de grandes quantités de données, y compris des informations personnelles et sensibles. Le respect des droits à la vie privée et des réglementations en matière de protection des données est essentiel au développement et au déploiement de l'IA. La minimisation de la collecte de données, la garantie d'un consentement éclairé et la protection des données contre tout accès non autorisé sont des considérations clés. Le respect, la protection et la promotion de la vie privée sont essentiels à la sauvegarde de la dignité humaine, de l'autonomie et du libre arbitre, tout au long du cycle de vie des systèmes d'IA.[146]
Fiabilité et responsabilité	Des lignes de responsabilité claires doivent être établies pour les résultats des systèmes d'IA, notamment l'identification des responsables des actions et des décisions prises par les technologies d'IA. Il est crucial de s'assurer qu'il existe des mécanismes pour remédier aux impacts négatifs potentiels des systèmes d'IA.
Sécurité et robustesse	Les systèmes d'IA doivent être conçus dans une logique de sécurité, pour prévenir les dommages involontaires. Des mesures doivent être prises pour s'assurer que les technologies d'IA sont robustes, fiables et capables de gérer des circonstances imprévues et des attaques contradictoires.
Autonomie et supervision humaines	L'IA doit être développée et utilisée pour améliorer l'autonomie et la prise de décision humaines, plutôt que de remplacer ou d'influencer indûment le jugement humain. Le maintien de la surveillance et du contrôle humains sur les systèmes d'IA est important pour préserver l'action humaine. Il est crucial de s'assurer que la responsabilité éthique et juridique peut être attribuée à des personnes physiques ou à des entités juridiques existantes, à chaque étape du cycle de vie du système d'IA. Cela inclut les cas où des recours sont nécessaires. La supervision humaine signifie plus qu'une simple supervision individuelle ; elle implique également un suivi public, inclusif au besoin.[147]

146 UNESCO (2021). Recommendation on the Ethics of Artificial Intelligence, disponible sur : https://unesdoc.unesco.org/ark:/48223/pf0000380455

147 *Ibid.*

Principes	Explication
Impacts sociaux, environnementaux et économiques	Les technologies d'IA peuvent avoir de profonds impacts sociaux et économiques. Les considérations éthiques comprennent la garantie d'un accès équitable aux avantages de l'IA, la minimisation des supressions d'emplois et la prise en compte d'implications sociétales plus larges, telles que l'inégalité des richesses et la fracture numérique.
Inclusion et diversité	Il est essentiel de donner la priorité au respect, à la protection et à la promotion de la diversité et de l'inclusion, lors du développement de systèmes d'IA, conformément au droit international et aux droits humains. Cela peut être réalisé en encourageant la participation active de tous les individus et groupes, indépendamment de leur race, couleur, ascendance, genre, âge, langue, religion, opinions politiques, origine nationale ou ethnique, origine sociale ou économique, handicap ou de tout autre facteur.[148]
Collaboration et approches multidisciplinaires	Aborder l'éthique de l'IA nécessite une collaboration entre diverses parties prenantes, notamment les chercheurs, les décideurs, les experts de l'industrie, les éthiciens et la société civile. Des perspectives multidisciplinaires et des voix diverses sont essentielles pour relever les défis éthiques complexes de l'IA.

Qui est un « acteur de l'IA » ?

Selon la Recommandation de l'UNESCO sur l'éthique de l'intelligence artificielle (2021), tout acteur impliqué dans au moins une étape du cycle de vie du système d'IA est appelé « acteur de l'IA ». Cela inclut les personnes physiques et morales, notamment les chercheurs, les programmeurs, les ingénieurs, les scientifiques des données, les utilisateurs finaux, les entreprises commerciales, les établissements universitaires et les entités publiques et privées.

Quels types de préoccupations éthiques les systèmes d'IA soulèvent-ils ?

Les systèmes d'IA soulèvent de nouvelles préoccupations éthiques, telles que celles liées à la prise de décision, à l'emploi et au travail, à l'interaction sociale, aux soins de santé, à l'éducation, aux médias, à l'accès à l'information, à la fracture numérique, aux données personnelles et à la protection des consommateurs, à l'égalité des sexes, à l'environnement, à la démocratie, à l'état de droit, à la sécurité et à la police, au double usage, ainsi qu'aux droits humains et aux libertés fondamentales, tels que le droit à la vie privée[149], à la liberté d'expression et à l'égalité devant la loi.

De plus, le potentiel des algorithmes d'IA pour reproduire et renforcer les préjugés préexistants et intensifier les formes existantes de discrimination, de préjugés et de stéréotypes présente des défis éthiques importants. À long terme, les systèmes d'IA peuvent saper la valeur ajoutée précédemment assurée par le sens unique de la faculté d'agir et de l'expérience humaine, apportant de nouvelles questions concernant la conscience de soi humaine, les interactions sociales, culturelles et environnementales, ainsi que l'autonomie, la faculté d'agir, la valeur et la dignité.[150]

148 *Ibid.*

149 A noteworthy document in the area of privacy and data protection issues for the Judiciary is the UNESCO (2022) Guidelines for Judicial Actors on Privacy and Data Protecti on, disponible sur : https://unesdoc.unesco.org/ark:/48223/pf0000381298

150 UNESCO (2021). Recommendation on the Ethics of Artificial Intelligence, disponible sur : https://unesdoc.unesco.org/ark:/48223/pf0000380455

Activité :

L'IA prend-elle de meilleures décisions que les humains ? Penser l'éthique de l'IA

Les participants à la formation regardent la vidéo et discutent de la façon dont l'IA et l'éthique interagissent, et de l'impact de l'IA sur l'éthique et les droits humains.

Source : UNESCO, https://youtu.be/2E7l1hdjHsg

Cadres pour l'éthique de l'IA

En plus de la Recommandation de l'UNESCO sur l'éthique de l'intelligence artificielle (2021), d'autres cadres sont brièvement présentés ci-dessous :

- L'Initiative mondiale sur l'éthique des systèmes autonomes et intelligents de l'Institute of Electrical and Electronics Engineers (IEEE) : La IEEE Standards Association a developpé une série de documents, notamment le Ethically Aligned Design framework[151] et le P7000 series of standards[152]. Ces ressources fournissent une approche globale de l'éthique de l'IA, couvrant des domaines tels que la transparence, la responsabilité et la priorisation des valeurs humaines.[153]
- Lignes directrices de la Commission européenne en matière d'éthique pour une IA digne de confiance : La Commission européenne a publié des lignes directrices qui définissent sept exigences clés pour une IA digne de confiance : la faculté d'agir et la surveillance humaines, la robustesse et la sécurité techniques, la confidentialité et la gouvernance des données, la transparence, la diversité, la non-discrimination et le bien-être sociétal et environnemental.[154]

151 IEEE (2019). Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (A/IS), disponible sur :https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf
152 Voir : https://sagroups.ieee.org/7000/
153 Voir : https://standards.ieee.org/industry-connections/ec/autonomous-systems/
154 Commission européenne (2019). Ethics guidelines for trustworthy AI, disponible sur : https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

- Le 3 novembre 2017, la Déclaration de Montréal pour une IA responsable a été annoncée à l'issue du Forum sur le développement socialement responsable de l'IA, qui s'est tenu à Montréal. Cet exemple d'effort de co-création a développé un ensemble de principes directeurs pour le développement et le déploiement responsables de l'IA à des fins publiques. Il s'agit d'un effort de collaboration impliquant une série de consultations publiques et d'assemblées de citoyens composée de 500 résidents, experts et principales parties prenantes. Avec plus de 2 200 citoyens et plus de 200 organisations signataires, il défend les principes suivants : bien-être, vie privée et intimité, respect de l'autonomie, responsabilité, participation démocratique, équité, solidarité, diversité et inclusion, prudence et développement durable.[155]
- Les principes de l'IA d'Asilomar : Ces principes ont été développés par un groupe de chercheurs, de décideurs et de penseurs en IA, lors de la Conférence d'Asilomar sur l'IA bénéfique. Ils couvrent divers aspects éthiques, notamment la garantie des avantages généraux de l'IA, la sécurité à long terme, le leadership en matière de recherche technique et l'orientation coopérative.[156]
- Les principes de l'OCDE en matière d'IA donnent la priorité au développement d'une IA digne de confiance, avec une approche centrée sur l'humain. Élaborés avec la contribution d'un panel de plus de 50 experts représentant les gouvernements, les universités, les entreprises, la société civile, les organisations internationales, la communauté technologique et les syndicats, ils se découpent en cinq principes centrés sur les valeurs pour une mise en œuvre responsable et digne de confiance de l'IA, et cinq recommandations pour les politiques publiques et la collaboration mondiale. Leur objectif est de fournir des orientations aux gouvernements, aux organisations et aux individus dans le développement et l'exploitation de systèmes d'IA qui donnent la priorité au bien-être des personnes, et de veiller à ce que les responsables de leur fonctionnement soient juridiquement reconnus comme tels.[157]

Le tableau 5, au début du module 4, donne un aperçu des principales initiatives en matière de réglementation, de politique et d'éthique de l'IA.

Comment concrétiser l'éthique de l'IA ?

Toute initiative d'IA dans le système judiciaire doit respecter les normes éthiques de responsabilité et d'ouverture. L'IEEE recommande de créer de nouvelles normes qui spécifient des degrés de transparence quantifiables et testables, afin que les systèmes puissent être évalués de manière impartiale et que le degré de conformité puisse être établi pour maintenir la transparence.

Pourtant, en raison des processus étroitement liés et en couches de la programmation algorithmique, le maintien de la transparence des algorithmes devient de plus en

155 Voir : https://gouai.cidob.org/resources/montreal-declaration-for-a-responsible-development-of-artificial-intelligence/
156 Future of Life Institute (2017). AI Principles, disponible sur : https://futureoflife.org/open-letter/ai-principles/
157 OCDE (2019). Forty-two countries adopt new OECD Principles on Artificial Intelligence, disponible sur : https://www.oecd.org/science/forty-two-countries-adopt-new-oecd-principles-on-artificial-intelligence.htm

plus difficile.[158] Les principes d'éthique des données codifiées ou les codes de conduite, les évaluations d'impact éthique et les évaluations d'impact sur la vie privée, la formation éthique pour les opérateurs judiciaires et les commissions d'examen éthique sont quelques exemples de méthodes d'examen éthique qui peuvent permettre une plus grande transparence et responsabilité dans l'utilisation des systèmes d'IA et d'ADM, au sein du système judiciaire.

En général, les évaluations de l'impact sur la vie privée permettent aux organisations et aux développeurs d'évaluer efficacement les risques posés (assurer le respect des exigences de confidentialité, identifier les mesures d'atténuation et classer efficacement les impacts de l'utilisation des données et des algorithmes). Il serait également idéal d'adopter une approche inclusive des parties prenantes qui met l'accent sur « l'inclusion proactive des utilisateurs ». De plus, le contexte d'utilisation des données doit constamment être pris en compte, nécessitant une intervention humaine et parfois une expertise spécifique au contexte.[159]

2. Qu'est-ce que le biais d'IA ?

Le biais d'IA est une différence systématique dans le traitement de certains objets, personnes ou groupes (par exemple, stéréotypes, préjugés ou favoritisme) par rapport à d'autres, par les algorithmes d'IA. Le biais d'IA peut avoir un impact sur la collecte et l'interprétation des données, la conception du système et la façon dont les utilisateurs interagissent avec lui.[160]

Les systèmes d'IA sont loin d'être des technologies neutres. Au contraire, ils peuvent refléter les préférences, les priorités et les préjugés (inconscients) de leurs créateurs. Les biais peuvent survenir de nombreuses manières, dans les systèmes d'IA. Les données de formation et les modèles d'IA peuvent être biaisés. Les groupes privilégiés peuvent être avantagés par rapport aux autres groupes, dans les décisions d'IA.

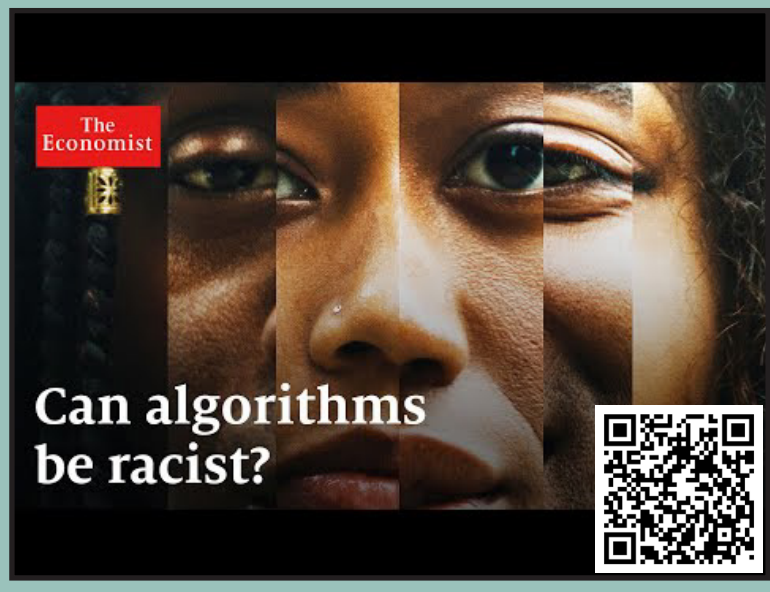
Même lorsque les développeurs de logiciels prennent grand soin de minimiser toute influence de leur propre biais, les données utilisées pour former un algorithme peuvent être une autre source importante de biais. Les systèmes d'IA peuvent renforcer ce qu'ils ont appris des données et augmenter les risques tels que les préjugés raciaux et sexistes.[161]

En outre, même un algorithme soigneusement construit fonde ses jugements sur des informations provenant d'une réalité imprévisible et imparfaite. Dans des situations nouvelles, les programmes d'IA sont susceptibles de commettre des erreurs de jugement.[162]

158 Voir : https://www.ieee.org
159 Morley J., Floridi L., Kinsey L., Elhalal A. (2019). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices, Eng Ethics, 26, 2141–2168, disponible sur : https://doi.org/10.1007/s11948-019-00165-5
160 Goole (2023). Machine Learning Glossary, disponible sur : https://developers.google.com/machine-learning/glossary/
161 Turner Lee N., Resnick P., Barton G (2019). Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms, disponible sur : https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/
162 Judge Dixon H. B. (2021). Artificial Intelligence: Benefits and Unknown Risks, disponible sur : https://www.americanbar.org/groups/judicial/publications/judges_journal/2021/winter/artificial-intelligence-benefits-and-unknown-risks/

Sujet de discussion

Les participants à la formation regardent la vidéo et discutent de la façon dont les préjugés liés à l'IA les ont affectés et pourquoi il est important d'en être conscient dans les contextes judiciaires



Source : The Economist, https://youtu.be/IzvgEs1wPFQ

Expérience de pensée : Biais liés aux données dans l'identification des chats et des chiens

Imaginez que vous créez un programme d'IA pour reconnaître les animaux de compagnie. Si l'algorithme est formé sur un million d'images de chiens, mais seulement quelques milliers d'images de chats, il peut avoir du mal à identifier avec précision les chats, en raison d'une compréhension moins développée de leur apparence. Il convient de noter que l'IA peut présenter des biais, car elle repose sur des données et des choix de formation qui peuvent être influencés par des préjugés humains.

Source : Utrecht University, Unboxing the black box of AI, disponible sur : https://www.uu.nl/en/organisation/in-depth/unboxing-the-black-box-of-ai

Certains des biais les plus controversés en IA se produisent dans la technologie de reconnaissance faciale. Une étude menée en 2016 à Oakland, en Californie, a révélé que, malgré les données d'enquête montrant une répartition égale de la consommation de drogues entre les groupes raciaux, les prédictions algorithmiques des arrestations policières se concentraient principalement sur les communautés afro-américaines, créant des boucles de rétroaction qui renforçaient les modèles de biais systémiques dans l'histoire des arrestations policières.[163] Les algorithmes de reconnaissance faciale peuvent également introduire des biais raciaux, lorsqu'ils

163 Banque mondiale WDR 2021. L'étude de 2016 menée par le Groupe d'analyse des données sur les droits humains à l'aide de données de 2010 et 2011 du service de police d'Oakland et d'autres sources a comparé une cartographie de la consommation de drogues basée sur des données d'enquête auprès des victimes d'actes criminels, avec une autre basée sur une analyse algorithmique des arrestations par la police. L'étude a montré que des données sources biaisées pourraient renforcer et potentiellement amplifier les préjugés raciaux dans les pratiques d'application de la loi. Les données sur les arrestations ont montré que les quartiers afro-américains connaissent en moyenne 200 fois plus d'arrestations liées à la drogue que les autres quartiers d'Oakland.

sont formés principalement sur des données provenant de visages caucasiens, ce qui réduit considérablement leur précision dans la reconnaissance d'autres ethnies.[164] Il est préoccupant de constater que diverses technologies ne fonctionnent pas avec précision pour les personnes à la peau plus foncée.

Par exemple, une étude menée par Georgia Tech a révélé que les voitures sans conducteur sont plus susceptibles de heurter les personnes de couleur, car les systèmes de détection d'objets qu'elles utilisent pour identifier les piétons ne fonctionnent pas aussi efficacement sur les personnes à la peau plus foncée. Ces exemples soulignent la nécessité d'une technologie plus inclusive et impartiale qui s'adresse à tous, quelle que soit la couleur de la peau.[165] L'industrie de la technologie est confrontée depuis longtemps à un problème de diversité de sa main-d'œuvre. Le Rapport sur la santé d'Internet 2020 de Mozilla suggère que près de 80 % des employés des grands géants de la technologie, comme Apple, Facebook, Google et Microsoft, sont des hommes. En outre, depuis 2014, on constate une croissance minime de la représentation des communautés noires, latines et autochtones ce qui constitue une préoccupation alarmante qui doit être abordée.[166]

164 Hill K. (2020). Wrongfully Accused by an Algorithm." New York Times, disponible sur : https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html

165 Kenny C. (2021). Artificial Intelligence: Can We Trust Machines to Make Fair Decisions? Data and Computer Scientists, Ecologists, Pathologists, and Legal Scholars Study AI's Biases, disponible sur: https://www.ucdavis.edu/curiosity/news/ais-race-and-gender-problem

166 Mozilla (2020). Internet Health Report, disponible sur : https://foundation.mozilla.org/en/insights/internet-health-report/

Approfondissement : Exemples de biais d'IA

Microsoft Tay a été créé pour plaire aux personnes âgées de 18 à 24 ans, et elle a fait ses débuts sur les réseaux sociaux avec un joyeux « Hello, world ! » (le « o » de *world* - monde, en français - était un émoji de la planète Terre). En douze heures, cependant, Tay s'est transformée en une négationniste raciste et grossière de l'Holocauste qui a déclaré que toutes les féministes « devraient mourir et brûler en enfer ». Tay, qui a été rapidement supprimé de Twitter, a été conçu pour apprendre des actions des autres utilisateurs de Twitter, et à cet égard, c'est une réussite. L'acceptation par Tay des pires caractéristiques de l'humanité est un exemple de biais algorithmique, qui se produit lorsque du code apparemment inoffensif adopte les biais de ses concepteurs ou des données qui l'alimente.

En 2015, Google Photos a identifié à tort plusieurs utilisateurs afro-américains comme des gorilles, déclenchant l'indignation des médias sociaux. L'architecte social en chef de Google et responsable de l'infrastructure pour Google Assistant a rapidement annoncé sur Twitter qu'une équipe était en cours de constitution pour résoudre le problème.

Source : Wired (2017). How to Keep Your AI From Turning Into a Racist Monster, disponible sur : https://www.wired.com/2017/02/keep-ai-turning-racist-monster/; voir également : https://www.bbc.com/news/.technology-33347866

Les outils d'IA peuvent être biaisés envers les personnes de couleur et les minorités

Une recherche menée en 2019 par le National Institute on Science and Technology (NIST) des États-Unis sur les technologies de reconnaissance faciale, souvent « basées sur l'IA », a révélé que les algorithmes étaient jusqu'à 100 fois plus susceptibles de produire un faux positif pour les personnes de couleur. Par exemple, le NIST a découvert que « pour une correspondance un-à-plusieurs, l'équipe a vu des taux plus élevés de faux positifs concernant les femmes afro-américaines », une conclusion « particulièrement importante parce que les répercussions pourraient inclure des allégations erronées ». Selon une deuxième étude menée par l'Université de Stanford et le MIT, le taux d'erreur pour les femmes à la peau foncée était de 34,7 %, contre 0,8 % pour les hommes à la peau claire. Une évaluation de Rekognition, un système de reconnaissance faciale appartenant à Amazon et vendu aux forces de l'ordre, a révélé des indicateurs de préjugés raciaux et a constaté que le système avait incorrectement reconnu 28 membres du Congrès américain comme des délinquants condamnés. De même, les systèmes d'IA et de prise de décision automatisée utilisés dans les dispositions préalables au procès, la détermination de la peine et les contextes carcéraux fournissent souvent des résultats erronés ou biaisés qui perpétuent les disparités existantes.

L'un des aspects les plus difficiles des préjugés liés à l'IA est que les ingénieurs et les développeurs en IA n'ont pas besoin d'être intentionnellement racistes ou sexistes. C'est une condition inquiétante, à une époque où les gens croient de plus en plus que la technologie est plus impartiale qu'eux. À mesure que l'industrie informatique développe l'IA, elle court le risque d'intégrer le racisme et d'autres préjugés dans un code qui fera des choix pendant des décennies. Et parce que l'apprentissage profond implique que c'est ce code, et non l'humain, qui rédigera du code, la nécessité d'éliminer les biais algorithmiques est encore plus grande.

Le biais d'IA peut être causé par diverses raisons, et voici quelques définitions et exemples de biais d'IA majeurs :

- **Biais de l'échantillon en raison de données de formation biaisées et non représentatives :** Si les règles extraites par l'algorithme d'apprentissage automatique d'un ensemble spécifique de données sont considérées comme légitimes, les préjugés et les omissions intégrés dans les exemples de données seront répétés dans le modèle prédictif. En d'autres termes, si les données utilisées pour former le modèle d'IA ne sont pas représentatives du contexte dans lequel le système d'IA est utilisé, le système d'IA peut produire des résultats biaisés. Par exemple, un système de reconnaissance faciale principalement développé à l'aide de photographies d'hommes blancs peut ne pas être en mesure d'identifier avec précision les femmes ou d'autres groupes raciaux. Les recherches montrent que dans le cas des femmes et des personnes de différentes origines raciales et culturelles, les niveaux de précision de ces modèles sont nettement inférieurs. On trouve un autre exemple dans les systèmes d'IA programmés pour identifier le cancer de la peau. Si l'ensemble de données initial n'est pas représentatif de la population, cette méthode fonctionnera mal pour les membres des groupes sous-représentés.[167]
- **Rappeler les biais lors de l'étiquetage des données :** Lorsque la solution d'IA utilise des données étiquetées, le processus d'étiquetage doit être cohérent entre les ensembles de données, sinon le résultat du modèle devient inexact. Par exemple, quelqu'un pourrait décrire une image de téléphone comme endommagé, et une autre comparable comme légèrement endommagé. L'ensemble de données sera incohérent dans cette situation, car il y aura deux étiquettes différentes faisant référence à des images similaires et comparables.
- **Biais d'association :** Il est important de noter que même les ensembles de données représentatifs reflètent des préjugés historiques et sociétaux, par exemple contre les minorités surreprésentées dans les populations carcérales ou les femmes occupant des emplois moins prestigieux. La « représentativité » des données peut donc perpétuer la discrimination et l'inégalité, alors qu'en fait, un ensemble de données consciemment adapté qui corrige ces inégalités sociales pourrait produire des résultats moins discriminatoires, à partir d'algorithmes formés sur cette base et ensuite appliqués à de nouveaux cas (pour informer la détermination de la peine privative de liberté ou bien l'examen automatisé des demandes d'emploi, par exemple). Le biais d'association le plus connu est le biais de genre. Par exemple, lorsque l'ensemble de données utilisé fait référence à un groupe de professions où tous les hommes travaillent en tant que médecins et toutes les femmes en tant qu'infirmières. Cela n'empêche pas les hommes de devenir infirmiers ou les femmes de devenir médecins. Cependant, selon le modèle de ML, il n'y a pas d'infirmiers ni de femmes médecins.

167 Mozilla (2020). Internet Health Report, disponible sur : https://foundation.mozilla.org/en/insights/internet-health-report/

- **Biais de mesure :** Il est causé par une mesure défectueuse par les sujets et/ou le chercheur. La source du biais de mesure est une imprécision faite lors de la collecte ou de la mesure des données. Par exemple, si les photos capturées par un appareil photo utilisé pour fournir des données pour un système de reconnaissance d'image sont de mauvaise qualité, cela peut entraîner des résultats biaisés quant à certaines données démographiques.[168] Une autre illustration est le jugement humain. Par exemple, un système de diagnostic médical peut être formé pour prédire la probabilité de maladie sur la base de mesures indirectes, telles que les visites chez le médecin plutôt que les symptômes réels.[169] Le biais de mesure peut également provenir du moment où les données pour certains groupes de population ne sont pas capturées du tout, en raison de leur existence en dehors du flux de collecte de données. Par exemple, l'utilisation des données des téléphones mobiles comme indicateur indirect de la capacité de l'utilisateur à rembourser les prêts peut désavantager les personnes ayant un accès limité ou inexistant aux téléphones mobiles. On peut imaginer une autre situation, où un algorithme conçu pour trouver des candidats pour des emplois pourrait utiliser le succès passé sur le lieu de travail comme un prédicteur du succès futur sur le nouveau, et extraire de cette information des critères de recrutement privilégiés spécifiques comme l'éducation et l'expérience. Les statistiques sous-jacentes, cependant, peuvent être dépassées, par exemple à une époque où les minorités ou les femmes étaient sous-représentées sur le marché du travail ou dans les normes d'admission à l'école. En conséquence, le système pourrait disqualifier les candidats qui pourraient surpasser l'ensemble de données « bon exécutant » du passé.[170]
- **Biais d'automatisation en raison de la dépendance non critique aux résultats générés par l'IA :** Une menace majeure posée par l'utilisation des systèmes d'IA dans l'administration de la justice est le dénommé biais d'automatisation, qui est la tendance des humains à considérer de manière non critique la solution offerte par l'IA comme correcte. Cela peut conduire à un manque de scepticisme envers les informations fournies par les algorithmes et à une tendance à agir automatiquement selon leurs suggestions. Détecter les biais d'automatisation peut être difficile, car ils sont souvent inconscients. On peut les repérer en prêtant attention à la façon dont nous nous appuyons sur les informations fournies par les systèmes automatisés et en nous demandant si nous critiquons ces informations ou si nous les acceptons sans poser de questions. Il est également important d'être conscient de nos propres préjugés et d'essayer d'être objectif, lors de l'évaluation des informations fournies par les systèmes automatisés. Par conséquent, le fait pour le juge de s'écarter de toute décision assistée ou automatisée ne doit impliquer aucune forme de représailles, de sanction, d'inspection ou de régime disciplinaire. Si la supervision et le contrôle humains prévalent, le contrôle doit être efficace (voir la section sur « Le principe de l'humain dans la boucle », module 1).

168 Hackernoon (2020). 7 Types of Data Bias in Machine Learning, disponible sur : https://hackernoon.com/7-types-of-data-bias-in-machine-learning-ubl3t3w.
169 Data Camp (2022). Different types of AI bias, disponible sur : https://www.datacamp.com/blog/data-demystified-the-different-types-of-ai-bias.
170 Baker J. E., Hobart L. N., Mittelstead M. G. (2021). AI for Judges. A Framework. Center for Security and Emerging Technology, disponible sur : https://www.armfor.uscourts.gov/ConfHandout/2022ConfHandout/Baker2021DecCenterForSecurityAndEmergingTechnology1.pdf

Les participants à l'activité de formation lisent l'histoire ci-dessous et évaluent l'impact éthique de la technologie, conformément à l'instrument d'évaluation de l'impact éthique de l'UNESCO, à l'annexe I (veuillez vous concentrer sur les parties qui traitent de l'équité, de la non-discrimination, de la diversité, de la protection des données et de la vie privée).

En 2020, JK a demandé un permis de conduire international à l'Office national des transports de Hambourg, une ville portuaire du nord de l'Allemagne. Elle a apporté tous les documents nécessaires à son rendez-vous, à l'exception d'une photo biométrique qu'elle souhaitait prendre dans le photomaton de l'office. Pour prendre une photo biométrique, elle devait placer son visage dans une zone spécifique de l'appareil photo, la photo n'étant prise qu'une fois le visage détecté par la machine. JK n'a pas été reconnue par le photomaton de l'Office national des transports, car il semblait que le logiciel de reconnaissance faciale de ce photomaton n'identifiait que les visages aux tons de peau clairs.

JK s'est souvenue qu'un membre du personnel lui avait dit que son teint de peau pourrait poser problème. Le bureau d'impression du gouvernement, propriétaire du photomaton a déclaré que, puisque le photomaton était équipé de la technologie la plus récente, le problème n'était pas lié au logiciel. Il a affirmé que la cause du problème était l'éclairage inadéquat de la cabine. Toutefois, selon une étude de Joy Buolamwini et Timnit Gebru, les technologies d'IA les plus récentes peuvent avoir des faiblesses qui entraînent des résultats discriminatoires et sexistes.

Source : Algorithm Watch. Automated Decision-Making Systems and Discrimination Understanding causes, recognizing cases, supporting those affected A guidebook for anti-discrimination counselling; Buolamwini J., Gebru T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR, 81, 77–91, disponible sur : https://proceedings.mlr.press/v81/buolamwini18a.html

Rappel !

Les outils d'IA intègrent les choix politiques des décideurs précédents, et donc le biais de ces décisions. Les outils de jugement préétablis, par exemple, peuvent introduire des préjugés, réduire le pouvoir discrétionnaire des juges et ne pas répondre aux difficultés spécifiques auxquelles sont confrontées les personnes appartenant à des groupes marginalisés et vulnérables. En conséquence, la compréhension de ces technologies et la poursuite de leur enquête et de leur évaluation permettront aux juges de participer pleinement à l'évolution des opérations judiciaires permise par l'IA.

Un mot de prudence sur les préjugés dans les systèmes d'IA conduisant à la discrimination

Même si un système d'IA semble neutre en surface, ses algorithmes peuvent conduire à des évaluations et des conséquences discriminantes. La discrimination peut souvent provenir de pratiques préjudiciables dans le monde réel qui alimentent les données utilisées par le système d'IA.

Lorsque les technologies de police axées sur les données sont des boîtes noires, il est difficile d'analyser les risques de taux d'erreurs, de faux positifs, de limitations des capacités de programmation, de données biaisées et même de défauts dans le code source qui influencent les résultats de recherche. Ces systèmes de boîtes noires perpétuent des cercles vicieux de préjugés.

Les systèmes de police prédictifs qui s'appuient sur des données historiques courent le risque de reproduire les résultats d'actes discriminatoires antérieurs. Cela peut entraîner des « boucles de rétroaction », dans lesquelles chaque nouveau choix basé sur des données antérieures génère plus de données, ce qui entraîne une suspicion et une incarcération disproportionnées des groupes marginalisés. Les algorithmes prédictifs peuvent contribuer à une prise de décision biaisée et à des conséquences discriminatoires, en fonction de la manière dont les crimes sont documentés, des crimes choisis pour être inclus dans l'étude et des méthodes analytiques utilisées.

Bien que de nombreuses personnes pensent que les données policières sont neutres, elles contiennent des préjugés politiques, sociaux et autres. Les données des services de police reflètent les procédures et les priorités du département, ainsi que les intérêts locaux, étatiques et fédéraux, et les préjugés institutionnels et individuels. Il n'y a pas de procédures définies pour l'utilisation des informations recueillies lors des opérations d'application de la loi dans le développement des systèmes d'IA. En outre, les pratiques policières peuvent jouir d'une transparence et d'une supervision limitées.[171]

De nombreuses études de recherche ont montré à maintes reprises que l'utilisation d'algorithmes prédictifs dans les services de police formés sur les données de criminalité passées réplique et amplifie les biais systémiques existants. Souvent, ce processus tient peu compte de la façon dont les différentes initiatives de réduction de la criminalité, la législation sur la criminalité, les tendances en matière de profilage ou les biais de détermination de la peine influencent les modèles détectés par ces algorithmes dans les données.

L'examen public accru de ces algorithmes a soulevé des questions sur la façon dont ils sont développés et mis en œuvre, la raison pour laquelle ils ne sont pas soumis à un examen plus approfondi, et s'il existe des mécanismes de gouvernance en place pour évaluer correctement leurs risques, leurs vulnérabilités et leur potentiel de préjudice sociétal accru.[172] Il a été démontré que le déploiement d'outils d'IA dans le système de justice pénale peut exacerber des pratiques policières déjà discriminatoires à l'égard des minorités.

171 Leslie D., Burr C., Aitken M., Cowls J., Katell M., Briggs M. (2021). Artificial intelligence, human rights, democracy, and the rule of law: a primer, Le Conseil de l'Europe, disponible sur : https://www.turing.ac.uk/sites/default/files/2021-03/cahai_feasibility_study_primer_final.pdf. « Les gouvernements adoptent de plus en plus de réglementations concernant l'utilisation des données, telles que le Règlement général sur la protection des données (RGPD) de l'Union européenne. Mais ceux-ci ont tendance à être axés sur l'utilisation des données par les entreprises et la mesure dans laquelle les protections accordées en vertu de ces règlements s'étendent aux LEA [agents d'application de la loi] est moins claire », disponible sur : UNESCO (2022). Manuel de formation mondial pour les agents des forces de l'ordre : Liberté d'expression, accès à l'information et sécurité des journalistes, disponible sur : https://unesdoc.unesco.org/ark:/48223/pf0000383978

172 Grupo de trabajo de la NACDL sobre vigilancia predictiva (2021). Garbage in, gospel out. How Data-Driven Policing Technologies Entrench Historic Racism and 'Tech-wash' Bias in the Criminal Legal System, disponible sur : https://www.nacdl.org/getattachment/eb6a04b2-4887-4a46-a708-dbdaade82125/garbage-in-gospel-out-how-data-driven-policing-technologies-entrench-historic-racism-and-tech-wash-bias-in-the-criminal-legal-system-11162021.pdf

La vérité sur les algorithmes. Les participants à la formation regardent la vidéo présentée par Cathy O'Neil et discutent de la manière et des raisons pour lesquelles les algorithmes sont biaisés. Les participants discutent également de la façon dont les biais algorithmiques pourraient avoir un impact sur leur travail.

Source : https://youtu.be/heQzqX35c9A

Après avoir regardé la formation vidéo, les participants discutent également du scénario suivant :

Scénario : biais algorithmique dans l'embauche

Dans un avenir pas si lointain, une grande entreprise, appelons-la « TechCo », décide de mettre en œuvre un système d'embauche algorithmique pour rationaliser son processus de recrutement et le rendre plus efficace. TechCo est fière de son engagement en faveur de la diversité et de l'inclusion, et la direction estime que l'utilisation d'outils d'embauche basés sur l'IA les aidera à atteindre ces objectifs. Ils embauchent une équipe de scientifiques des données et d'ingénieurs en apprentissage automatique pour développer le système.

Voici comment se déroule le scénario :

1. Collecte de données :
   - L'équipe commence par collecter les données historiques des processus d'embauche passés de TechCo. Cet ensemble de données comprend les CV, les commentaires lors des entretiens et les décisions d'embauche de la dernière décennie.
   - Malheureusement, les données historiques reflètent certains biais qui ont existé au sein de l'entreprise. Par exemple, un nombre disproportionné de candidats masculins sont embauchés à des postes techniques, et les candidats de certaines universités prestigieuses sont favorisés.

2. Formation des modèles :
   - Les scientifiques des données utilisent ces données historiques pour former l'algorithme. Ils cherchent à identifier les modèles et les critères qui prédisent les candidats retenus.
   - En raison des données historiques biaisées, l'algorithme commence à détecter ces biais. Par exemple, il pourrait apprendre que les candidats d'universités prestigieuses ont plus de chances de réussir, même si cette préférence est basée sur un parti pris historique plutôt que sur un mérite objectif.
3. Biais involontaire :

   Lorsque l'algorithme commence à traiter de nouvelles demandes d'emploi, il perpétue par inadvertance les biais présents dans les données de formation. Les curriculum vitae des femmes, des candidats issus de milieux sous-représentés et de ceux des écoles moins prestigieuses reçoivent des notes inférieures, ce qui les écarte ou les renvoie au bas de la liste d'embauche.
4. Plaintes et préoccupations éthiques :
   - Au fil du temps, les demandeurs d'emploi qui estiment avoir été injustement rejetés commencent à exprimer leurs préoccupations. Ils remarquent un schéma où l'algorithme désavantage systématiquement certains groupes.
   - Les organisations de défense des droits civils et les médias ont vent de ces problèmes et commencent à enquêter sur les pratiques d'embauche de TechCo, les accusant de partialité algorithmique et de discrimination.
5. Conséquences juridiques et réputationnelles :
   - TechCo fait face à des défis juridiques et à des poursuites potentielles pour pratiques d'embauche discriminatoires. Ils subissent également un impact significatif sur leur réputation, les clients et les partenaires exprimant leur préoccupation quant à leur engagement en faveur de la diversité et de l'inclusion.
   - La direction de l'entreprise se rend compte du problème de biais algorithmique et décide d'arrêter temporairement l'utilisation de l'algorithme d'embauche pendant qu'elle enquête sur le problème.
6. Audit algorithmique et mesures correctives :
   - TechCo fait appel à des auditeurs externes et à des éthiciens des données pour évaluer l'algorithme et son impact. Les auditeurs identifient les données biaisées et les failles du modèle.
   - L'entreprise prend des mesures pour recycler l'algorithme avec un ensemble de données plus diversifié et représentatif, supprimer les fonctionnalités biaisées et mettre en œuvre des mesures de protection contre les biais futurs.
7. Reconstruire la confiance :

   TechCo s'excuse publiquement pour le biais algorithmique et la discrimination. Elle décrit son engagement à remédier au problème et à garantir des pratiques d'embauche équitables.

L'entreprise investit dans des mesures de transparence, publie régulièrement des rapports sur les performances de son algorithme d'embauche et recherche un contrôle externe, afin de regagner la confiance perdue.

Plusieurs candidats à un emploi qui croient avoir été lésés par la partialité inhérente au système d'embauche déposent plainte. Que décideriez-vous et quels facteurs prendriez-vous en compte lors de votre prise de décision ?

Les risques liés aux biais posés par l'IA et l'ADM sont devenus omniprésents, comme dans les systèmes de reconnaissance faciale dans les espaces publics, qui permettent une surveillance de masse[173] ou dans le déploiement de systèmes ADM fortement biaisés pour la détection de la fraude à l'aide sociale, comme le système néerlandais SyRI, abordé dans l'encadré ci-dessous.[174] Les systèmes d'IA peuvent fonctionner de manière imprévisible, et même les systèmes qui semblent effectuer des tâches « simples » ou routinières peuvent entraîner des résultats involontaires et souvent dommageables. Cela rend les risques encore plus élevés, comme décrit dans l'encadré ci-dessous, qui met en évidence des exemples de biais algorithmiques dans le système judiciaire.

Études de cas : Exemples de biais algorithmiques dans le système judiciaire et le gouvernement

COMPAS

Le profil de gestion des délinquants correctionnels pour les sanctions alternatives (COMPAS) utilisé par le pouvoir judiciaire aux États-Unis n'inclut pas la race ou l'origine ethnique comme critères, mais des recherches ont montré qu'il attribue systématiquement des notes de risque plus élevés aux accusés noirs qu'aux accusés blancs, ce qui les rend moins susceptibles d'être libérés.[175] Il y a eu des cas où des prisonniers ayant des antécédents pratiquement parfaits, comme Glen Rodriguez[176], se sont vus refuser une libération conditionnelle en raison d'une note COMPAS inexacte, ce qui leur laissait peu de recours pour contester la décision ou même savoir comment elle avait été calculée. Une analyse réalisée en 2016 par ProPublica a révélé que les COMPAS utilisés par les tribunaux de Floride contenaient des préjugés raciaux. ProPublica a examiné 7 000 cas et a découvert que la note était extraordinairement peu fiable pour prédire les crimes violents : seules 20 % des personnes susceptibles de commettre des crimes violents l'ont fait. Les chercheurs ont également découvert que l'algorithme était plus susceptible de désigner les défendeurs de couleur comme de futurs criminels que les défendeurs blancs, et que les défendeurs blancs étaient plus souvent étiquetés à tort comme présentant un faible risque que les défendeurs de couleur.[177] En réponse à l'étude, le propriétaire de COMPAS, Northpointe, a publié un démenti soutenant que le rapport de ProPublica était « basé sur des statistiques et une analyse de données erronées et ne montrait pas que le COMPAS lui-même était biaisé sur le plan racial, et encore moins que d'autres instruments de risque étaient biaisés ».[178]

173 Big Brother Watch (2019). UK MASS SURVEILLANCE CHALLENGED IN EUROPE'S HIGHEST HUMAN RIGHTS COURT, disponible sur : https://bigbrotherwatch.org.uk/2019/07/uk-mass-surveillance-challenged-in-europes-highest-human-rights-court/

174 Algorithm Watch (2020). How Dutch activists got an invasive fraud detection algorithm banned, disponible sur : https://algorithmwatch.org/en/syri-netherlands-algorithm/

175 Corbett-Davies S., Pierson E., Feller A., Goel S., Huq A. (2017). Algorithmic decision making and the cost of fairness, disponible sur : https://arxiv.org/pdf/1701.08230.pdf

176 Wexler R. (2017). When a computer program keeps you in jail: How computers are harming criminal justice, disponible sur : https://www.nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html

177 Criswell B. (2020). Algorithms Deciding the Future of Legal Decisions, disponible sur : https://montrealethics.ai/algorithms-deciding-the-future-of-legal-decisions/

178 Angwin J., Larson J., Mattu S., Kirchner L. (2016). Machine Bias, disponible sur : https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing; voir aussi : https://www.uscourts.gov/federal-probation-journal/2016/09/false-positives-false-negatives-and-false-analyses-rejoinder.

Le système SyRI

Afin de détecter la fraude à la protection sociale, le gouvernement néerlandais a déployé un système nommé SyRI, pour « indication du risque système », afin de croiser les informations personnelles des résidents provenant de différentes bases de données et d'identifier les « profils de citoyens improbables » qui nécessitent un examen plus approfondi. Le système fonctionnait comme suit : si un organisme gouvernemental (par exemple, les municipalités, la banque de sécurité sociale, les autorités fiscales) détectait une fraude aux prestations, aux allocations ou aux impôts dans un certain quartier, il pouvait utiliser SyRI. SyRI a aidé à identifier les résidents qui devaient faire l'objet d'une enquête pour fraude plus approfondie.

Cette pratique a été contestée par l'Autorité néerlandaise de protection des données et le Conseil d'État, qui ont soulevé des préoccupations concernant le droit à la vie privée ainsi que les droits à une procédure régulière, tels que la présomption d'innocence. En outre, le système manquait de transparence, car ses algorithmes n'ont pas été publiés et il n'a fait l'objet d'aucun audit technique. Par ailleurs, son ciblage des quartiers défavorisés pourrait constituer une discrimination fondée sur le statut socio-économique ou migratoire des résidents. De plus, SyRI a été utilisé principalement dans les quartiers à faibles revenus, ce qui exacerbe la discrimination et les préjugés, si le gouvernement utilise exclusivement l'analyse des risques de SyRI dans ces quartiers.

En 2020, le tribunal de La Haye a ordonné l'arrêt immédiat de SyRI[179], concluant que la législation qui l'instituait offrait une protection insuffisante contre les intrusions dans la vie privée, en raison des mesures disproportionnées adoptées pour prévenir et punir la fraude, dans l'intérêt du bien-être économique. La Cour a conclu que SyRI avait violé l'article 8 de la Convention européenne des droits de l'homme (CEDH), qui protège le droit au respect de la vie privée et familiale.

Source : Algorithm Watch (2020) How Dutch activists got an invasive fraud detection algorithm banned, disponible sur : https://algorithmwatch.org/en/syri-netherlands-algorithm/. Voir aussi : https://towardsdatascience.com/fighting-back-on-algorithmic-opacity-30a0c13f0224; https://iapp.org/news/a/digital-welfare-fraud-detection-and-the-dutch-syri-judgment/; https://pace.coe.int/en/files/28715/html

Réglementation du droit à l'explication dans l'UE, dans le cadre de l'ADM

Des règles telles que le « droit à l'explication » du Règlement général sur la protection des données (RGPD) de l'UE ont été promulguées, en réponse à des problèmes liés à la transparence et à la responsabilité de l'IA.[180] Les articles 13(2)(f), 14(2)(g) et 15(1)(h) du RGPD obligent les responsables du traitement à informer les personnes concernées de l'existence d'ADM, notamment le profilage, visé à l'article 22(1) et (4), et à fournir des informations utiles sur la logique en cause, ainsi que sur l'importance et les conséquences pour la personne concernée.[181]

180 Casey B., Farhangi A., Vogl R. (2018). Rethinking Explainable Machines: The GDPR's 'Right to Explanation' Debate and the Rise of Algorithmic Audits in Enterprise, Berkeley Technology Law Journal, 34, disponible sur : https://ssrn.com/abstract=3143325

181 Une personne concernée est une personne qui peut être identifiée, directement ou indirectement, par un identifiant tel qu'un nom, un numéro d'identification ou des données de localisation, ou par des facteurs personnels liés à son identité physique, physiologique, génétique, mentale, économique, culturelle ou sociale. Voir aussi : https://academic.oup.com/idpl/article/7/4/233/4762325

179 Algorithm Watch (2020). How Dutch activists got an invasive fraud detection algorithm banned, disponible sur : https://algorithmwatch.org/en/syri-netherlands-algorithm/.

L'article 22(1) du RGPD précise que les personnes concernées ont le droit de ne pas faire l'objet d'une décision fondée exclusivement sur un traitement automatisé, notamment le profilage, qui crée des effets juridiques les concernant ou les affecte de manière significative. L'article 22(2) – (4) décrit les conditions limitées dans lesquelles la prise de décision automatisée est autorisée, et décrit certaines protections pour garantir que les personnes concernées peuvent exercer leurs droits avec succès.[182]

Étude de cas : Législation sur le droit à l'explication en Estonie

L'article 23(4) de la loi estonienne sur l'assurance-chômage permet à la Caisse d'assurance-chômage de prendre des décisions concernant l'attribution des prestations de chômage aux demandeurs, de manière entièrement automatisée. Les demandeurs sont immédiatement informés que la décision a été prise automatiquement, qu'ils ont le droit d'être entendus et qu'ils peuvent déposer une demande de révision interne.

De telles pratiques permettent aux personnes dont la demande a été soumise à une prise de décision automatisée de comprendre comment les décisions ont été prises et de faire appel de ces décisions.

Source : https://fpf.org/blog/gdpr-and-the-ai-act-interplay-lessons-from-fpfs-adm-case-law-report%ef%bf%bc/

Biais d'IA et égalité des sexes

Par exemple, les technologies de reconnaissance automatisée du genre (AGR) suppriment le droit à l'auto-identification et infèrent le genre en fonction des données acquises sur les personnes. Les technologies d'AGR utilisent des informations telles que le nom légal et les traits du visage d'une personne, pour simplifier l'identité de genre en binaire. Cela manque d'une compréhension scientifique des diverses identités de genre.[183] Cet effacement systématique et renforcé sur le plan technologique a des effets réels sur les droits fondamentaux des personnes ayant diverses identités de genre et affecte la jouissance de leurs droits liés à l'assistance sociale, tels que le logement, le travail et les soins de santé.[184] En outre, la conception des ensembles de données peut affecter l'identité des individus. Un ensemble de données qui capture le genre de manière binaire, par exemple, se trompe sur le genre des individus ayant diverses identités de genre.[185]

182 *Ibid.*

183 Sun S. D. (2019). Stop Using Phony Science to Justify Transphobia, disponible sur : https://blogs.scientificamerican.com/voices/stop-using-phony-science-to-justify-transphobia/; see also UN, OHCHR and the human rights of LGBTI people, available at: https://www.ohchr.org/en/sexual-orientation-and-gender-identity. Voir aussi : UNESCO (2022). Glossary: Understanding concepts around gender equality and inclusion in education, disponible sur : https://unesdoc.unesco.org/ark:/48223/pf0000380971

184 Leufer D. (2021). Computers are binary, people are not: how AI systems undermine LGBTQ identity, disponible sur : https://www.accessnow.org/how-ai-systems-undermine-lgbtq-identity/

185 Conseil des droits de l'homme des Nations Unies (2021). The right to privacy in the digital age. Report of the United Nations High Commissioner for Human Rights, disponible sur : https://www.ohchr.org/en/HRBodies/HRC/RegularSessions/Session48/Documents/A_HRC_48_31_AdvanceEditedVersion.docx

AymurAI : Une intelligence artificielle responsable, pour une justice ouverte et sensible au genre

AymurAI est une initiative visant à promouvoir une justice ouverte et sensible au genre, en Amérique latine. Cette initiative vise à aider les fonctionnaires des tribunaux pénaux et les juges qui souhaitent promouvoir l'ouverture des données dans leurs tribunaux pénaux. AymurAI est un logiciel basé sur l'intelligence artificielle (IA) qui identifie semi-automatiquement les informations importantes dans les décisions judiciaires et crée des ensembles de données ouverts axés sur les données relatives à la violence basée sur le genre (VBG). Il dispose également d'un outil d'anonymisation qui détecte les informations sensibles dans les décisions des tribunaux pénaux et les rédige. « AymurAI » signifie « récolte », en quechua. Cet outil vise à « récolter » les données des résolutions judiciaires en général, avec un accent particulier sur les cas de violence basée sur le genre. Il est « semi-automatisé », car il ne fonctionne pas de manière autonome sans intervention humaine et ne prend pas de décision. AymurAI aide à détecter les informations pertinentes et rationalise la collecte des condamnations judiciaires, mais la validation humaine des résultats du logiciel est cruciale pour garantir des résultats précis et fiables.

AymurAI est une application de bureau qui lit la résolution du tribunal, détecte les informations pertinentes, les présente à l'utilisateur pour validation, puis les stocke dans un ensemble de données pouvant être publié. L'outil utilise des règles et la reconnaissance des entités nommées (NER) pour extraire des informations essentielles des documents judiciaires. Dans les cas de violence basée sur le genre, les balises peuvent représenter le type de violence, le lieu, le genre, la relation avec l'auteur, la décision du juge dans cette affaire et d'autres données pertinentes. Ces balises passent par un processus de validation et, une fois approuvées, les informations collectées sont structurées en un ensemble de données ouvert. Tout cela est réalisé en quatre étapes simples.

Le projet est né du manque de données unifiées sur la violence sexiste en Argentine (la seule base de données officielle ouverte étant celle du Bureau de la violence domestique de la Cour suprême de justice et les rapports du Registre unique des cas de violence sexiste, qui ne dispose de données que jusqu'en 2018). À l'inverse, AymurAI peut aider à partager des informations sur la violence basée sur le genre et sur la manière dont elle est traitée dans différents jugements.

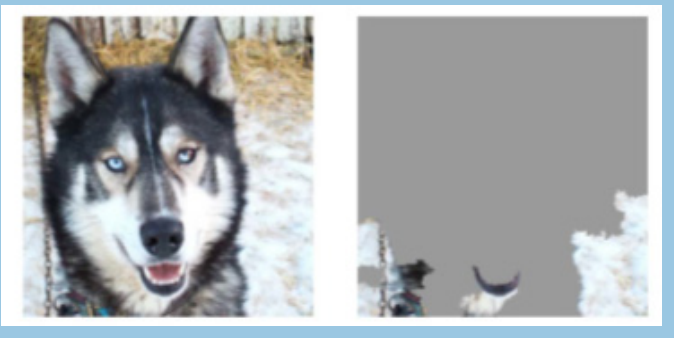
AymurAI est actuellement mis en œuvre au tribunal pénal 10 de la ville de Buenos Aires. Cette Cour pénale, dirigée par Pablo Casas, promeut, conçoit et permet l'application de politiques de justice ouverte, à travers sa base de données publique. Cette base de données est tenue à jour par les personnes qui travaillent au tribunal. La base de données compte environ cinq mille décisions judiciaires anonymisées depuis août 2016, notamment de nombreux cas de VBG. Il contient 64 catégories avec des informations détaillées sur chaque décision de justice, comme le type de violence subie par la victime dans chaque cas, conformément à la loi argentine n° 26.485. La base de données comprend également des données contextuelles (par exemple, les variables socio-économiques des personnes impliquées dans le conflit, si le défendeur a des enfants avec la victime et les phrases utilisées lors des agressions). Les employés du tribunal 10 utilisent différents outils pour entretenir la base de données. Par exemple, ils utilisent un outil pour anonymiser les décisions de justice appelé IA2.[186]

186 https://www.aymurai.info.

Activité :

L'IA peut créer un risque imprévu qui peut avoir des conséquences potentiellement mortelles. Lisez l'exemple ci-dessous et discutez des questions avec les participants.

Un algorithme volontairement défectueux a été développé par des chercheurs de l'Université de Washington qui ont catégorisé les photos de chiens husky et de loups. L'algorithme a exploité la présence ou l'absence de neige pour faire la distinction entre les huskies domestiques et les loups sauvages. Sur l'ensemble de données d'entraînement, les loups sont apparus dans la neige plus fréquemment que les huskies. Par conséquent, toutes les images de chiens lupins avec de la neige ont été classées comme des loups par le système. En conséquence, l'IA risquait de fournir des résultats incorrects 50 % du temps.[187]



Parce que les pixels qui définissent les loups sont ceux de la toile de fond enneigée (à droite), un husky (à gauche) est confondu avec un loup. Cet artefact résulte d'une base d'apprentissage mal représentée.

Source : Besse P., Castets-Renard C., Garivier A., Loubes J.-M. (2018). L'IA quotidienne peut-elle être éthique ? Machine Learning Algorithm Fairness, disponible sur : https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3391288

Cet exemple montre qu'il pourrait être fatal que les systèmes d'IA utilisés dans des domaines à enjeux élevés soient formés en utilisant une base d'apprentissage mal représentée.[188] Par exemple, dans le système de santé, les données de groupes de population spécifiques ont tendance à manquer dans les données avec lesquelles les outils de ML apprennent, ce qui signifie que l'outil pourrait fonctionner moins bien pour ces communautés. Pour illustrer cela, une équipe de scientifiques britanniques a constaté que presque tous les ensembles de données sur les maladies oculaires proviennent de patients en Amérique du Nord, en Europe et en Chine, ce qui signifie que les algorithmes de diagnostic des maladies oculaires sont moins sûrs de bien fonctionner pour les groupes raciaux des pays sous-représentés.[189] Un autre exemple est que les algorithmes de détection du cancer de la peau ont tendance à être moins précis lorsqu'ils sont utilisés sur des patients noirs, parce que les modèles de ML sont formés principalement sur des images de patients à la peau claire.[190]

187 Pearson D. (2021). AI biopsy dilemma: Wolf or husky, equity or bias?, disponible sur : https://healthexec.com/topics/precision-medicine/ai-biopsy-dilemma-wolf-or-husky-equity-or-bias.
188 Access Now (2018). Human rights in the age of artificial intelligence, disponible sur : https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf
189 Knight W. (2020). AI Can Help Diagnose Some Illnesses—If Your Country Is Rich, disponible sur : https://www.wired.com/story/ai-diagnose-illnesses-country-rich/
190 Lashbrook A. (2018). AI-Driven Dermatology Could Leave Dark-Skinned Patients Behind, disponible sur : https://www.theatlantic.com/health/archive/2018/08/machine-learning-dermatology-skin-color/567619/

Questions pour la discussion :

1. Quels étaient les principaux facteurs utilisés par le système pour différencier les huskies domestiques des loups sauvages ?
2. Y avait-il des failles dans cette analyse et pourquoi ?
3. Que se passerait-il si les processus de prise de décision de l'IA déployés dans les systèmes de justice utilisaient des algorithmes aussi défectueux ?

3. Pourquoi la transparence et la responsabilité algorithmiques sont importantes dans le contexte du pouvoir judiciaire ?

Le manque de transparence algorithmique est une question de premier plan des discussions sur l'IA et les droits humains. Le déploiement de systèmes d'IA dans le système judiciaire soulève des préoccupations quant à la manière d'évaluer en profondeur les effets à court et à long terme, dont les intérêts sont servis par les algorithmes, et s'ils sont sensibles au contexte pour faire face au contexte socioculturel dans différents pays.

Cette opacité des systèmes d'IA est alarmante. Un débat politique éclairé est impossible sans la capacité à comprendre le fonctionnement des systèmes d'IA. L'opacité dans la façon dont les systèmes d'IA arrivent à leurs décisions et la difficulté à déterminer la responsabilité de leurs actions signifient que des atteintes aux droits humains peuvent se produire lorsque de tels systèmes sont utilisés.[191]

Dans le même temps, il se peut également que, même lorsque les décisions basées sur l'IA peuvent être expliquées, les personnes concernées par la décision ne soient pas d'accord avec le résultat. Dans de telles situations, les parties concernées devraient avoir droit à un recours juridique. Contrairement aux procédures robustes qui existent dans de nombreux contextes juridiques pour promouvoir la responsabilité des décisions humaines dans les gouvernements - des lois sur la liberté d'information aux protections juridictionnelles et aux procédures d'appel - les algorithmes fonctionnent principalement dans une zone exempte de responsabilité. Cette section traitera de la transparence algorithmique et de la responsabilité dans le contexte des opérations judiciaires.

Transparence algorithmique

Lorsqu'il s'agit d'un système d'IA, la transparence fait référence à la quantité d'informations mises à la disposition de l'utilisateur. La structure du modèle, ses utilisations prévues, comment et quand les décisions de déploiement ont été prises et par qui en font partie, tout comme les décisions de conception et les données de formation.[192]

191 Deeks A. (2019). The Judicial Demand for Explainable Artificial Intelligence, 119 Colum. L. Rev. Virginia Public Law and Legal Theory Research Paper No. 2019-51, disponible sur : https://ssrn.com/abstract=3440723

192 Malek Md. A. (2021). Transparency in Predictive Algorithms: A Judicial Perspective, disponible sur : https://doi.org/10.31124/advance.14699937.v2

Les utilisateurs d'un système d'IA déployé dans le système judiciaire (par exemple, les plaignants et les défendeurs) ne savent souvent pas comment le système d'IA a été formé et comment il prend ses décisions. Par conséquent, lorsqu'il s'agit d'intenter une action en justice contre les résultats erronés et néfastes du système d'IA, il est difficile pour les personnes affectées par son utilisation de les contester, en l'absence de transparence sur la façon dont le système a été conçu et son fonctionnement.[193]

Le besoin de transparence algorithmique comprend les demandes adressées aux entreprises de divulguer leurs algorithmes propriétaires afin qu'ils puissent être examinés par des auditeurs indépendants, des régulateurs ou le grand public, avant leur mise en œuvre. Cependant, il est peu probable que les algorithmes ou le code logiciel sous-jacent soient mis à la disposition du public, car les entreprises privées considèrent leur algorithme comme un actif propriétaire clé et ne souhaitent pas le divulguer.

La Cour de justice européenne a déclaré que les entreprises ne peuvent pas déclarer et faire valoir devant les tribunaux qu'elles ne sont pas autorisées ou ne peuvent pas divulguer leurs algorithmes en raison de considérations de propriété intellectuelle (PI) ou de secret commercial, afin d'échapper à leur responsabilité d'expliquer l'IA (en vertu de l'article 22 du RGPD), à l'exception de l'IA qui sert un objectif de sécurité nationale ou d'affaires pénales. Il convient toutefois de noter qu'une transparence adéquate des systèmes automatisés est compliquée et difficile à atteindre, en raison des fréquents changements d'algorithmes. Par exemple, Google modifie son algorithme des centaines de fois par an.[194] De plus, le risque de manipuler des algorithmes augmente s'ils sont rendus publics.

193 Felzmann H., Fosch-Villaronga E., Lutz C., Tamò-Larrieux A. (2020). Towards Transparency by Design for Artificial Intelligence, Sci Eng Ethics 26, 3333–3361, disponible sur : https://doi.org/10.1007/s11948-020-00276-4

194 https://searchengineland.com/google-seo-news-google-algorithm-updates.

Étude de cas : La transparence algorithmique en pratique

- Royaume-Uni : Le Central Digital and Data Office et le Centre for Data Ethics and Innovation (CDEI) du Royaume-Uni ont publié l'une des premières directives nationales sur la transparence algorithmique dans le monde, en 2021. La norme consiste en un modèle que les organisations du secteur public sont encouragées à suivre pour tout outil algorithmique qui engage directement le public (tel qu'un chatbot) ou répond à des exigences spécifiques basées sur les risques. Les informations collectées sur les outils d'IA sont disponibles dans un registre public.[195]

- France, Pays-Bas et Nouvelle-Zélande : Ces trois pays ont également élaboré des directives pour aider les responsables du secteur public à naviguer dans l'utilisation responsable des algorithmes. L'Etalab français soutient les agences gouvernementales dans la mise en œuvre du cadre juridique pour la responsabilité et la transparence des algorithmes du secteur public.[196]

- États-Unis : Plusieurs gouvernements locaux aux États-Unis ont mis en place des interdictions ou des arrêts temporaires de l'utilisation de technologies algorithmiques, telles que les technologies de reconnaissance faciale (TRF), pour l'application de la loi et la surveillance. L'objectif principal de ces lois est de répondre aux préoccupations concernant la vie privée, mais il existe également des intersections importantes avec les questions de responsabilité algorithmique. Ces interdictions sont généralement établies par la législation, mais certaines lois ont prévu des exceptions limitées à l'interdiction, telles que des informations de tiers obtenues par le biais de TRF. Par exemple, un projet de loi à San Francisco interdisant l'utilisation des TRF ne s'applique qu'aux utilisations par les agences municipales et exclut l'utilisation par les agences fédérales, telles que celles des ports et des aéroports.[197]

- Chili : GobLab, un laboratoire d'innovation au sein de l'École de gouvernement de l'Université Adolfo Ibañez à Santiago, a mené des recherches approfondies sur l'utilisation des algorithmes par le gouvernement chilien, en collaboration avec le Conseil chilien de la transparence. Avec le financement de la Banque interaméricaine de développement, le groupe a rédigé et proposé un règlement que le gouvernement est sur le point d'adopter, après les premiers tests auprès de divers organismes publics. Le règlement fera du Chili le premier pays d'Amérique latine à adopter des normes sur la transparence algorithmique.[198]

- Initiatives au niveau des villes : La transparence algorithmique dans l'UE a été introduite *ex ante* au niveau local depuis octobre 2020, dans les villes d'Amsterdam[199], de Helsinki[200], et de Nantes[201], établissant des registres décrivant les algorithmes utilisés dans les administrations municipales. Pour s'assurer que l'IA utilisée par les services publics est centrée sur l'humain, les registres indiquent, entre autres, comment les données sont traitées, quels dangers sont impliqués et si les technologies sont soumises à une surveillance humaine.[202]

195 Centre for Data Ethics and Innovation (2023). Algorithmic Transparency Recording Standard Hub. gov.uk, disponible sur : https://www.gov.uk/government/collections/algorithmic-transparency-recording-standard-hub
196 Turak H. (2020). Open algorithms: Experiences from France, the Netherlands, and New Zealand. Open Government Partnership, disponible sur :https://www.opengovpartnership.org/stories/open-algorithms-experiences-from-france-the-netherlands-and-new-zealand/.
197 Haataja M, van de Fliert L., Rautio P. (2020). Public AI Registers: Realising AI transparency and civic participation in government use of AI Saidot, disponible sur : https://openresearch.amsterdam/en/page/73074/public-ai-registers
198 Aránguiz Villagrán M. (2022). Algorithmic Audit for Decision-Making or Decision Support Systems. Inter-American Development Bank, disponible sur : http://dx.doi.org/10.18235/0004154
199 Voir : https://algoritmeregister.amsterdam.nl/en/ai-register
200 Voir : https://ai.hel.fi/en/ai-register/
201 Voir : https://data.nantesmetropole.fr/pages/algorithmes_nantes_metropole/
202 *Ibid.*

La transparence est encore compliquée par le problème de la boîte noire des systèmes d'IA (abordé dans le module 1). Même fournir le code source de l'algorithme peut ne pas suffire. Il est nécessaire d'expliquer comment les résultats d'un algorithme sont générés.[203] L'un des objectifs réglementaires les plus importants pour l'utilisation sûre et responsable des algorithmes dans le secteur public est d'établir des normes d'explicabilité.

Étude de cas : La transparence algorithmique du point de vue de l'élaboration des politiques publiques : l'exemple de la France

En France, la loi de 2016 pour une République numérique stipule que chaque fois qu'un organisme public soumet des résidents à un traitement algorithmique, ces derniers ont le droit d'être informés : 1) du degré auquel le traitement algorithmique contribue à la prise de décision ; 2) des données traitées ; 3) des paramètres de traitement ; 4) des opérations auxquelles un tel traitement est appliqué. Les informations doivent être communiquées à toute personne, sur demande, dans une langue intelligible et sans porter atteinte aux secrets protégés par la loi.

En 2018, lorsque le Conseil constitutionnel a débattu d'un projet de loi visant à aligner la loi française sur la protection des données sur le RGPD, il a statué que si un organisme public ne peut pas communiquer les principes de fonctionnement d'un algorithme sans compromettre les secrets protégés, aucune décision ne peut être prise uniquement sur la base d'un tel algorithme. Ainsi, si une entité publique fonde sa décision uniquement sur un algorithme, le secret commercial ne peut pas être invoqué pour éviter de divulguer le fonctionnement de l'algorithme.

Source : Décret n° 2017-330 du 14 mars 2017 relatif aux droits des personnes faisant l'objet de décisions individuelles prises sur le fondement d'un traitement algorithmique, disponible sur : https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000034194929?r=EILBrO52Ri ; voir également : https://www.conseil-constitutionnel.fr/decision/2018/2018765DC.htm.

Responsabilité algorithmique

La responsabilité algorithmique fait référence à la capacité de ceux qui conçoivent, construisent, achètent ou mettent en œuvre l'algorithme à être tenus responsables de leurs actions et de leur impact, conformément aux politiques et aux lois concernant l'utilisation de l'algorithme. Un système de gouvernance doté d'un acteur responsable exige que l'acteur soit capable d'expliquer et de justifier ses décisions concernant l'algorithme, et de faire face à des conséquences si ses actions sont contraires à la loi.[204]

Responsabilité dès la conception

« Tous les systèmes d'IA doivent être conçus pour faciliter la responsabilisation et l'auditabilité de bout en bout. Cela nécessite à la fois des humains responsables dans la boucle tout au long de la chaîne de conception et de mise en œuvre, ainsi que des protocoles de surveillance des activités qui permettent une surveillance et un examen de bout en bout. »

Source : Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector, The Alan Turing Institute, disponible sur : https://doi.org/10.5281/zenodo.3240529

203 *Ibid.*

204 Ada Lovelace Institute, AI Now Institute and Open Government Partnership (2021). Algorithmic Accountability for the Public Sector, disponible sur : https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/

Les défis de la responsabilité algorithmique peuvent être liés au fait que le décideur (par exemple, le juge) ne contrôle pas les sources de données (données obtenues par l'intermédiaire de courtiers en données ou par les autorités chargées de l'application de la loi, à l'aide d'outils d'évaluation des risques). Ils pourraient également découler du fait qu'il est très difficile de traduire des concepts algorithmiques complexes (par exemple, les résultats des algorithmes de regroupement qui segmentent les populations en fonction d'un grand nombre de variables d'entrée) en concepts compréhensibles par l'homme (par exemple, l'affiliation raciale). Cela pourrait entraîner une interprétation inexacte des résultats algorithmiques. Les défis de responsabilité algorithmique peuvent également être déclenchés par des asymétries d'information. Par exemple, l'opacité des algorithmes de ML pourrait empêcher les personnes concernées de connaître et de comprendre les résultats du processus de prise de décision automatisée (ADM) ou même d'être conscientes qu'elles y ont été soumises. En outre, des problèmes peuvent survenir au stade de la mise en œuvre, lorsque des données contradictoires sont injectées dans le système pour le tromper en commettant des erreurs. Veuillez vous référer aux discussions du module 1 sur les questions de cybersécurité.[205]

4. Pleins feux sur l'identification biométrique, les technologies de reconnaissance faciale et les deepfakes

L'adoption de technologies à haut risque, telles que la reconnaissance faciale et l'identification biométrique, présente des défis aggravés pour les décideurs et les régulateurs du monde entier. Les ONG de défense des droits humains ont également dénoncé le manque de protection adéquate de la vie privée dans de nombreux systèmes nationaux d'identité biométrique, où l'accès aux prestations sociales et à d'autres services gouvernementaux était subordonné à l'enregistrement auprès du système.[206]

La résolution de l'Assemblée générale des Nations Unies sur le droit à la vie privée à l'ère numérique (2020) a fait référence au « piratage et à l'utilisation illégale des technologies biométriques », comme « des actes hautement intrusifs qui violent le droit à la vie privée », interfèrent avec la liberté d'expression et d'opinion, la liberté de réunion et d'association pacifiques et la liberté de religion ou de conviction, et « peuvent contredire les principes d'une société démocratique, y compris lorsqu'ils sont entrepris de manière extraterritoriale ou à grande échelle ».[207]

Un rapport du Haut-Commissaire des Nations Unies aux droits de l'homme de 2021, intitulé « Le droit à la vie privée à l'ère numérique », a appelé à un moratoire sur l'utilisation des technologies de reconnaissance faciale dans les espaces publics, jusqu'à ce que les gouvernements puissent montrer qu'il n'y a pas de problèmes substantiels liés à l'exactitude ou aux impacts discriminatoires, et que ces technologies respectent des normes

205 Parlement européen (2019). A governance framework for algorithmic accountability and transparency, disponible sur : https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2019)624262

206 https://www.ohchr.org/Documents/Issues/Poverty/DigitalTechnology/AmnestyInternational.pdf

207 Assemblée générale des Nations Unies (2020). The right to privacy in the digital age, disponible sur : https://documents-dds-ny.un.org/doc/UNDOC/GEN/N20/371/75/PDF/N2037175.pdf?OpenElement

strictes en matière de confidentialité et de protection des données.[208]

La reconnaissance biométrique est basée sur la comparaison de la représentation numérique d'une personne, de son visage, de son empreinte digitale, de son iris, de sa voix ou de son mouvement avec d'autres représentations similaires stockées dans une base de données. Sur cette base, le système décide de la probabilité qu'un individu corresponde bien à la personne à identifier. Les autorités du monde entier utilisent de plus en plus la reconnaissance faciale à distance en temps réel, comme une forme de reconnaissance biométrique.[209]

La Haut-Commissaire des Nations Unies aux droits de l'homme a indiqué que « la reconnaissance biométrique en temps réel soulève de graves préoccupations en vertu du droit international relatif aux droits de l'homme ».[210] Certaines de ces préoccupations reflètent des problèmes liés aux techniques prédictives, tels que la probabilité d'une identification incorrecte des personnes et des effets disproportionnés sur les membres de certains groupes (le plus souvent marginalisés).[211] Les individus peuvent être profilés à l'aide de la technologie de reconnaissance faciale en fonction de leur race, origine ethnique, origine nationale, sexe/genre et d'autres caractéristiques.[212]

La reconnaissance biométrique à distance est associée à une interférence significative avec le droit à la vie privée. Les informations biométriques d'une personne sont l'un des aspects clés de sa personnalité, car elles exposent des caractéristiques spécifiques qui la distinguent des autres.[213] La reconnaissance biométrique à distance permet aux autorités gouvernementales d'identifier et de suivre systématiquement les individus dans les espaces publics, ce qui peut avoir un impact négatif sur l'exercice des droits à la liberté d'expression, de réunion pacifique, d'association et de libre circulation.[214]

208 Conseil des droits de l'homme des Nations Unies (2021). The right to privacy in the digital age. Report of the United Nations High Commissioner for Human Rights, disponible sur : https://www.ohchr.org/en/HRBodies/HRC/RegularSessions/Session48/Documents/A_HRC_48_31_AdvanceEditedVersion.docx

209 *Ibid.*

210 Conseil des droits de l'homme des Nations Unies (2020). Impact of new technologies on the promotion and protection of human rights in the context of assemblies, including peaceful protests, disponible sur : https://documents-dds-ny.un.org/doc/UNDOC/GEN/G20/154/35/PDF/G2015435.pdf?OpenElement

211 Conseil des droits de l'homme des Nations Unies (2021). The right to privacy in the digital age. Report of the United Nations High Commissioner for Human Rights, disponible sur : https://www.ohchr.org/en/HRBodies/HRC/RegularSessions/Session48/Documents/A_HRC_48_31_AdvanceEditedVersion.docx.

212 Conseil des droits de l'homme des Nations Unies (2020). Racial discrimination and emerging digital technologies: a human rights analysis, par. 39–40, disponible sur : https://www.ohchr.org/sites/default/files/HRBodies/HRC/RegularSessions/Session44/Documents/A_HRC_44_57_AdvanceEditedVersion.docx

213 Conseil des droits de l'homme des Nations Unies (2020). Impact of new technologies on the promotion and protection of human rights in the context of assemblies, including peaceful protests, para. 33, disponible sur : https://documents-dds-ny.un.org/doc/UNDOC/GEN/G20/154/35/PDF/G2015435.pdf?OpenElement. Voir aussi Cour européenne des droits de l'homme, Reklos et Davourlis c. Grèce, requête n ° 1234/05, arrêt du 15 avril 2009, par. 40.

214 Voir : Comité européen de la protection des données et Contrôleur européen de la protection des données (2021). Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), par. 30, disponible sur : https://edpb.europa.eu/system/files/2021-06/edpb-edps_joint_opinion_ai_regulation_en.pdf; Conseil des droits de l'homme des Nations Unies (2020). Impact of new technologies on the promotion and protection of human rights in the context of assemblies, including peaceful protests, disponible sur : https://documents-dds-ny.un.org/doc/UNDOC/GEN/G20/154/35/PDF/G2015435.pdf?OpenElement.; Conseil des droits de l'homme des Nations Unies (2019). Surveillance and human rights, disponible sur : https://documents-dds-ny.un.org/doc/UNDOC/GEN/G19/148/76/PDF/G1914876.pdf?OpenElement

Études de cas

Le RGPD et les données biométriques

Le RGPD de l'UE limite le traitement des données biométriques dans une certaine mesure. Ce n'est que lorsque les données sont liées à une personne spécifique qu'elles deviennent des données à caractère personnel et sont donc couvertes par le présent règlement. Selon le RGPD, les données biométriques sont « des données personnelles résultant d'un traitement technique spécifique relatif aux caractéristiques physiques, physiologiques ou comportementales d'une personne physique, qui permettent ou confirment l'identification unique de cette personne physique ». Ainsi, si la reconnaissance biométrique ne vise pas à identifier (mais plutôt à catégoriser, profiler ou affecter la reconnaissance), elle peut ne pas relever de la définition du RGPD.

Selon le considérant 51 du RGPD, « le traitement de photographies [est considéré comme] des données biométriques uniquement lorsqu'il est traité par un moyen technique spécifique permettant l'identification ou l'authentification unique d'une personne physique ».

Source : https://rm.coe.int/prems-107320-gbr-2018-compli-cahai-couv-texte-a4-bat-web/1680a0c17a

Le cas de Clearview AI

L'Autorité de protection des données de l'État allemand de Hambourg a décidé que Clearview AI traitait illégalement les données biométriques obtenues et mises à disposition en tant que service. En outre, il n'existait aucune base juridique valide pour le traitement des données. La Cour a noté que Clearview AI a traité des données biométriques (en vertu de l'article 4(14) du RGPD), car elle « utilise une procédure mathématique spécialement développée pour générer une valeur de hachage unique de la personne concernée qui permet l'identification ». Le litige a été initié par une plainte de la personne concernée, celle-ci n'ayant pas donné son consentement pour le traitement de ses données biométriques. L'Autorité de protection des données a déterminé que même si Clearview AI n'était pas établie dans l'UE, elle était soumise au RGPD par le biais de la surveillance de l'activité en ligne des personnes concernées (article 3 (2)(b) du RGPD), car elle « n'offre pas un instantané [des individus], mais de toute évidence, archive également les sources sur une période de temps ». Clearview AI a reçu l'ordre de supprimer toutes les données personnelles du plaignant.

Source : Future of Privacy Forum (2022). GDPR and the AI Act interplay: Lessons from FPF's ADM Case Law Report, disponible sur : https://fpf.org/blog/gdpr-and-the-ai-act-interplay-lessons-from-fpfs-adm-case-law-report

Les technologies de reconnaissance faciale utilisent des images numériques pour identifier et valider les visages humains. Ces technologies fonctionnent en identifiant les caractéristiques du visage dans une image source et en les comparant à travers un ensemble de données. Les technologies de reconnaissance faciale ont un large éventail d'utilisations, bien qu'elles soient le plus souvent utilisées à des fins de sécurité, telles que les activités de police et de sécurité nationale (par exemple, la lutte contre le terrorisme). Les progrès de l'IA ont amélioré la capacité et la sophistication de ces technologies au cours des dernières années, ce qui en fait une composante standard des biens de consommation tels que les téléphones mobiles, qui permettent aux utilisateurs de « se connecter » en utilisant leur visage.[215]

215 Hill D., O'Connor C. D., Slane A. (2022). Police use of facial recognition technology: The potential for engaging the public through co-constructed policy-making, International Journal of Police Science & Management, 24(3), 325–335, disponible sur : https://doi.org/10.1177/14613557221089558

Les technologies de reconnaissance faciale font polémique dans le secteur privé

Plusieurs entreprises, dont Microsoft et IBM, ont été critiquées pour avoir déployé un logiciel de reconnaissance faciale plus précis pour certaines données démographiques que d'autres. Plus précisément, ces systèmes ont tendance à identifier avec précision les hommes à la peau claire beaucoup plus souvent que les femmes à la peau plus foncée.

De même, une controverse a surgi lorsque le logiciel de marquage automatique de photos de Google a identifié de nombreuses photos d'Afro-Américains comme étant des « gorilles » ou des « singes ». La cause de ces erreurs pourrait résider dans le développement des modèles algorithmiques. Les modèles ont été formés avec des ensembles de données de photos de personnes principalement d'origine caucasienne, et n'ont donc pas été formés avec suffisamment de données pour identifier les personnes non blanches, en particulier les femmes. Le travail de Joy Buolamwini, informaticienne au MIT et fondatrice de l'Algorithmic Justice League, a incité de nombreuses entreprises à publier des déclarations répondant aux critiques et à réformer leurs modèles.

Source : https://www.poetofcode.com/

En novembre 2021, Meta a annoncé qu'elle « fermait le système de reconnaissance faciale sur Facebook », citant des règles peu claires des régulateurs. De même, IBM va cesser de proposer son logiciel de reconnaissance faciale pour certaines activités, dont la surveillance de masse.

Source : Gouvernement britannique (2022). Policy paper Establishing a pro-innovation approach to regulating AI, disponible sur : https://www.gov.uk/government/publications/establishing-a-pro-innovation-approach-to-regulating-ai/establishing-a-pro-innovation-approach-to-regulating-ai-policy-statement

Le recours à l'identification biométrique et aux technologies de reconnaissance faciale dans les opérations judiciaires peut devenir une boîte de Pandore pour différents types de biais, tels que ceux basés sur la race ou le genre. Le cas des données ImageNet peut être utilisé comme exemple illustratif. Il s'agit d'un ensemble de données clé pour le développement d'applications de vision par ordinateur, qui contient plus de 45 % des images provenant des États-Unis, contre seulement 3 % de la Chine et de l'Inde combinées. Ce manque de diversité contribue aux lacunes des algorithmes de reconnaissance d'image, qui interprètent les yeux asiatiques comme clignotant en continu, étiquettent l'image d'une mariée américaine traditionnelle vêtue de blanc comme « mariée », « robe », « femme » et « mariage », mais celle d'une mariée indienne comme « art de la performance » et « costume », et identifient à tort le sexe/genre des femmes à la peau plus foncée avec un taux d'erreur de 35 %, tout en identifiant à tort le sexe/genre des hommes à la peau plus claire avec un taux d'erreur de 0 %.[216]

Bien que la surveillance de masse basée sur l'IA par le biais de la reconnaissance faciale implique la collecte, le stockage et le traitement de données personnelles (biométriques, en l'occurrence nos visages), elle a également un impact sur

216 Parlement européen (2019). A governance framework for algorithmic accountability and transparency, disponible sur : https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2019)624262

notre vie privée, notre identité et notre autonomie, en ouvrant la possibilité d'être observé, suivi et reconnu.[217] Les gens peuvent se sentir obligés d'adhérer à une norme particulière, en raison de l'effet de « refroidissement » psychologique, modifiant l'équilibre des pouvoirs entre la personne et le gouvernement ou l'entreprise privée utilisant la technologie de reconnaissance faciale.

Bien que la reconnaissance faciale puisse avoir un effet plus prononcé sur le droit à la vie privée et à l'intégrité psychologique, on pourrait soutenir que le suivi numérique de tous les aspects de la vie humaine (via les données de localisation, les données IdO des montres intelligentes, les trackers de santé, les haut-parleurs intelligents, les thermostats, les véhicules, etc.) pourrait avoir un impact similaire. La fréquence cardiaque, la température corporelle et d'autres types de reconnaissance biométrique basée sur l'IA mesurent ou même prédisent notre comportement, notre état mental et nos émotions. Cela peut avoir de graves répercussions sur le droit à la vie privée dans l'environnement en ligne.[218]

Approfondissement : les systèmes de reconnaissance faciale peuvent mal identifier le genre

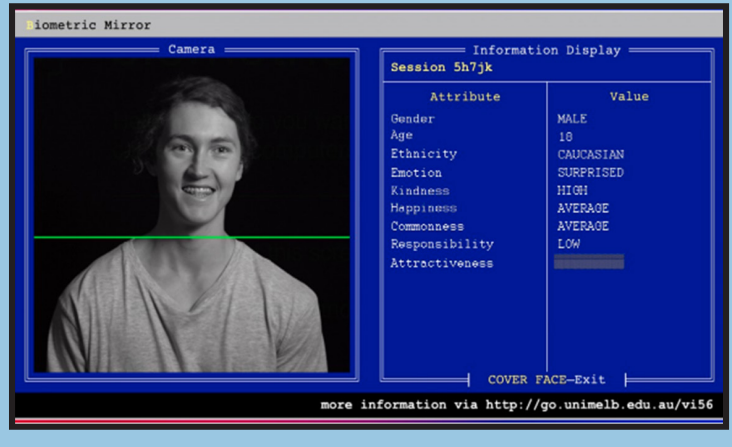
Les systèmes d'IA destinés à « genrer » les individus dans des contextes publics ne sont pas futuristes, ils sont déjà utilisés dans le monde entier. À São Paulo, au Brésil, l'Institut brésilien pour la protection des consommateurs (IDEC) a contesté l'installation et l'utilisation de panneaux d'affichage intelligents qui prétendent anticiper l'émotion, l'âge et le sexe des usagers du métro, pour leur fournir de « meilleures publicités ».[219]

217 Secrétariat du CAHAI (2020). Towards Regulation of AI Systems. Global perspectives on the development of a legal framework on Artificial Intelligence (AI) systems based on the Council of Europe's standards on human rights, democracy and the rule of law, Council of Europe Study, DGI/2020/16, disponible sur : https://rm.coe.int/prems-107320-gbr-2018-compli-cahai-couv-texte-a4-bat-web/1680a0c17a.
218 *Ibid.*
219 Voir : https://idec.org.br/sites/default/files/acp_viaquatro.pdf

Activité :

Les participants à la formation regardent la vidéo et discutent des implications sociétales des technologies d'IA et de reconnaissance faciale. Les participants discutent également de l'impact que ces technologies pourraient avoir sur leur travail. Comment les technologies de reconnaissance faciale affectent-elles les droits humains ? Quels sont les groupes les plus vulnérables et les plus exposés aux violations des droits humains par les technologies de reconnaissance faciale ?

Des chercheurs basés à Melbourne ont demandé à des volontaires humains de juger des milliers de photos sur les mêmes caractéristiques, puis ont utilisé cet ensemble de données pour en créer le miroir biométrique. Celui-ci utilise l'IA pour analyser le visage d'une personne en le scannant, et affiche plus tard 14 caractéristiques à son sujet, tels que son âge, sa race et son niveau d'attractivité perçu. Il utilise un ensemble de données ouvertes de milliers d'évaluations faciales et participatives. Cependant, cette analyse est souvent fausse, car l'IA génère l'analyse sur la base d'informations subjectives et biaisées fournies par des volontaires humains initiaux.[220]



Lien vers la vidéo : https://youtu.be/fb_sfhT0mrg

220 Houser K. (2018). Biased AI biometric mirror, disponible sur : https://futurism.com/the-byte/biased-ai-biometric-mirror.

Deepfakes

Les deepfakes sont une technologie d'IA particulièrement dangereuse qui a un impact sur les droits humains. Un deepfake est toute forme de média (vidéo, audio ou autre) qui a été modifié ou entièrement ou partiellement créé à partir de zéro.[221] Les machines peuvent apprendre à effectuer des tâches en regardant des exemples, à l'aide de réseaux neuronaux. Il existe plusieurs technologies qui peuvent être appliquées à cela, mais la plus populaire est basée sur les réseaux antagonistes génératifs (GAN) et les modèles de diffusion.[222]

221 Van der Sloot B., Wagensveld Y. (2022). Deepfakes: regulatory challenges for the synthetic society. Computer Law & Security Review, disponible sur : https://www.sciencedirect.com/science/article/pii/S0267364922000632, disponible sur : https://doi.org/10.1016/j.clsr.2022.105716

222 *Ibid.*

Réseaux antagonistes génératifs (GAN)

Les GAN sont une approche non supervisée de l'apprentissage profond qui peut générer du matériel hyperréaliste. Les GAN sont utilisés pour des techniques d'apprentissage en profondeur non supervisées, telles que la génération d'images réalistes ou d'ensembles de données d'image, la traduction de texte en image et d'image en texte, le vieillissement des visages et la création d'émojis. Les GAN utilisent deux réseaux neuronaux : un générateur qui génère de nouvelles instances et un discriminateur qui cherche à différencier ces images fausses, souvent de mauvaise qualité ou irréalistes des données d'image réelles entrées dans le système d'IA. Grâce à cette interaction, le générateur apprend à produire des images de plus en plus convaincantes et de haute qualité, qui finissent par tromper le discriminateur en lui faisant croire qu'elles font partie des données d'image réelles.[223]

Modèles de diffusion

Les modèles de diffusion sont des modèles génératifs plus avancés que les GAN sur la synthèse d'images. Plus récemment, les modèles de diffusion ont été utilisés dans DALL-E 2, le modèle de génération d'images d'OpenAI et Imagen de Google.[224] L'accès public à DALL-E est contrôlé via une longue liste d'attente et un paywall après plusieurs invites, tandis qu'Imagen de Google est interdit au public. La sortie de DALL-E est filtrée, ce qui rend difficile la génération d'images contenant de la violence, de la nudité ou des visages réalistes.[225]

Cependant, le nouveau programme de conversion de texte en image nommé Stable Diffusion, développé par Stability AI[226], offre une génération d'image en libre accès, non filtrée, libre d'utilisation pour tout le monde. Voici une image créée par Stable Diffusion à partir du texte exact « Photo de Bernie Sanders dans Mad Max Fury Road (2015), explosions, cheveux blancs, lunettes, vêtements en lambeaux, traits du visage symétriques détaillés, éclairage spectaculaire ».[227]

Image : Reddit / Licovoda

Source : The Verge (2022). Anyone can use this AI art generator — that's the risk https://www.theverge.com/2022/9/15/23340673/ai-image-generation-stable-diffusion-explained-ethics-copyright-data

Comme déjà indiqué, en 2023, Getty Images a intenté une action en contrefaçon contre Stability AI aux États-Unis, affirmant que l'entreprise avait copié 12 millions d'images « sans autorisation… ni compensation » pour former son modèle d'IA.

223 AAAS. Artificial Intelligence and the Courts: Materials for Judges, disponible sur : https://www.aaas.org/ai2/projects/law/judicialpapers
224 O'Connor R. (2022). Introduction to Diffusion Models for Machine Learning, disponible sur : https://www.assemblyai.com/blog/diffusion-models-for-machine-learning-introduction/.
225 Voir : https://labs.openai.com/policies/content-policy
226 Voir : https://stability.ai/
227 Vincent J. (2022). Anyone can use this AI art generator — that's the risk https://www.theverge.com/2022/9/15/23340673/ai-image-generation-stable-diffusion-explained-ethics-copyright-data

Le vrai problème associé aux deepfakes est la simplicité de générer tout un écosystème de fausses informations. Une fausse vidéo, de faux sites Web qui hébergent la vidéo et génèrent de la désinformation et de l'anti-information sur ce qui est affiché dans la vidéo, de faux comptes Twitter qui renvoient à la vidéo, de faux comptes sur des forums de discussion qui commentent le contenu de la vidéo, de faux comptes Instagram qui génèrent des mèmes de la fausse vidéo. Il s'avère extrêmement difficile de pénétrer un environnement de tromperie multicouche et complexe, et d'en obtenir des informations fiables.[228]

Activité :

Les participants regardent la vidéo et discutent de la façon dont les deepfakes pourraient affecter le travail des opérateurs judiciaires.

Lien vers la vidéo : https://youtu.be/oxXpB9pSETo

Les deepfakes et l'ensemble de l'écosystème falsifié qu'ils créent mettent en péril les droits à un procès équitable, à un recours effectif et à la présomption d'innocence. Ils pourraient être utilisés comme de fausses preuves devant les tribunaux. Les parties peuvent toujours faire valoir que les preuves présentées à leur encontre sont fausses et artificielles, et que les procès prendront plus de temps. Les deepfakes soulèvent également la possibilité qu'un juge accepte par erreur des preuves fabriquées comme fiables.[229] Par conséquent, le secteur judiciaire devrait commencer à investir dans des outils numériques qui facilitent l'évaluation médico-légale des preuves vidéo et audio, pour s'assurer que les preuves n'ont pas été générées par les GAN et les autoencodeurs variationnels. D'autre part, l'IA a le potentiel de vérifier l'authenticité des preuves numériques en détectant de faux algorithmes ou des données manipulées. L'utilisation de l'IA pour analyser une image ou une vidéo pourrait déterminer si elle a été manipulée d'une manière ou d'une autre. Cependant, il s'agit toujours d'un domaine de recherche en développement.

228 *Ibid.*
229 *Ibid.*

Ces activités de groupe visent à encourager les participants à la formation à discuter des divers défis juridiques et éthiques du déploiement de l'IA dans le système judiciaire.

Activité 1

Les participants à la formation lisent « State c. Loomis », dans le module 4, et répondent à la question suivante : Pensez-vous qu'il est approprié que le tribunal permette à un algorithme, sur lequel les acteurs du système juridique ont une visibilité limitée, de jouer un rôle, même mineur, dans la privation de liberté d'une personne ? Évaluez l'impact éthique de cette décision conformément à l'instrument d'évaluation de l'impact éthique de l'UNESCO, à l'annexe I [veuillez vous concentrer sur les parties qui traitent de l'équité, de la non-discrimination, de la diversité, de la protection des données et de la vie privée].

Activité 2

Les participants à la formation parcourent le matériel sur les modèles de diffusion présentés ci-dessus et lisent également l'article suivant : https://www.theverge.com/2023/2/6/23587393/ai-art-copyright-lawsuit-getty-images-stable-diffusion.

Les participants discutent ensuite des implications juridiques des poursuites intentées par Getty Images contre Stability AI. L'action en justice s'appuiera sur l'interprétation de la doctrine américaine de l'utilisation équitable, qui autorise l'utilisation non autorisée d'œuvres protégées par le droit d'auteur dans certaines circonstances. La notion d '« utilisation transformatrice » peut également constituer un aspect important. Le résultat de Stable Diffusion est-il suffisamment distinct de ses données d'entraînement ? Une étude récente a révélé que le programme mémorise certaines de ses images d'entraînement et peut les répéter presque à l'identique, bien que dans un nombre relativement limité de cas.

Les participants à la formation discutent de l'impact du développement et du déploiement de l'IA sur les règles du droit d'auteur dans leurs propres juridictions.

Activité 3

Discutez des implications juridiques et éthiques de cette affaire.

Cas australien - Maire victorien en costume ChatGPT

Un maire victorien, Brian Hood, se prépare à poursuivre OpenAI, si celui-ci ne corrige pas les fausses allégations de ChatGPT selon lesquelles il aurait purgé une peine de prison pour corruption. Les avocats de Hood ont envoyé une lettre de préoccupation à OpenAI le 21 mars, leur donnant 28 jours pour corriger les erreurs, mais OpenAI n'a pas encore répondu. Les fausses allégations étaient liées à un scandale de corruption transnationale impliquant une filiale de la Reserve Bank of Australia, au

230 Byron K. (2023). Victorian mayor readies defamation lawsuit over ChatGPT content, disponible sur : https://www.afr.com/technology/vinoctorian-mayor-readies-defamation-lawsuit-over-chatgpt-content-20230405-p5cyh5

début des années 2000, mais Hood n'a jamais été inculpé d'un quelconque crime.[230]

Activité 4

Les participants à la formation lisent le texte ci-dessous sur la façon dont les technologies de reconnaissance faciale peuvent envahir le droit à la vie privée et regardent les vidéos. Ensuite, les participants à la formation discutent de la façon dont les technologies de reconnaissance faciale et leurs risques peuvent être contestés en vertu de leurs lois nationales sur la protection des données et de la vie privée.

En mai 2020, l'American Civil Liberties Union (ACLU) a intenté une action en justice[231] au nom d'organisations représentant les victimes de violence domestique, les immigrants illégaux et les travailleurs du sexe. L'organisation a accusé Clearview, une société de technologie qui développe une technologie de reconnaissance faciale, d'avoir enfreint la Biometric Information Privacy Act (BIPA)[232] de l'Illinois, une loi de l'État qui empêche les entreprises commerciales d'exploiter les identifiants physiques des citoyens, y compris la cartographie informatique de leurs visages, sans leur consentement.[233]

La plainte a été déposée devant le tribunal de l'État de l'Illinois, à Chicago, après que le New York Times a révélé, en janvier 2020, que Clearview développait un système de suivi et de surveillance basé sur des identifiants biométriques. La technologie de reconnaissance faciale a permis à Clearview d'acquérir plus de trois milliards d'empreintes faciales à partir de photographies Web.[234]

Clearview a fourni l'accès à ces informations à des sociétés privées, à des personnes fortunées et à des organismes d'application de la loi fédéraux, étatiques et locaux. L'entreprise affirme qu'en utilisant cette grande base de données, elle peut identifier instantanément des personnes avec une précision inégalée, ce qui permet une surveillance clandestine et à distance étendue des Américains.[235]

La BIPA exige que les entreprises qui collectent, capturent ou obtiennent un identifiant biométrique d'un résident de l'Illinois, tel qu'une empreinte digitale, une empreinte faciale ou un scan de l'iris, doivent d'abord en informer le sujet et obtenir son consentement écrit. Cela est dû au fait que l'acquisition forcée d'identifiants biométriques immuables présente plus de dangers pour la sécurité, la vie privée et la sûreté d'un individu que la capture d'autres identifiants, tels que les noms et les adresses. Et l'enregistrement de l'empreinte faciale d'une personne — comparable à l'établissement de son profil ADN à partir de matériel génétique inévitablement versé sur une bouteille d'eau, mais distinct de la publication ou de la transmission d'une photographie — est un comportement, pas un discours, et est donc légitimement régi par la loi. Clearview n'a pas respecté la BIPA, privant plusieurs citoyens de l'Illinois de leurs droits à la vie privée.[236]

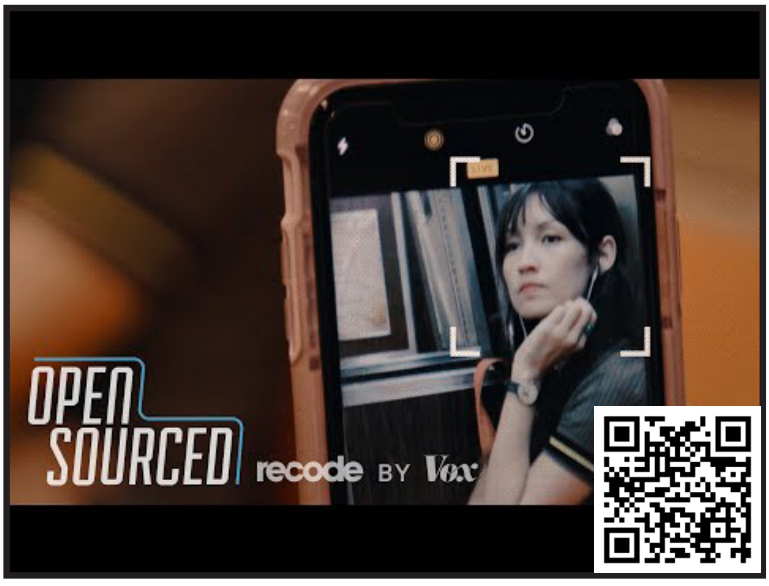
Cette action en justice a été la première à se concentrer sur les dommages que la technologie de Clearview causerait aux survivants d'abus domestiques et sexuels, aux migrants sans papiers, aux communautés de couleur et aux membres

231 Alba D. (2020). A.C.L.U. Accuses Clearview AI of Privacy 'Nightmare Scenario', disponible sur : https://www.nytimes.com/2020/05/28/technology/clearview-ai-privacy-lawsuit.html.
232 Voir : https://www.aclu-il.org/en/campaigns/biometric-information-privacy-act-bipa
233 Mac R., Hill K. (2022). Clearview AI settles suit and agrees to limit sales of facial recognition database. The facial recognition software maker is largely prohibited from selling its database of photos to private companies, disponible sur : https://www.nytimes.com/2022/05/09/technology/clearview-ai-suit.html
234 *Ibid.*
235 *Ibid.*
236 ACLU (2022). ACLU v. Clearview AI, disponible sur : https://www.aclu.org/cases/aclu-v-clearview-ai

d'autres populations vulnérables. Les membres, les clients et les participants au programme des organisations demanderesses ont été exposés à l'impression faciale par Clearview sans leur consentement et risquent de subir certains des effets les plus graves du programme de surveillance inégalé de Clearview.[237]

Le 11 mai 2022, après que les parties ont négocié un accord de règlement, le tribunal a approuvé une ordonnance de consentement rejetant cette affaire. L'élément fondamental du règlement interdit les opérations de Clearview non seulement dans l'Illinois, mais dans l'ensemble des États-Unis, interdisant définitivement à Clearview de rendre sa base de données d'empreintes faciales accessible aux organisations privées. En outre, la société se voit interdire de vendre l'accès à sa base de données à toute agence dans l'Illinois, y compris les autorités de l'État et municipales, pendant cinq ans.[238]

Source : https://youtu.be/s44EFtBoRxY



Source : https://youtu.be/cc0dqW2HCRc

237 *Ibid.*
238 ACLU, EXIBIT 2. signed settlement agreement, disponible sur : https://www.aclu.org/legal-document/exhibit-2-signed-settlement-agreement

Activité 5

Les participants à la formation explorent une affaire judiciaire hypothétique impliquant un biais d'IA et répondent à la question de savoir comment ils décideraient de l'affaire si elle était jugée dans leur juridiction.

Titre de l'affaire hypothétique : Smith c. AI Financial Services

Contexte : John Smith, un Afro-Américain, a intenté une action en justice contre AI Financial Services, une importante institution de prêt, alléguant un parti pris racial dans le système automatisé d'approbation des prêts de la société. Il affirme que le système d'IA a injustement rejeté sa demande de prêt hypothécaire, ce qui a entraîné une détresse financière et émotionnelle.

Détails de l'affaire :

1. **Argumentation du demandeur :** John Smith soutient que le système d'IA d'approbation des prêts utilisé par AI Financial Services refuse de manière disproportionnée les prêts aux Afro-Américains, comme en témoignent les données montrant une disparité importante dans les taux d'approbation des prêts entre les groupes raciaux.
2. **Réponse du défendeur :** AI Financial Services défend son système d'IA, affirmant qu'il repose sur des critères financiers objectifs et ne considère pas la race comme un facteur dans les décisions de prêt. Ils font valoir que toute disparité dans les approbations de prêts est due à des différences dans les antécédents financiers et la solvabilité des demandeurs.

Examen du système d'IA : Au cours du procès, les deux parties font venir des témoins experts pour examiner le système d'IA :

1. **Expert du demandeur :** Un expert en éthique de l'IA témoigne que les données de formation du système d'IA comportaient des préjugés raciaux inhérents, qui ont influencé sa prise de décision. Il présente des preuves de cas similaires où les systèmes d'IA ont montré un comportement discriminatoire.
2. **Expert du défendeur :** L'expert en IA du défendeur soutient que le système d'IA a été conçu pour être neutre sur le plan racial et que tout biais dans les données d'entraînement n'était pas intentionnel. Il met en évidence les processus de test et de validation rigoureux que l'IA a subis avant le déploiement.

Rôle de la Cour : Le juge doit déterminer si la partialité de l'IA a joué un rôle dans le refus de prêt de John Smith et, le cas échéant, si AI Financial Services est responsable de discrimination. Voici les principaux facteurs à prendre en considération :

1. **Transparence du système d'IA :** Le tribunal évalue la transparence du processus décisionnel du système d'IA et si le défendeur a correctement divulgué son utilisation de l'IA aux demandeurs de prêt.

2. **Intention vs impact :** Le juge fait la distinction entre la discrimination intentionnelle et l'impact disparate résultant du biais d'IA, qui reste illégal, en vertu des lois anti-discrimination.
3. **Efforts d'atténuation :** Le tribunal examine si AI Financial Services a pris des mesures raisonnables pour atténuer les préjugés dans son système d'IA et s'il a rapidement résolu les problèmes identifiés.

Résultat : Le tribunal se prononce en faveur de John Smith, jugeant que le système d'IA utilisé par AI Financial Services présentait un biais qui a eu un impact disparate sur les demandeurs afro-américains. Le jugement comprend une compensation financière pour John Smith et une injonction obligeant AI Financial Services à examiner et à réviser ses algorithmes d'IA pour assurer la conformité avec les lois anti-discrimination.

Ce cas hypothétique met en évidence les problèmes juridiques complexes entourant le biais d'IA dans l'attribution de prêts, et l'importance de la transparence, de l'équité et de la responsabilité dans l'utilisation des systèmes d'IA, en particulier lorsqu'ils ont une incidence sur les droits des individus et l'accès aux services financiers.

6. Ressources

1. Alang N. (2017). Turns Out Algorithms are Racist, disponible sur : https://newrepublic.com/article/144644/turns-algorithms-racist/
2. Angwin J., Larson J., Mattu S., Kirchner L. (2016). Machine bias, disponible sur : https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
3. Buolamwini J., Gebru T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR, 81, 77–91, disponible sur : https://proceedings.mlr.press/v81/buolamwini18a.html
4. Commission Nationale de l'Informatique et des Libertés (2022). Asking the right questions before using an artificial intelligence system, disponible sur : https://www.cnil.fr/en/asking-right-questions-using-artificial-intelligence-system
5. Edwards L., Veale M. (2017). Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For, 16 Duke Law & Technology Review, 18, disponible sur : https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2972855
6. European Parliamentary Research Service (2019). A governance framework for algorithmic accountability and transparency, disponible sur : https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU(2019)624262_EN.pdf
7. Green B., Chen Y. (2019). Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments, Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, 90–99, disponible sur : https://doi.org/10.1145/3287560.3287563
8. Hart R. (2017). If you're not a white male, artificial intelligence's use in healthcare could be dangerous, disponible sur : https://qz.com/1023448/if-youre-not-a-white-male-artificial-intelligences-use-in-healthcare-could-be-dangerous
9. Kleinberg J. , Lakkaraju H., Leskovec J., Ludwig J., Mullainathan S. (2017). Human Decisions and Machine Predictions, disponible sur : https://www.cs.cornell.edu/home/kleinber/w23180.pdf
10. Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector, The Alan Turing Institute, disponible sur : https://doi.org/10.5281/zenodo.3240529
11. UTS Human Technology Institute report (2022). Outlining a Model Law for facial recognition, disponible sur : https://www.uts.edu.au/human-technology-institute/projects/facial-recognition-technology-towards-model-law
12. Whittlestone J., Nyrup R., Alexandrova A., Cave S. (2019). The role and limits of principles in AI ethics: Towards a focus on tensions, Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19), Association for Computing Machinery, 195–200, disponible sur : https://doi.org/10.1145/3306618.3314289

Module 4

Droits humains et IA

Le module 4 offre une analyse approfondie de certains droits humains impactés par l'IA, tels que (i) le droit à l'accès au tribunal, à un procès équitable et à une procédure régulière, (ii) à un recours effectif, (iii) le droit à la protection contre la discrimination, (iv) la liberté d'expression, (v) le droit à la vie privée et à la protection des données, et (vi) l'accès à l'information. Le module donne également un aperçu des principales approches de gouvernance de l'IA : fondées sur les risques et sur les droits humains.

Qu'allez-vous apprendre ?

Après avoir terminé ce module, les participants seront en mesure de :

- Comprendre et expliquer les cas de possibles violations des droits humains à travers l'utilisation d'ADM et d'IA : (i) le droit à l'accès au tribunal, à un procès équitable et à une procédure régulière, (ii) à un recours effectif, (iii) le droit à la protection contre la discrimination, (iv) la liberté d'expression, (v) le droit à la vie privée et à la protection des données, et (vi) l'accès à l'information.
- Comprendre les principales approches de gouvernance de l'IA : fondées sur les risques et sur les droits humains.

1. Introduction aux droits humains et à l'IA

Il existe une forte corrélation entre la démocratie, l'état de droit et les droits humains. Des institutions démocratiques robustes et responsables, des processus décisionnels inclusifs et transparents, et un pouvoir judiciaire indépendant et impartial qui respecte l'état de droit sont des conditions préalables au respect des droits humains.

Les droits humains sont les libertés et les droits fondamentaux détenus par toute personne, de la naissance à la mort. Les droits humains défendent la dignité inaliénable de chaque personne, indépendamment de sa race, de son origine ethnique, de son genre, de son âge, de son orientation sexuelle, de sa classe, de sa religion, de son niveau de handicap, de sa langue, de sa nationalité ou de tout autre attribut. Les gouvernements sont tenus de respecter, de défendre et de faire respecter les droits humains. Les individus ont droit à des recours juridiques qui prévoient la réparation de toute violation des droits humains.

La Charte internationale des droits[239] représente un corpus de droit international des droits humains qui comprend neuf grands traités relatifs aux droits humains, des instruments régionaux relatifs aux droits humains dans les Amériques, en Afrique et en Europe. Elle a été incorporée dans les constitutions et les lois nationales ainsi que la jurisprudence coutumière et jurisprudentielle.[240]

Les instruments intergouvernementaux non contraignants tels que les Principes directeurs des Nations Unies relatifs aux entreprises et aux droits de l'homme[241] ont également abordé la question de la responsabilité des parties prenantes du secteur privé dans le contexte des droits humains.

Les droits humains offrent un ensemble de normes de base mondiales fondées sur des principes tels que l'égalité, l'autonomie et la dignité humaine. Ces principes et le cadre juridique qui les accompagne imposent aux nations des obligations juridiquement contraignantes de respecter, de défendre et de faire respecter les droits humains.

Le droit international relatif aux droits humains exige que les États-nations prévoient un recours effectif lorsqu'un individu subit une violation des droits humains. Les recours efficaces comprennent les recours judiciaires et administratifs, tels que l'ordonnance d'indemnisation ou d'excuses, et les mesures préventives qui peuvent inclure des changements à la loi, à la politique et à la pratique. Les obligations en matière de droits humains exigent également des États qu'ils mettent en place des mécanismes efficaces pour empêcher que les droits humains ne soient violés.[242]

239 Comprised of the Universal Declaration of Human Rights, the International Covenant on Civil and Political Rights and the International Covenant on Economic, Social and Cultural Rights.

240 Baluarte D. C., De Vos C. M. (2010). From Judgment to Justice: Implementing International and Regional Human Rights Decisions, Open Society Justice Initiative, Open Society Foundations: New York, disponible sur : https://www.justiceinitiative.org/uploads/62da1d98-699f-407e-86ac-75294725a539/from-judgment-to-justice-20101122.pdf

241 Conseil des droits de l'homme des Nations Unies (2011). Guiding Principles on Business and Human Rights, disponible sur : https://www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_en.pdf

242 Le paragraphe 3 de l'article 2 du Pacte international relatif aux droits civils et politiques exige de chaque État partie qu'il veille à ce qu'une personne dont les droits énoncés dans le Pacte ont été violés dispose d'un recours effectif et à ce que ce recours soit appliqué. Voir aussi : Comité des droits de l'homme de l'ONU (2004). The Nature of the General Legal Obligation Imposed on States Parties to the Covenant, disponible sur : https://www.refworld.org/docid/478b26ae2.html

Le cadre du droit international relatif aux droits humains est un moyen établi d'assurer la protection des droits en général et dans l'environnement numérique, notamment les droits à l'égalité et à la non-discrimination. Sa nature d'ensemble de normes exploitables se prête particulièrement bien aux technologies qui transcendent les frontières nationales, telles que l'IA. Une approche fondée sur les droits humains fournit des orientations normatives aux développeurs d'IA pour défendre la dignité humaine, quelle que soit leur juridiction.

Le droit relatif aux droits humains peut éclairer l'élaboration de garanties techniques et politiques dans le déploiement de l'IA. Dans cet esprit, en 2019, le Conseil des droits de l'homme (CDH) a adopté la première résolution (41/11) sur « Les technologies numériques nouvelles et émergentes et les droits de l'homme ».[243] La résolution reconnaît la nécessité de mieux tenir compte de l'ensemble des implications des nouvelles technologies en matière de droits humains, pour rester pertinent à l'ère numérique.

En 2021, le Conseil a adopté la résolution 47/23, soulignant l'importance d'une approche fondée sur les droits humains pour développer et déployer des technologies numériques innovantes. La résolution note que les nouvelles technologies ont le potentiel d'offrir de multiples opportunités pour faire progresser les droits humains en contribuant positivement à la construction d'institutions démocratiques et à la résilience de la société civile, ainsi qu'à la réalisation des objectifs de développement durable (ODD). Les défenseurs des droits humains et les développeurs de technologies, ainsi que les gouvernements, doivent rester alertes pour s'attaquer aux problèmes de droits humains posés par l'IA, en utilisant des protections et des instruments basés sur les normes et cadres existants en matière de droits humains.[244]

Pour que l'IA profite au bien public, sa conception et sa mise en œuvre doivent, au minimum, éviter de porter atteinte aux valeurs humaines fondamentales garanties par le droit international relatif aux droits humains, qui fournit un cadre solide pour la protection de ces valeurs. Si des garanties adéquates sont mises en œuvre, l'IA pourrait également être un facteur clé dans l'amélioration et la promotion des droits humains.

Comment l'IA peut-elle aider à la protection et à l'application des droits humains ?

Les systèmes d'IA ont de nombreuses applications qui peuvent aider à l'application des droits humains. Par exemple, les systèmes d'IA sont utilisés pour analyser les modèles de pénurie alimentaire, afin de lutter contre la faim, d'améliorer le diagnostic et le traitement médicaux, ou de rendre les services de santé plus accessibles.

Le module 2 a donné un aperçu de la façon dont l'IA peut aider les opérateurs judiciaires, grâce à la découverte électronique et à l'examen des documents, à l'analyse prédictive et au soutien ADM, aux outils d'évaluation des risques,

243 Conseil des droits de l'homme des Nations Unies (2019). New and emerging digital technologies and human rights, disponible sur : https://digitallibrary.un.org/record/3834165

244 DiPLO (2022). Promoting and Protecting Human Rights in the Digital Era, disponible sur : https://www.diplomacy.edu/event/promoting-and-protecting-human-rights-in-the-digital-era/

à la résolution des litiges, à l'IA générative, à la reconnaissance et à l'analyse linguistiques, et à la gestion numérique des dossiers et des affaires. Le pouvoir judiciaire, en tant qu'institution publique, est tenu à un niveau plus élevé en matière de comportement des opérateurs judiciaires, et des juges en particulier, envers les individus et la société. Cela s'est reflété dans les principes de l'état de droit tels que la justification, la proportionnalité et l'égalité. D'une part, l'IA peut accroître l'efficacité des opérateurs judiciaires, d'autre part, elle peut également éroder la légitimité procédurale et la confiance dans les institutions démocratiques et l'autorité de la loi.

Sans garde-fous appropriés, l'IA pourrait également empiéter sur les droits humains

Par exemple, un biais non détecté peut exister dans les algorithmes de ML qui prédisent la récidive. Ou bien le déploiement de l'IA pourrait être utilisé pour limiter la liberté d'expression des personnes ou leur capacité à s'engager dans des activités politiques ou à identifier des dissidents politiques.[245] L'IA pourrait également porter atteinte aux droits humains, dans des situations d'utilisation de données de formation de mauvaise qualité, de conception de système ou d'interactions complexes entre le système d'IA et son environnement. Par exemple, l'exacerbation algorithmique du discours de haine ou l'incitation à la violence en ligne. On peut également penser à l'amplification de la désinformation et de l'anti-information, qui pourrait avoir un impact sur le droit de participer aux affaires politiques et publiques, en particulier pendant les élections. L'ampleur et l'impact probables du préjudice seront liés à l'ampleur et à l'impact potentiel des décisions prises par un système d'IA spécifique. Dans le même temps, il est important de noter que l'IA peut être utilisée pour identifier les discours de haine et aider à supprimer les contenus liés à la promotion du terrorisme.

Les applications de l'IA peuvent affecter directement l'égalité d'accès aux droits fondamentaux, y compris le droit à la vie privée et à la protection des informations personnelles, le droit à l'accès à la justice et le droit à un procès équitable, en particulier en ce qui concerne la présomption d'innocence et la charge de la preuve, le droit à l'emploi, à l'éducation, au logement et à la santé, ainsi que le droit aux services publics et à la protection sociale. Si elles ne sont pas accompagnées de garanties adéquates contre les préjugés, les technologies de l'IA pourraient contribuer à refuser l'accès aux droits qui affectent de manière disproportionnée les femmes, les minorités et ceux qui sont déjà les plus vulnérables et les plus marginalisés.[246]

Par exemple, l'utilisation de systèmes de reconnaissance biométrique ou faciale dans les espaces publics pourrait permettre une surveillance de masse empiétant sur les droits humains.[247] Selon le rapport de l'AI Now Institute « Regulating Biometrics » (2020)[248], les technologies de reconnaissance faciale ne remplace pas adéquatement les empreintes digitales. Les technologies de reconnaissance faciale montrent des résultats médiocres et des taux d'erreur

245 Assemblée générale des Nations Unies (2018). Promotion and protection of the right to freedom of opinion and expression. Note by the Secretary-General, disponible sur : https://documents-dds-ny.un.org/doc/UNDOC/GEN/N18/270/42/PDF/N1827042.pdf?OpenElement

246 Conseil de l'Europe (2019). Preventing discrimination caused by the use of artificial intelligence, disponible sur : https://pace.coe.int/en/files/28809.

247 Human Rights Watch (2020). Argentina: Child Suspects' Private Data Published Online, disponible sur : https://www.hrw.org/news/2020/10/09/argentina-child-suspects-private-data-published-online

248 Kak A. (2020). Regulating Biometrics. Global Approaches and Urgent Questions, disponible sur : https://ainowinstitute.org/publication/regulating-biometrics-global-approaches-and-open-questions

élevés pour « les femmes noires, les minorités de genre, les jeunes et les personnes âgées, les personnes porteuses de handicaps et les travailleurs manuels ».[249]

Souvent, le déploiement de l'IA par les organismes d'application de la loi peut empiéter sur le droit à une procédure régulière et à l'égalité des droits en matière de protection. Par exemple, si le système d'IA est utilisé pour les tests ADN qui impliquent le traitement de données de santé sensibles et les évaluations des risques de la justice pénale qui pourraient être biaisées en faveur de certaines populations en fonction du genre, de la race, de l'origine ethnique, etc.

Rappel !

Comme nous l'avons vu, les outils de police prédictive ou de reconnaissance faciale ne peuvent être une prédétermination de culpabilité ou une preuve suffisante pour réfuter la présomption d'innocence. Une prédiction statistique ne peut pas être une cause d'arrestation ou, en vertu de la common law, un soupçon raisonnable, ou une cause probable plus élevée, et est loin d'être une preuve *prima facie*, et encore moins une preuve inculpatoire. Sa valeur de renseignement ne peut excéder celle donnée aux informations policières ou aux informations de renseignement et n'aurait donc aucune valeur probante. L'utiliser comme seule source violerait le principe de la présomption d'innocence.

L'utilisation de l'IA doit être orientée vers le principe de bienfaisance, l'amélioration et le progrès de l'humanité. Ainsi, le développement et l'utilisation des systèmes d'IA doivent être orientés vers le bénéfice et le bien-être de la société et de la civilisation humaine, pour l'amélioration des conditions de vie, de la santé, du travail, du développement des capacités physiques et mentales.

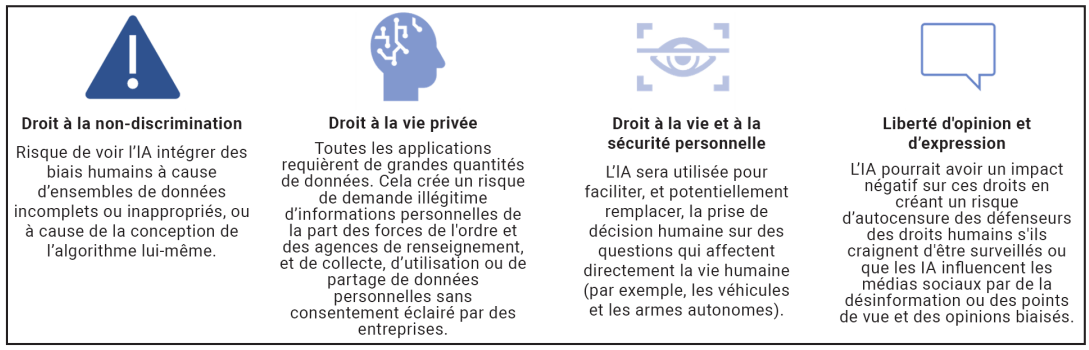
Bien que l'on puisse s'attendre à ce que la structure de base et le cadre institutionnel de la protection des droits humains, bien établis et universellement reconnus, développent des réponses efficaces à de nombreuses menaces et défis causés par la puissance croissante de l'automatisation numérique et de l'intelligence artificielle, il existe plusieurs raisons pour lesquelles les mécanismes existants d'application des droits humains peuvent nécessiter une revitalisation, s'ils veulent assurer une protection efficace. Premièrement, de nombreux droits sont difficiles à faire valoir dans la pratique, en raison de l'opacité de nombreux systèmes sociotechniques dans lesquels ces technologies sont intégrées. Deuxièmement, notre compréhension de la portée et du contenu des droits existants a été développée à une époque pré-réseautée. Ainsi conçus, ces droits pourraient ne pas fournir une protection complète contre toute la gamme de menaces et de risques auxquels ces technologies peuvent donner lieu, en particulier en ce qui concerne la discrimination et les tentatives illégitimes de tromper et de manipuler les individus à l'aide de « technologies persuasives »[250].

249 *Ibid.*

250 La technologie persuasive est une « technologie créée spécifiquement pour changer les opinions, les attitudes ou les comportements de ses utilisateurs afin d'atteindre ses objectifs », voir : Centre for Humane technology (2021). Persuasive Technology. How does technology use design to influence my behavior?, disponible sur : https://assets.website-files.com/5f0e1294f002b15080e1f2ff/612f8e3e010ff2e211c92019_2%20-%20Persuasive%20Technology%20Issue%20Guide.pdf

La figure 12 ci-dessous donne un aperçu de certains droits humains couverts par ce manuel de formation qui pourraient être impactés par le déploiement de l'IA en général.

Figure 12. Sélectionner les droits humains impactés par l'IA



Source : OCDE, https://www.oecd-ilibrary.org/sites/ba682899-en/

Activité:

Les participants à la formation regardent la vidéo et discutent de la façon dont l'IA peut avoir un impact sur les droits humains.

Source : https://youtu.be/TbBMeFrR7H8

251 Voir : https://www.oecd-ilibrary.org/sites/969ff07f-en/index.html?itemId=/content/component/969ff07f-en

Avantages de l'approche des droits humains pour le développement et le déploiement de l'IA

Les mécanismes institutionnels du droit relatifs aux droits humains fournissent l'orientation et la base pour assurer le développement et l'utilisation éthiques et centrés sur l'homme de l'IA dans la société. Les opérateurs judiciaires peuvent recommander une diligence raisonnable en matière de droits humains, telle que des évaluations d'impact sur les droits de l'homme (HRIA), pour mesurer et évaluer les risques posés par le déploiement de l'IA sur les droits humains. Plus le risque pour les droits humains est élevé, plus l'IA peut être jugée impropre à l'utilisation.

Les évaluations de l'impact sur les droits humains peuvent aider à identifier les groupes ou les communautés vulnérables ou à risque en ce qui concerne l'IA. Certaines personnes ou communautés peuvent être sous-représentées en raison, par exemple, de l'utilisation limitée des smartphones et de l'absence de leurs données dans les ensembles de données utilisés pour former les systèmes d'IA. Une approche fondée sur les droits humains peut offrir un recours à ceux dont les droits sont violés. Des exemples de recours comprennent la cessation d'activité, l'élaboration de nouveaux processus ou politiques, des excuses ou une compensation monétaire.

Il existe cinq avantages clés à tirer parti des cadres des droits humains dans le contexte de l'IA.[251]

- Au fil du temps, une vaste infrastructure internationale, régionale et nationale des droits humains a été développée, et il existe des institutions établies qui peuvent aider à l'application des droits humains dans le contexte de l'intelligence artificielle. Cette infrastructure comprend des organisations intergouvernementales, des tribunaux, des ONG, des établissements universitaires et d'autres institutions et communautés où les droits humains peuvent être revendiqués et où des réparations peuvent être demandées.
- Un ensemble complet de lois nationales, régionales et internationales a rendu opérationnelle l'application des droits humains dans le domaine numérique.
- Les droits humains constituent un langage universel quant aux questions qui transcendent les frontières nationales, telles que l'IA. L'infrastructure des droits humains peut aider à atteindre et à inclure un plus large éventail de parties prenantes.
- Les droits humains jouissent d'une légitimité et d'un soutien généralisés dans le monde. La simple perception qu'un acteur puisse violer les droits humains peut revêtir une grande importance, dans la mesure où cela peut entraîner des répercussions en termes de réputation.
- De nombreux États sont dotés d'une certaine forme de cadre des droits humains, même s'ils n'ont pas de cadre de protection des données - par conséquent, utiliser le cadre des droits humains comme base rendrait le processus plus inclusif.

Un défi lié à l'approche des droits humains pour le développement et le déploiement de l'IA est le fait que leur application est liée aux juridictions. Les demandeurs doivent souvent démontrer leur capacité juridique dans une juridiction particulière. Lorsque les problèmes impliquent de grandes entreprises internationales et des systèmes d'IA qui couvrent de nombreuses juridictions, ces approches peuvent ne pas être optimales.[253]

252 *Ibid.*

Tableau 5. Principaux instruments internationaux relatifs au droit à la vie privée en général, et dans l'environnement en ligne en particulier.

Traités	
1	Déclaration universelle des droits de l'homme[253]
2	Pacte international relatif aux droits civils et politiques[254]
3	Convention internationale sur l'élimination de toutes formes de discrimination raciale[255]
4	Lignes directrices régissant la protection de la vie privée et les flux transfrontières de données de caractère personnel[256]
5	Convention du Conseil de l'Europe pour la protection des personnes à l'égard du traitement automatisé des données à caractère personnel (Convention 108 / Convention 108+)[257]
Normes	
6	Les Directives des Nations Unies concernant les fichiers informatisés de données personnelles (ONU, 1990)[258]
7	Les normes internationales sur la vie privée et la protection des données (la résolution de Madrid)[259]
8	La Recommandation de l'OCDE sur la gestion des risques liés à la sécurité numérique pour la prospérité économique et sociale[260]
9	Lignes directrices régissant la protection de la vie privée et les flux transfrontières de données de caractère personnel[261]
10	Principes des Nations Unies sur la protection des données personnelles et de la vie privée (2018)[262]
11	La résolution de l'Assemblée générale des Nations Unies sur le droit à la vie privée à l'ère numérique de 2014[263]
Autres documents	
12	Rapport du Rapporteur spécial sur la promotion et la protection des droits de l'homme et des libertés fondamentales dans le cadre de la lutte contre le terrorisme[264]
13	Promotion et protection du droit à la liberté d'opinion et d'expression de l'ONU 2018[265]
14	La résolution de l'Assemblée générale des Nations Unies sur le droit à la vie privée à l'ère numérique (2020) a fait référence au « piratage et à l'utilisation illégale des technologies biométriques », comme « des actes hautement intrusifs qui violent le droit à la vie privée », interfèrent avec la liberté d'expression et d'opinion, la liberté de réunion et d'association pacifiques et la liberté de religion ou de conviction, et « peuvent contredire les principes d'une société démocratique, y compris lorsqu'ils sont entrepris de manière extraterritoriale ou à grande échelle ».[266]
15	Un rapport du Haut-Commissaire des Nations Unies aux droits de l'homme de 2021, intitulé « Le droit à la vie privée à l'ère numérique », a appelé à un moratoire sur l'utilisation des technologies de reconnaissance faciale dans les espaces publics, jusqu'à ce que les gouvernements puissent montrer qu'il n'y a pas de problèmes substantiels liés à l'exactitude ou aux impacts discriminatoires, et que ces technologies respectent des normes strictes en matière de confidentialité et de protection des données.[267]
16	UNSDG Data Privacy, Ethics and Protection: Guidance Note on Big Data for Achievement of the 2030 Agenda (2017)[268]
17	UN Compendium of data protection and privacy policies[269]

253 Voir : https://www.un.org/en/about-us/universal-declaration-of-human-rights
254 Bureau des droits de l'homme des Nations Unies (1976). Pacte international relatif aux droits civils et politiques, disponible sur : https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights
255 Bureau des droits de l'homme des Nations Unies (1965). Convention internationale sur l'élimination de toutes les formes de discrimination raciale, disponible sur : https://www.ohchr.org/en/instruments-mechanisms/instruments/international-convention-elimination-all-forms-racial
256 OCDE (2002). Lignes directrices de l'OCDE sur la protection de la vie privée et les flux transfrontiers de données à caractère personnel, disponibles sur : https://www.oecd-ilibrary.org/docserver/9789264196391-en.pdf?expires=1695655643&id=id&accname=ocid195767&checksum=923738DCA1AEE95B3D260E41902AC30D
257 Conseil de l'Europe (2018). Convention pour la protection des personnes à l'égard du traitement automatisé des données à caractère personnel (Convention 108+), disponible sur : https://www.coe.int/en/web/data-protection/convention108-and-protocol
258 Joinet L. (1988). Guidelines for the Regulation of Computerized Personal Data Files: final report, disponible sur : https://digitallibrary.un.org/record/43365?ln=en
259 Voir : https://www.dataguidance.com/opinion/international-madrid-resolution
260 OCDE (2015). Digital Security Risk Management for Economic and Social Prosperity: OECD Recommendation and Companion Document, disponible sur : https://www.oecd.org/publications/digital-security-risk-management-for-economic-and-social-prosperity-9789264245471-en.htm
261 Voir : https://www.oecd.org/sti/ieconomy/oecdguidelinesontheprotectionofprivacyandtransborderflowsofpersonaldata.htm
262 Voir : https://unsceb.org/privacy-principles
263 Conseil des droits de l'homme des Nations Unies (2014). The right to privacy in the digital age: report of the Office of the United Nations High Commissioner for Human Rights, disponible sur : https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Fwww.ohchr.org%2Fsites%2Fdefault%2Ffiles%2FDocuments%2FIssues%2FDigitalAge%2FA-HRC-27-37_en.doc&wdOrigin=BROWSELINK
264 Voir : https://www.ohchr.org/en/special-procedures/sr-terrorism
265 Assemblée générale des Nations Unies (2018). Promotion and protection of the right to freedom of opinion and expression, disponible sur : https://documents-dds-ny.un.org/doc/UNDOC/GEN/N18/270/42/PDF/N1827042.pdf?OpenElement
266 Assemblée générale des Nations Unies (2020). The right to privacy in the digital age, disponible sur : https://documents-dds-ny.un.org/doc/UNDOC/GEN/N20/371/75/PDF/N2037175.pdf?OpenElement
267 Conseil des droits de l'homme des Nations Unies (2021). The right to privacy in the digital age. Report of the United Nations High Commissioner for Human Rights, disponible sur : https://www.ohchr.org/en/HRBodies/HRC/RegularSessions/Session48/Documents/A_HRC_48_31_AdvanceEditedVersion.docx
268 Voir : https://unsdg.un.org/resources/data-privacy-ethics-and-protection-guidance-note-big-data-achievement-2030-agenda
269 PNUD (2021). COMPENDIUM OF DATA PROTECTION AND PRIVACY POLICIES AND OTHER RELATED GUIDANCE WITHIN THE UNITED NATIONS ORGANIZATION AND OTHER SELECTED BODIES OF THE INTERNATIONAL COMMUNIT, disponible sur : https://unstats.un.org/legal-identity-agenda/documents/Paper/data_protecton_%20and_privacy.pdff

Droit d'accès au tribunal, à un procès équitable et à une procédure régulière

« Tous sont égaux devant les tribunaux et les cours de justice. ... toute personne a droit à ce que sa cause soit entendue équitablement et publiquement par un tribunal compétent, indépendant et impartial, établi par la loi, qui décidera soit du bien-fondé de toute accusation en matière pénale dirigée contre elle, soit des contestations sur ses droits et obligations de caractère civil [...] Toute personne accusée d'une infraction pénale est présumée innocente jusqu'à ce que sa culpabilité ait été légalement établie. »

– Article 14 du PIDCP

En ce qui concerne l'application de la loi et le système juridique, le potentiel de l'IA de renforcer ou d'amplifier les préjugés existants est une préoccupation majeure. Les droits à la liberté, à la sécurité et à un procès équitable peuvent être violés lorsque la liberté physique ou la sécurité personnelle d'un individu est en jeu, par exemple avec la police prédictive, l'évaluation des risques de récidive et la détermination de la peine. Comme déjà discuté, les systèmes d'IA « à boîte noire » empêchent les professionnels du droit tels que les juges, les avocats et les procureurs de comprendre la raison d'être des résultats du système, ce qui complique la justification et l'appel de la décision.[270]

L'IA et la prise de décision automatisée (ADM) ont un impact substantiel sur la vie des gens, et elles peuvent souvent restreindre le droit de participer, de contester ou de remettre en cause de quelque manière que ce soit le résultat de la décision ou ses contributions. Souvent, les systèmes d'IA, en raison de leur nature de « boîte noire », sont incapables de produire une explication intelligible et compréhensible par l'homme de leurs décisions. Ces systèmes peuvent également comporter des biais intégrés qui limitent l'accès des données invisibles et des groupes marginalisés aux tribunaux et à la justice.

Des outils d'évaluation des risques criminels, par exemple, sont proposés comme instruments pour aider les juges dans les décisions de condamnation. Bien que les autorités attribuent un niveau de culpabilité potentielle en classant une personne comme présentant un risque élevé ou faible de récidive, cela pourrait être contraire au droit à un jury impartial et à la présomption d'innocence. Les logiciels de police prédictive peuvent également refléter les préjugés sociétaux et présenter le risque d'utiliser des données historiques

270 CAHAI Secretariat (2020). Towards Regulation of AI Systems. Global perspectives on the development of a legal framework on Artificial Intelligence (AI) systems based on the Council of Europe's standards on human rights, democracy and the rule of law, Étude du Conseil de l'Europe DGI/2020/16, disponible sur : https://rm.coe.int/prems-107320-gbr-2018-compli-cahai-couv-texte-a4-bat-web/1680a0c17a.

pour introduire des préjugés et attribuer faussement la culpabilité.[271] Il existe plusieurs cas documentés où l'utilisation d'algorithmes d'IA dans la police prédictive, l'évaluation des risques et la détermination de la peine a conduit à des résultats sous-optimaux dans le système de justice pénale. Dans de nombreux cas, l'utilisation de l'IA pour la notation des risques des accusés et les efforts de police prédictive sont annoncés comme des tentatives bien intentionnées d'éliminer la partialité humaine potentielle des juges dans leurs décisions de condamnation et de mise en liberté sous caution, tout en allouant des ressources policières limitées pour prévenir la criminalité. Cependant, ces systèmes d'IA, s'ils ne sont pas conçus avec des préoccupations éthiques à l'esprit, peuvent finir par exacerber le biais même qu'ils cherchent à atténuer, en incorporant directement des facteurs biaisés ou en utilisant des procurations pour les biais dans leurs recommandations.[272] Cela peut entraîner de graves conséquences, y compris la perpétuation de la discrimination contre certains groupes.

Par conséquent, lorsque les systèmes d'IA sont biaisés et opaques, ils soulèvent des préoccupations concernant les normes de procès équitable, telles que la présomption d'innocence, le droit d'être informé rapidement de l'origine et de la nature d'une accusation, le droit à un procès équitable et la capacité de se défendre en personne. L'opacité de la prise de décision par les systèmes d'IA soulève également des préoccupations concernant la privation arbitraire de liberté et le droit de ne pas être puni sans loi.[273]

> « L'utilisation d'outils d'évaluation des risques pour prendre des décisions équitables sur la liberté humaine nécessiterait de résoudre de profonds défis éthiques, techniques et statistiques, notamment en veillant à ce que les outils soient conçus et construits pour atténuer les préjugés à la fois au niveau du modèle et des couches de données, et que des protocoles appropriés soient en place pour promouvoir la transparence et la responsabilité. Les outils actuellement disponibles et à l'étude pour une utilisation généralisée souffrent de plusieurs de ces défaillances ».[274]

271 AccessNow (2018). AI and human rights, disponible sur : https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf

272 Par exemple, selon les dossiers publics, la police de la Nouvelle-Orléans a utilisé le logiciel créé par Palantir pour les enquêtes criminelles d'une manière qui allait au-delà de la portée initiale prévue du logiciel. À la suite d'une série de rapports d'enquête et de réactions négatives importantes du public, la ville a mis fin à son contrat de six ans avec Palantir, en mars 2018.

273 CAHAI (2020). The Impact of Artificial Intelligence on Human Rights, Democracy and the Rule of Law, disponible sur : https://rm.coe.int/cahai-2020-06-fin-c-muller-the-impact-of-ai-on-human-rights-democracy-/16809ed6da.

274 Partnership on AI, Report on Algorithmic Risk Assessment Tools in the US Criminal Justice System, disponible sur : https://partnershiponai.org/wp-content/uploads/2021/08/Report-on-Algorithmic-Risk-Assessment-Tools.pdf

Activité :

Les participants à la formation lisent une jurisprudence sélective qui traite des boîtes noires algorithmiques dans les ADM et les systèmes d'IA, et discutent de la façon dont l'IA et les progrès technologiques affectent les droits humains à l'accès au tribunal, à un procès équitable et à une procédure régulière.[275]

État c. Loomis aux États-Unis

Dans l'affaire État c. Loomis, la Cour suprême du Wisconsin a déterminé que l'utilisation de l'algorithme COMPAS, un outil exclusif d'évaluation des risques, lors de la détermination de la peine, ne violait pas les droits du défendeur à une procédure régulière. COMPAS a été initialement développé pour aider les commissions des libérations conditionnelles à déterminer le risque de récidive. Cependant, le résultat de COMPAS - une note d'évaluation des risques - a été utilisé à la fois par l'État et par le tribunal de première instance, lors de la détermination de la peine. Northpointe, Inc., la société qui a créé COMPAS, a refusé de révéler sa méthodologie au tribunal ou au prisonnier. Le tribunal de condamnation a condamné le défendeur à une peine de six ans au lieu de la libération conditionnelle, car l'algorithme a déterminé qu'il avait une probabilité significative de récidive.[276]

Bien que la Cour ait confirmé la validité de COMPAS, de nombreuses limitations ont été imposées à son application. L'algorithme ne pouvait pas être utilisé pour évaluer si un criminel purgerait une peine de prison ou pour estimer la durée de sa peine. Tous les rapports d'enquête pré-sentencielle, y compris la note, devaient inclure un avertissement élaboré en cinq parties sur les limites de l'algorithme. Son utilisation nécessitait également une justification distincte de la sentence. La Cour suprême a refusé de prendre l'affaire en appel du défendeur.[277]

Il reste à savoir s'il est approprié que le tribunal permette à un algorithme, dans lequel les opérateurs judiciaires ont une visibilité limitée, de jouer un rôle, même mineur, dans la privation de liberté d'une personne. La décision de la Cour suprême du Wisconsin et les documents d'appel révèlent des erreurs fondamentales concernant le fonctionnement potentiel d'un algorithme comme COMPAS et les protections nécessaires pour le rendre utile dans la détermination de la peine. Ces malentendus offrent un aperçu d'un cadre plus prometteur, qui permettrait aux algorithmes de renforcer le système judiciaire sans poser de problèmes juridiques, technologiques ou éthiques.[278]

People c. Alvin Davis aux États-Unis

Dans cette affaire, deux témoins ont affirmé avoir vu un homme noir dans la cinquantaine sur la propriété, la veille du meurtre d'une femme plus âgée qui y avait été agressée sexuellement et assassinée. Dans les mois qui ont précédé le meurtre, des dizaines de personnes, dont M. Davis et une autre personne, avaient visité la résidence de la victime. M. Davis est un Afro-Américain atteint de la maladie de Parkinson, âgé de 70 ans au moment du meurtre. Une deuxième personne qui correspondait à la description des témoins avait des antécédents d'infractions sexuelles.

De nombreux sites et objets sur la scène du crime ont été échantillonnés pour l'ADN. Bon nombre de ces objets, y compris une canne qui aurait été utilisée pour agresser sexuellement la victime, ne contenaient pas l'ADN de M. Davis. Bien que STRMix, un logiciel utilisé pour l'analyse de l'ADN, ait réussi à faire correspondre M. Davis à l'échantillon d'ADN prélevé sur un lacet probablement utilisé pour attacher la victime, le logiciel d'ADN traditionnel n'a pas été en mesure de le faire. L'accusation a beaucoup insisté sur STRMix, devant le jury. Atteint de la maladie de Parkinson, M. Davis est confiné dans un fauteuil roulant. Le premier procès contre lui s'est terminé par une suspension de jury. Après un deuxième procès, il a été reconnu coupable et condamné à une peine d'emprisonnement à perpétuité sans possibilité de libération conditionnelle.

Dans l'affaire People c. Alvin Davis en Californie, l'Electronic Frontier Foundation (EFF) est intervenue en faveur de la capacité de M. Davis à consulter le code source de STRMix, le programme d'ADN médico-légal utilisé pendant son procès. Dans plusieurs affaires, dont la plus récente est celle-ci, L'EFF a affirmé qu'un défendeur a le droit de réviser le logiciel d'analyse d'ADN. Dans deux de ces affaires, États-Unis c. Ellis et State c. Pickett, les tribunaux ont convenu

275 Grimm P., Grossman M. R., Cormack G. V, Artificial Intelligence as Evidence, Artificial Intelligence as Evidence, 19 Nw. J. Tech. & Intell. Prop. 9, disponible sur : https://scholarlycommons.law.northwestern.edu/njtip/vol19/iss1/2

276 Israni E. (2017). Algorithmic Due Process: Mistaken Accountability and Attribution in State v. Loomis, disponible sur : https://jolt.law.harvard.edu/digest/algorithmic-due-process-mistaken-accountability-and-attribution-in-state-v-loomis-1.

277 *Ibid.*

278 *Ibid.*

avec l'EFF que les défendeurs avaient droit au code source TrueAllele, l'un des principaux rivaux de STRMix.[279]

Pour s'assurer que le résultat du logiciel de mise en correspondance de l'ADN utilisé contre eux est exact, les accusés doivent être autorisés à examiner le fonctionnement du logiciel. Comme il peut y avoir des failles de codage, l'accès au code source ne peut remplacer un témoignage sur le fonctionnement du logiciel. Cela est particulièrement vrai pour les logiciels d'ADN médico-légaux les plus récents, tel que STRMix et TrueAllele, en proie à des problèmes d'exactitude et de fiabilité.[280] En réalité, STRMix a déjà été examiné, et des défauts de programmation susceptibles d'avoir entraîné des résultats erronés dans 60 cas dans le Queensland, en Australie, y ont été constatés.[281]

État du New Jersey c. Pickett ; États-Unis c. Ellis

Dans les affaires New Jersey c. Pickett[282] et United States c. Ellis[283], la défense a demandé l'accès au code source du logiciel d'une entreprise (TrueAllele). TrueAllele est utilisé pour effectuer une étude de génotypage probabiliste sur des échantillons d'ADN. Dans les deux cas, les tribunaux ont conclu que l'accès au code devrait être accordé à la défense sous réserve d'une ordonnance de protection. Dans l'affaire Picket, le tribunal a souligné que « rien de moins qu'un accès complet contrevient aux principes fondamentaux d'équité, ce qui compromet indubitablement le droit d'un défendeur à présenter une défense complète ». Bien que ces outils soient différents des technologies d'IA basées sur des données, les décisions montrant que le code source du logiciel est accessible dans les procédures pénales créent un précédent encourageant pour d'autres technologies avancées revendiquant la protection des secrets commerciaux.[284]

État du New Jersey c. Francisco Arteaga aux États-Unis

Le cas New Jersey c. Arteaga est un exemple d'affaire qui souligne l'importance de la découvrabilité des algorithmes d'IA et de leurs entrées de données dans les affaires judiciaires. En 2019, une entreprise de West New York, dans le New Jersey, a été cambriolée sous la menace d'une arme, et Francisco Arteaga a ensuite été identifié comme suspect et inculpé de vol qualifié. Avant l'identification de M. Arteaga, l'enquête de la police du New Jersey a révélé que les témoins du crime étaient incapables d'identifier le délinquant, et la recherche de reconnaissance faciale menée par le Centre régional de renseignement des opérations du New Jersey n'a donné aucun résultat.

Après cette tentative infructueuse d'identification des suspects, les services de la police de New York ont effectué une recherche par reconnaissance faciale à l'aide de photos tronquées issues de caméras de surveillance dans la rue. M. Arteaga figurait parmi les résultats de recherche, et l'analyste de la reconnaissance faciale du NYPD l'a identifié comme la « possible correspondance ». La police a ensuite intégré la photo de M. Arteaga à plusieurs autres, où deux témoins l'ont finalement identifié, malgré les processus défectueux utilisés pour mener les séances d'identification. En dépit de l'importance de la correspondance basée sur l'algorithme pour cette affaire, la défense n'a reçu aucune information concernant l'algorithme qui l'a généré. M. Arteaga a réclamé la divulgation de la technologie de reconnaissance faciale utilisée par la police de New York, de la photo originale et de toutes les modifications apportées par la police de New York avant d'effectuer une recherche, ainsi que des informations concernant l'analyste qui a effectué la recherche qui a permis son identification. Le tribunal de district du New Jersey a rejeté sa requête de divulgation.

EPIC, en collaboration avec l'Electronic Frontier Foundation (EFF) et la National Association of Criminal Defense Lawyers (NACDL), a déposé une note informant le tribunal de la façon dont les erreurs se produisent dans les systèmes de reconnaissance faciale, du potentiel de biais dans ces systèmes. Ils ont fait valoir que la divulgation constitue le dernier recours pour corriger ces erreurs. La note décrivait la séquence des procédures nécessaires pour effectuer une recherche par reconnaissance faciale, qui impliquent toutes des décisions humaines pouvant ajouter des inexactitudes et augmenter la probabilité d'erreur dans l'identification. La note soutient que l'examen humain qui suit une perquisition ne peut être considéré comme un recours aux erreurs algorithmiques.[285] L'affaire est maintenant devant le juge de la Cour d'appel.[286]

279 Zhao H. (2021). EFF tells California Court that Forensic Software Source Code Must Be Disclosed to the Defendant, disponible sur : https://www.eff.org/deeplinks/2021/05/eff-tells-california-court-forensic-software-source-code-must-be-disclosed

280 Zhao H. (2021). How Your DNA—or Someone Else's—Can Send You to Jail, disponible sur : https://www.eff.org/deeplinks/2021/05/how-your-dna-or-someone-elses-can-send-you-jail.

281 Murray D. (2015). Queensland authorities confirm 'miscode' affects DNA evidence in criminal cases, disponible sur : http://www.couriermail.com.au/news/queensland/queensland-authorities-confirm-miscode-affects-dna-evidence-in-criminal-cases/news-story/833c580d3f1c59039efd1a2ef55af92b

282 State of New Jersey v Corey Pickett, disponible sur : https://law.justia.com/cases/new-jersey/appellate-division-published/2021/a4207-19.html.

283 EFF, United States v. Ellis, disponible sur : https://www.eff.org/cases/united-states-v-ellis

284 NACDL'S task force on predictive policing (2021). Garbage in, gospel out. How Data-Driven Policing Technologies Entrench Historic Racism and 'Tech-wash' Bias in the Criminal Legal System, disponible sur : https://www.nacdl.org/Document/GarbageInGospelOutDataDrivenPolicingTechnologies

285 EPIC Amicus Brief, New Jersey v. Arteaga, disponible sur : https://epic.org/documents/new-jersey-v-arteaga/

286 Murphy R. (2022). Lawyers and digital rights advocates want the facial recognition process exposed in court, disponible sur : https://localtoday.news/nj/lawyers-and-digital-rights-advocates-want-the-facial-recognition-process-exposed-in-court-52064.html

L'une des plus grandes menaces générées par les systèmes d'IA dans l'administration de la justice est ce que l'on appelle le biais d'automatisation, qui est la tendance des humains à considérer la solution proposée par l'intelligence artificielle comme correcte par défaut, entraînant une validation automatique de la part des humains. Il s'agit d'un risque particulièrement aberrant dans l'administration de la justice, car il peut conduire à une confiance aveugle dans les décisions proposées par le système, à considérer que la seule jurisprudence existante est celle proposée par la machine, ou à considérer qu'une évaluation de la possibilité de récidive est correcte. Au fil du temps, cela conduirait à un changement dans le raisonnement des décisions visant à justifier les raisons pour lesquelles le résultat donné par le système n'est pas suivi. Cette possibilité est aggravée par la charge de travail disproportionnée de la plupart de nos tribunaux, qui mène à un système de travail dans lequel la quantité et la rapidité priment sur la qualité. Ainsi, le fait pour le juge de s'écarter de toute décision, qu'elle soit assistée ou automatisée, ne saurait entraîner aucune forme de représailles, de sanction, d'inspection ou de régime disciplinaire. Si la supervision et le contrôle humains prévalent, le contrôle doit être efficace. Questions clés que nous devrions poser à cet égard :

(i) Comment le réglement d'une affaire par un système d'IA, au lieu d'un juge, affecte-t-il le droit effectif à l'accès au tribunal, à un procès équitable et à une procédure régulière ?

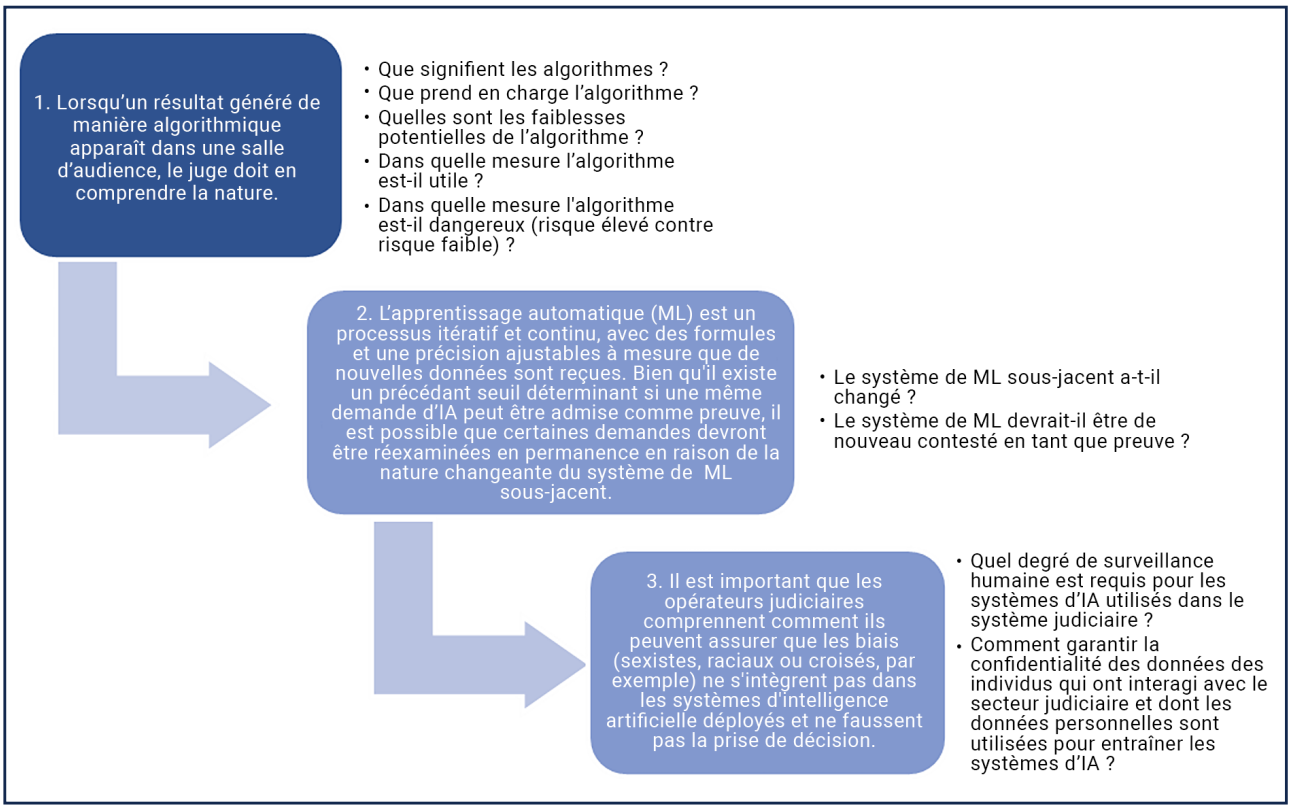
(ii) Comment s'articulera la motivation des décisions judiciaires ? Les citoyens ont le droit de connaître la motivation des jugements et les juges ont le devoir de les motiver. Dans le cas d'une boîte noire, le raisonnement logique de la conclusion n'est ni transparent ni accessible.

(iii) Dans le cas de l'existence d'une proposition de projet de décision ou de l'application de la jurisprudence par un système d'IA qui alimente une décision/un jugement rendu en dernier ressort par un juge humain, les parties ont-elles le droit de connaître le raisonnement du système d'IA, et cet argument pourrait-il être utilisé comme motif d'appel ou comme argument à l'appui de l'appel ? Le droit à la transparence de l'algorithme et les délibérations secrètes du tribunal sont deux questions distinctes qui ne doivent pas être confondues.

Les systèmes d'IA doivent être considérés comme des outils auxiliaires et de soutien, sans se voir attribuer de valeur décisive ou être surestimés, sans oublier la motivation judiciaire nécessaire et l'individualisation essentielle des peines. Le droit de ne pas faire l'objet d'une décision uniquement automatisée, le droit d'être informé de la décision automatisée, le droit de contester ou de réviser des décisions automatisées ou algorithmiques et le droit de demander une supervision et une intervention humaines doivent être garantis.

La figure 13 ci-dessous décrit quelques étapes que les opérateurs judiciaires pourraient suivre pour décider des affaires impliquant l'IA et les droits humains :

Figure 13 Étapes que les opérateurs judiciaires pourraient suivre pour décider des affaires impliquant l'IA et les droits humains



Source : Auteurs

Activité :

S'assurer que les systèmes d'IA sont utilisés d'une manière qui respecte les principes d'un procès équitable est essentiel au maintien de l'intégrité du système juridique. Voici un exemple de cas hypothétique qui illustre l'importance de l'IA pour assurer un procès équitable. Veuillez examiner les éléments de l'affaire et discuter des lois qui se seraient appliquées si l'affaire avait été jugée dans votre juridiction. Quelle en aurait été l'issue ?

Titre de l'affaire : L'État contre John Doe

Contexte : John Doe fait face à des accusations criminelles liées à un vol qualifié survenu chez un dépanneur. L'accusation s'appuie sur des images de caméras de surveillance comme pièces à conviction. Cependant, la défense soutient que les images ne sont pas concluantes et que John Doe est accusé à tort.

Rôle de l'IA dans la garantie d'un procès équitable :

1. Analyse vidéo de l'IA : L'accusation introduit un système d'analyse vidéo basé sur l'IA qui prétend améliorer et analyser les images de surveillance. On dit que ce système d'IA a la capacité d'identifier les visages, d'améliorer la qualité de l'image et de détecter les comportements suspects.
2. Préoccupations soulevées par la défense : La défense soulève des préoccupations quant à la précision et aux biais potentiels du système d'IA. Elle soutient que l'IA a peut-être été entraînée sur des ensembles de données biaisés et que ses résultats pourraient ne pas être fiables.
3. Témoins experts : L'accusation et la défense appellent des témoins experts pour témoigner sur les capacités et les limites du système d'IA. Le témoin expert de la défense remet en question la précision de l'IA et met en évidence les biais potentiels.
4. Transparence et explicabilité : La défense demande que les algorithmes et les processus décisionnels du système d'IA soient divulgués pour examen. Elle soutient que sans transparence et explicabilité, on ne peut pas faire confiance aux conclusions de l'IA.
5. Examen indépendant : Le tribunal ordonne un examen indépendant des résultats et des algorithmes du système d'IA par un tiers neutre. Cet examen vise à évaluer l'exactitude et l'équité des conclusions de l'IA.
6. Jurisprudence : L'affaire attire l'attention sur la nécessité de normes et de directives juridiques concernant l'utilisation de l'IA dans les procès criminels. La Cour examine si l'utilisation de l'IA dans cette affaire est conforme aux normes juridiques existantes et aux principes d'équité.

Résultat :

Le tribunal décide finalement d'admettre l'analyse vidéo améliorée par l'IA comme preuve, sous conditions :

- Les algorithmes et les processus décisionnels du système d'IA doivent être divulgués à la défense et au réviseur indépendant.
- Le tribunal reconnaît que les systèmes d'IA peuvent introduire des biais et des erreurs, et que les témoignages d'experts concernant les limites de l'IA seront autorisés.
- Le réviseur indépendant évaluera les conclusions de l'IA et fournira un rapport au tribunal.

Ce cas hypothétique souligne l'importance de la transparence, de l'équité et de la responsabilité, lors de l'utilisation de l'IA dans le système judiciaire. Il souligne également la nécessité de normes et de directives juridiques pour garantir que les technologies de l'IA ne compromettent pas les principes d'un procès équitable, notamment le droit à la défense, le droit de contestation des preuves et le droit d'interrogation et de contre-interrogation des témoins.

Recours effectif

L'utilisation de systèmes d'IA dans des situations où les droits humains sont en jeu peut présenter des difficultés pour assurer le droit de recours. Étant donné l'opacité de nombreux systèmes d'IA, les individus peuvent ne pas savoir comment les décisions affectant leurs droits ont été prises ou si le processus était discriminatoire. Souvent, l'opérateur judiciaire utilisant le système d'IA peut être incapable d'expliquer le processus décisionnel automatisé. Ces problèmes sont aggravés par le déploiement de systèmes d'IA qui recommandent, prennent ou exécutent des décisions au sein du pouvoir judiciaire, les institutions chargées de protéger les droits elles-mêmes, y compris le droit à un recours effectif.[287]

Contestabilité

Les individus et les groupes concernés doivent disposer de moyens efficaces de contester les résolutions et les décisions pertinentes. Comme condition préalable nécessaire, l'existence, le processus, la justification, le raisonnement et le résultat possible des systèmes algorithmiques aux niveaux individuel et collectif doivent être expliqués et clarifiés de manière opportune, impartiale, facilement lisible et accessible aux personnes dont les droits ou les intérêts légitimes peuvent être affectés, ainsi qu'aux autorités publiques compétentes. La contestation doit inclure la possibilité d'être entendu, un examen approfondi de la décision et la possibilité d'obtenir une décision non automatisée. Ce droit n'est pas dérogeable et doit être abordable et facilement exécutoire avant, pendant et après le déploiement, y compris par la mise à disposition de points de contact et de lignes directes facilement accessibles.

Source : Recommandation CM/Rec (2020) du Comité des Ministres du Conseil de l'Europe aux États membres sur l'impact des systèmes algorithmiques sur les droits humains (adoptée par le Comité des Ministres le 8 avril 2020 lors de la 1373e réunion des Délégués des Ministres)

Les processus décisionnels automatisés constituent de véritables enjeux pour la capacité des individus à obtenir un recours efficace. Il s'agit notamment de l'opacité de la décision elle-même, de son fondement et de la question de savoir si les individus ont consenti à l'utilisation de leurs données pour prendre cette décision ou s'ils sont même conscients de son impact sur eux. En raison de la difficulté à attribuer la responsabilité de la décision, il n'est pas clair pour les personnes désireuses de contester la décision de savoir à qui s'adresser. En raison de la nature des jugements rendus automatiquement, sans ou avec peu de contribution humaine, et en mettant l'accent sur l'efficacité plutôt que sur le raisonnement contextuel humain, les organisations déployant des systèmes d'ADM ont une obligation encore plus grande de fournir aux personnes touchées un moyen de demander réparation.[288] Dans ce contexte, il convient de mentionner que la proposition de directive européenne sur la responsabilité en matière d'IA créerait une « présomption de causalité » réfutable, afin d'alléger la charge de la preuve pour établir les dommages causés par un système d'IA. Cela réduirait certains des obstacles rencontrés lors de l'introduction d'une réclamation pour les préjudices causés par un système d'IA. En outre, cela donnerait aux tribunaux nationaux le pouvoir d'ordonner la divulgation de preuves concernant des systèmes d'IA soupçonnés d'avoir causé des préjudices.[289]

287 Toronto Declaration, disponible sur : https://www.torontodeclaration.org/declaration-text/english/

288 Committee of Experts on Internet Intermediaries (MSI-NET) (2018). Algorithms and Human Rights: Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications, Étude du Conseil de l'Europe, DGI/2017/12, disponible sur : https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5

289 Proposal for a Directive on adapting non contractual civil liability rules to artificial intelligence, disponible sur : https://commission.europa.eu/business-economy-euro/doing-business-eu/contract-rules/digital-contracts/liability-rules-artificial-intelligence_en

Activité :

Les participants à la formation lisent la jurisprudence sélective qui traite des boîtes noires algorithmiques dans les ADM et les systèmes d'IA, et discutent de la façon dont l'IA et les progrès technologiques affectent le droit de recours.

People c. Chubbs (2015) aux États-Unis

Une cour d'appel de Californie a confirmé un secret commercial dans une affaire pénale, en 2015, pour empêcher la divulgation du code source de TrueAllele à la défense. La décision dans l'affaire People v. Chubbs est référencée aux États-Unis pour refuser aux défendeurs l'accès à des preuves secrètes.[290] Le tribunal a statué qu'un défendeur n'a pas droit au code source d'un algorithme d'ADN utilisé pour identifier le défendeur, *prima facie*. Le propriétaire d'un secret commercial a le droit de refuser de le divulguer si l'octroi de ce droit ne sert pas à cacher la fraude ou à promouvoir l'injustice.[291] Dans ce cas, la Cour d'appel de Californie a étendu un privilège de preuve en matière de secret commercial dans une affaire pénale. Cela a permis au développeur de retenir « entièrement » le code source. L'affaire Chubbs a constitué la base d'une nouvelle jurisprudence aux États-Unis, qui refuse l'accès au code source sous-jacent des algorithmes utilisés dans l'ensemble du système de justice pénale.[292]

Affaire Uber concernant l'utilisation du programme de détection des fraudes « Mastermind » en Europe

Une affaire récente contre Uber s'est appuyée sur l'article 22 du RGPD, qui stipule que les individus « ont le droit de ne pas être soumis à une décision reposant uniquement sur un traitement automatisé, y compris le profilage, qui produit des effets juridiques à leur égard ou qui les affectent de manière significative ».[293] Les requérants ont demandé au tribunal de district d'Amsterdam d'analyser Mastermind, le programme sophistiqué de détection des fraudes d'Uber.

Invoquant les garanties du RGPD contre la prise de décision automatisée, les chauffeurs Uber au Royaume-Uni et au Portugal ont affirmé avoir été licenciés à tort par l'algorithme anti-fraude de l'entreprise. Les requérants ont affirmé que l'algorithme utilisé par Uber était automatisé (aucune intervention humaine significative) et avait entraîné la cessation de leur activité chez Uber, sans leur donner la possibilité de contester la décision prise par l'entreprise.[294]

L'objectif affiché de Mastermind est d'aider Uber à surveiller efficacement sa plateforme. La poursuite alléguait qu'Uber n'avait pas démontré que son personnel était suffisamment informé des intrants de son système de lutte contre la fraude pour prévoir les résultats ou expliquer les décisions de l'algorithme. Il indique également qu'Uber est tenu de fournir aux chauffeurs des informations précises sur toute violation présumée. Selon la plainte, les lettres de désactivation d'Uber étaient pour la plupart génériques et omettaient des informations concernant la fraude présumée. De plus, les chauffeurs n'ont pas eu la possibilité de réfuter les allégations.[295]

Un tribunal de district d'Amsterdam a ordonné à Uber de réintégrer les chauffeurs licenciés à tort par l'algorithme de l'entreprise. Il a également ordonné à Uber d'indemniser les chauffeurs avec plus de 100 000 € de dommages et intérêts.[296]

290 Milner-Smith H., Copper D. (2017). When a computer program keeps you in jail, disponible sur : https://www.nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html

291 People c. Superior Court of Los Angeles County (Chubbs) (Cal. Ct. App. 2015), disponible sur : https://www.quimbee.com/cases/people-v-chubbs

292 Chaney G. (2019). The Criminal Justice System's Algorithms Need Transparency, disponible sur : https://www.law360.com/articles/1143086/the-criminal-justice-system-s-algorithms-need-transparency

293 https://ekker.legal/wp-content/uploads/2020/10/Court-request-Uber-account-deactivation-unofficial-translation.pdf. L'article 22 du RGPD dispose que « la personne concernée a le droit de ne pas être soumise à une décision reposant uniquement sur un traitement automatisé, y compris le profilage, qui produit des effets juridiques à son égard ou qui l'affectent de manière significative ».

294 Huseinzade N. (2021). Algorithm Transparency: How to Eat the Cake and Have It Too, disponible sur : https://europeanlawblog.eu/2021/01/27/algorithm-transparency-how-to-eat-the-cake-and-have-it-too/

295 Claburn T. (2020). Uber drivers take ride biz to European court over 'Kafkaesque' algorithmic firings by Mastermind code, disponible sur : https://www.theregister.com/2020/10/26/uber_algorithmic_lawsuit/

296 Nawrat A. (2021). HR tech gone wrong? Uber told to reinstate drivers after 'robo-firing', disponible sur : https://www.unleash.ai/hr-technology/court-rules-against-uber-robo-firing-employee-surveillance/.

Le cas de Robodebt en Australie

En 2016, le gouvernement australien a introduit « Robodebt », un système automatisé de mise en correspondance des données, pour remplacer l'examen humain des données sur le revenu des bénéficiaires de l'aide sociale. L'objectif était de détecter les trop-payés et les fraudes. Cependant, les personnes signalées par l'algorithme comme suspectes étaient tenues de fournir des preuves de leur innocence via un formulaire en ligne, faute de quoi, elles risquaient de perdre l'ensemble de leurs avantages. Ce processus a eu des effets néfastes sur de nombreuses personnes.

Cependant, l'algorithme a pris les données de l'administration fiscale (basées sur une année complète) et les a comparées au revenu bimensuel, ignorant le fait que les revenus des bénéficiaires de l'aide sociale sont souvent très irréguliers, en raison de contrats à court terme ou de travail saisonnier. En conséquence, des milliers de personnes ont été privées à tort de prestations sociales, et beaucoup d'entre elles n'ont pas pu contester ces décisions, car soit les notifications automatisées étaient envoyées à une ancienne adresse, soit l'accès au portail via lequel elles auraient pu transmettre les preuves requises était bloqué.

Dans de nombreux cas, les gens se sont soudainement retrouvés gravement endettés, et certains cas de suicide ont même été signalés. Certaines sources calculent que les autorités ont tenté de réclamer près de 600 millions de dollars australiens (360 millions d'euros) aux citoyens sur la base de ce système, qui a souvent généré des erreurs, mais dans le cadre duquel la charge de la preuve a été transférée à l'individu. Les résultats étaient très difficiles à contester. Cette affaire a relancé le débat sur la manière dont les algorithmes et l'appariement des données sont utilisés pour éclairer les décisions.[297]

Le règlement proposé pour une action collective contre le Commonwealth d'Australie concernant son utilisation de Robodebt a été approuvé par la Cour fédérale, le 11 juin 2021. Conformément au règlement, le Commonwealth paiera 112 millions de dollars (incluant les frais de justice) à certains membres du groupe, au titre d'intérêts, s'abstiendra de contracter, d'exiger ou de recouvrer des dettes invalides de certains membres du groupe et acceptera les déclarations judiciaires selon lesquelles certaines de ses décisions administratives n'ont pas été valablement prises.[298]

297 Human Rights Law Centre (2021). The Federal Court approves a $112 million settlement for the failures of the Robodebt system, disponible sur : https://www.hrlc.org.au/human-rights-case-summaries/2021/9/30/the-federal-court-approves-a-112-million-settlement-for-the-failures-of-the-robodebt-system.

298 *Ibid.* Voir également : Katherine Prygodicz & Ors v. The Commonwealth of Australia (No 2) [2021] FCA 634 (11 June 2021).

« Toutes les personnes sont égales devant la loi et ont droit sans discrimination à une égale protection de la loi. À cet égard, la loi doit interdire toute discrimination et garantir à toutes les personnes une protection égale et efficace contre toute discrimination, notamment de race, de couleur, de sexe, de langue, de religion, d'opinion politique et de toute autre opinion, d'origine nationale ou sociale, de fortune, de naissance ou de toute autre situation. » - Article 26 du PIDCP

« Dans les États où il existe des minorités ethniques, religieuses ou linguistiques, les personnes appartenant à ces minorités ne peuvent êtres privées du droit d'avoir, en commun avec les autres membres de leur groupe, leur propre vie culturelle, de professer et pratiquer leur propre religion, ou d'employer leur propre langue. » - Article 27 du PIDCP

« Les États parties au présent Pacte s'engagent à assurer le droit égal des hommes et des femmes au bénéfice de tous [...] les droits énoncés dans le présent Pacte. »
– Article 3 du PIDCP et du PIDESC

Les droits à la protection contre la discrimination peuvent être violés par les systèmes d'IA, en raison (i) du potentiel de biais de la part des développeurs d'algorithmes ; (ii) du biais intégré dans le modèle sur lequel les systèmes d'IA sont construits ; (iii) du biais intégré dans les ensembles de données utilisés pour former les modèles ; ou (iv) du biais introduit lorsque de tels systèmes sont appliqués dans des contextes réels. Ces risques s'exacerbent dans les situations où des systèmes d'IA sont déployés pour assister les opérateurs judiciaires dans leurs activités quotidiennes.

La conception des systèmes d'IA et leur utilisation dans les procédures judiciaires doivent être régies dans le but de produire des résultats conformes aux droits humains et non discriminatoires. Des normes et des garanties minimales doivent être établies ; si elles ne peuvent pas être respectées, le système d'IA en question ne doit pas être utilisé.

En outre, l'IA doit être réglementée de manière à être suffisamment transparente et explicable pour permettre un examen indépendant efficace. La conception et le déploiement des systèmes d'IA doivent respecter et donner effet au droit à l'accès aux tribunaux, au droit à la présomption d'innocence et au droit à la liberté, entre autres.

Aucun être humain ne doit être exposé à une décision automatisée aboutissant à un casier judiciaire, et les technologies de l'IA ne doivent pas compromettre le droit à un procès équitable devant un tribunal impartial et indépendant. Les systèmes d'IA ne doivent pas cataloguer de prime abord des individus comme criminels sans procès, ni permettre aux autorités de prendre des mesures injustifiées et disproportionnées contre des individus sans soupçon raisonnable.

Lorsque les systèmes d'IA éclairent les décisions sur les privations de liberté, ils doivent être ajustés pour créer des résultats qui favorisent la libération, et ils ne doivent pas faciliter la détention, sauf en dernier recours. Pour s'assurer que les systèmes d'IA atteignent l'effet souhaité de réduction des taux de détention provisoire, ils doivent être soumis à des tests rigoureux.[299]

Questions qui doivent être prises en compte par les opérateurs judiciaires, lors de l'évaluation de l'impact potentiel et du risque de l'IA sur les droits à la protection contre la discrimination

- Comment, le cas échéant, le système d'IA pourrait-il entraîner une discrimination, avoir des impacts discriminatoires sur les titulaires de droits ou fonctionner différemment pour différents groupes de manière discriminatoire ou préjudiciable ?
- Comment, le cas échéant, l'utilisation du système d'IA pourrait-elle exacerber les inégalités ou la discrimination existantes dans les populations qu'il affecte ?
- Le cas échéant, de quelles manières supplémentaires l'utilisation de ce système pourrait-elle contribuer ou exacerber les inégalités ?

Source : Leslie D., Burr C., Aitken M., Cowls J., Katell M., Briggs M. (2021). Artificial intelligence, human rights, democracy, and the rule of law: a primer, Le Conseil de l'Europe, disponible sur : https://www.turing.ac.uk/sites/default/files/2021-03/cahai_feasibility_study_primer_final.pdf.

Les systèmes d'IA doivent être conçus pour s'assurer qu'ils ne produisent pas de résultats discriminatoires, en veillant à ce que les suspects et les accusés ne soient pas désavantagés, directement ou indirectement, en fonction de leurs caractéristiques, telles que la race, l'origine ethnique, la nationalité, la minorité ou le statut socio-économique. Les systèmes d'IA doivent être soumis à des tests obligatoires avant et après le déploiement, afin d'identifier et de corriger tout effet discriminatoire. Veuillez vous référer au module 3 qui traite en détail des biais algorithmiques.[300]

Les systèmes d'IA doivent être transparents et compréhensibles afin que leurs principaux utilisateurs, tels que les décideurs, les parties à un litige, les défendeurs, puissent les comprendre et les examiner. Les intérêts commerciaux ou de propriété, tels que les secrets commerciaux, doivent être mis en balance avec les exigences liées à la transparence. Chaque système d'IA doit pouvoir être audité par un auditeur indépendant et ses processus doivent pouvoir être reproduits à cette fin.[301]

299 Fair Trials, Regulating Artificial Intelligence for Use in Criminal Justice Systems in the EU Policy Paper, disponible sur : https://www.fairtrials.org/sites/default/files/Regulating%20Artificial%20Intelligence%20for%20Use%20in%20Criminal%20Justice%20Systems%20-%20Fair%20Trials.pdf

300 *Ibid.*

301 *Ibid.*

Activité :

Les participants à la formation lisent les éléments des affaires Deliveroo et Foodinho, et discutent de la manière dont l'opacité des algorithmes d'IA et leur fonctionnement en tant que boîtes noires affectent la protection contre la discrimination et les préjugés personnels.

Affaire Deliveroo (2021)[302]

Deliveroo est un service de livraison de repas qui fonctionne comme un marché à trois facettes, via une application en ligne. Il met en relation les consommateurs locaux, les restaurants et épiceries, et les livreurs. Trois syndicats ont contesté Deliveroo devant les tribunaux italiens pour violation des lois régionales du travail. Dans cette affaire, le tribunal de Bologne a statué que l'algorithme d'évaluation de la réputation de Deliveroo était discriminatoire à l'égard des préparateurs ou des livreurs de repas.[303] L'algorithme de ML examiné par le tribunal aurait été utilisé pour estimer la « fiabilité » d'un livreur. Le tribunal a noté que l '« indice de fiabilité » du livreur en pâtirait si ce dernier n'annulait pas une mission pré-réservée via l'application au moins 24 heures avant l'heure de début. Étant donné que l'algorithme donne la priorité sur les missions en heure de pointe à des livreurs plus fiables, ceux qui ne peuvent pas les effectuer, même en cas d'urgence ou de maladie grave, auront moins de propositions d'emploi, à l'avenir. Selon le tribunal, le fait que l'algorithme de ML n'ait pas examiné la cause de l'annulation constituait une discrimination et pénalisait injustement les livreurs qui avaient des raisons juridiquement valables de ne pas travailler. Deliveroo a été condamné à indemniser les demandeurs à hauteur de 50 000 €.[304]

Le tribunal a également noté que les critères de fonctionnement de l'algorithme n'étaient ni définis sur l'application au-delà des aspects génériques de fiabilité et de participation, ni fournis au tribunal par la société défenderesse, ce qui empêchait une évaluation approfondie de la question.[305]

Affaire Foodinho (2021)

Foodinho, un autre service de livraison de nourriture basé en Italie, a été condamné à verser 2,6 millions d'euros par l'Autorité italienne de protection des données (Garante) pour avoir utilisé des algorithmes de mesure de la performance discriminatoires à l'égard de ses employés. L'autorité a déclaré Foodinho en violation des principes de transparence, de sécurité et de protection de la vie privée par défaut et par conception, et elle a tenu la société pour responsable de ne pas avoir pris les mesures appropriées pour protéger les droits et libertés de ses employés (i.e., les livreurs) contre l'ADM discriminatoire. En termes de gestion algorithmique des travailleurs à la tâche, la décision de la Garante est une première en son genre. Elle a affirmé que la direction de Foodinho avait violé l'article 22(3) du RGPD.[306]

Dans sa décision, la Garante a déclaré que Foodinho s'engage dans deux types différents d'activités de traitement automatisé : l'une relève du « système d'excellence » et l'autre est une composante du système qui distribue des commandes basées sur un algorithme interne, connu sous le nom de « Jarvis ». La méthode de notation interne utilisée par Foodinho pour fournir des créneaux de livraison à ses livreurs est connu sous le nom de « système d'excellence », qui évalue chaque livreur. Les livreurs ayant des notes plus élevées sont prioritaires lors de la détermination des créneaux de livraison.

302 Colossa A. (2021). Algorithms, biases, and discrimination in their use: About recent judicial rulings on the subject, disponible sur : https://www.ciat.org/ciatblog-algorithms-biases-and-discrimination-in-their-use-about-recent-judicial-rulings-on-the-subject/?lang=en

303 Lomas N. (2021). Italian court rules against 'discriminatory' Deliveroo rider-ranking algorithm, disponible sur : https://techcrunch.com/2021/01/04/italian-court-rules-against-discriminatory-deliveroo-rider-ranking-algorithm/.

304 Geiger G. (2021). Court Rules Deliveroo Used 'Discriminatory' Algorithm, disponible sur : https://www.business-humanrights.org/en/latest-news/court-rules-deliveroo-used-discriminatory-algorithm/.

305 *Ibid.*

306 Milner-Smith et al. (2021). Italian Supervisory Authority Fines Foodinho Over Its Use of Performance Management Algorithms, disponible sur : https://www.insideprivacy.com/gdpr/italian-supervisory-authority-fines-foodinho-over-its-use-of-performance-management-algorithms/.

En pratique, cela signifie que les livreurs « moins excellents » sont exclus de l'attribution des créneaux de livraison, si les « plus excellents » ont déjà pris tous les créneaux de livraison disponibles. La « note d'excellence » est déterminée par un processus statistique automatisé qui prend principalement en compte les commentaires des clients et des partenaires commerciaux ainsi que les taux de livraison. Il est important de noter que les commentaires positifs ont moins de poids que les commentaires négatifs, et que le système pénalise les livreurs qui ne respectent pas les niveaux de livraison requis. L'algorithme (« Jarvis ») qui attribue les commandes utilise des informations telles que l'emplacement géographique des livreurs tel que déterminé par leurs appareils GPS, le lieu de prise en charge, l'adresse de livraison, toute exigence de commande spéciale et le type de véhicule utilisé. Jarvis attribue des commandes et automatise entièrement le traitement de ces données. Cependant, Foodinho n'a pas spécifiquement expliqué à la Garante comment cet algorithme est lié au système d'excellence.[307]

Liberté d'expression et accès à l'information

« Toute personne a droit à la liberté de pensée, de conscience et de religion. Ce droit implique la liberté d'avoir ou d'adopter une religion ou une conviction de son choix, ainsi que la liberté de manifester sa religion ou sa conviction, individuellement ou en commun, tant en public qu'en privé, par le culte et l'accomplissement des rites, les pratiques et l'enseignement. Nul ne subira de contrainte pouvant porter atteinte à sa liberté d'avoir ou d'adopter une religion ou une conviction de son choix. »

– Article 18 du PIDCP et article 18 de la DUDH

« Nul ne peut être inquiété pour ses opinions. Toute personne a droit à la liberté d'expression ; ce droit comprend la liberté de rechercher, de recevoir et de répandre des informations et des idées de toute espèce, sans considération de frontières, sous une forme orale, écrite, imprimée ou artistique, ou par tout autre moyen de son choix. »

- Article 19 du PIDCP

Plusieurs cadres juridiques et principes directeurs internationaux établissent que les droits humains à la liberté d'expression et à l'accès à l'information s'étendent à Internet. En 2011, le Comité des droits de l'homme des Nations Unies a publié l'observation générale n° 34[308] déclarant que l'article 19 du PIDCP[309] protège toutes les formes d'expression et les moyens de leur diffusion, y compris toutes

307 *Ibid.*
308 ONU (2011). General Comment No 34, disponible sur : https://www2.ohchr.org/english/bodies/hrc/docs/gc34.pdf
309 ONU (1976). International Covenant on Civil and Political Rights, disponible sur : https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights

les formes de modes d'expression électroniques et sur Internet (y compris l'accès à l'information en ligne). Cela signifie que le principe de sauvegarde du droit à la liberté d'expression s'étend à l'espace en ligne tout comme il le fait dans le monde hors ligne.[310] En 2012, le Conseil des droits de l'homme des Nations Unies a adopté une résolution révolutionnaire 20/8[311] pour promouvoir, protéger et assurer la jouissance des droits humains en ligne. Cette résolution affirme l'importance de faire respecter les droits humains à l'ère numérique : « Les mêmes droits que les personnes ont hors ligne doivent également être protégés en ligne, en particulier la liberté d'expression, qui est applicable sans considération de frontières et par tout moyen de leur choix, conformément aux articles 19 de la Déclaration universelle des droits de l'homme et du Pacte international relatif aux droits civils et politiques.[312] De même, la résolution de 2018 du Conseil des droits de l'homme des Nations Unies sur la promotion, la protection et la jouissance des droits humains sur Internet a déclaré que « les mêmes droits que les personnes ont hors ligne doivent également être protégés en ligne, en particulier la liberté d'expression »[313] et a appelé tous les États à garantir ces droits.

Les rapports annuels et thématiques du rapporteur spécial abordent diverses questions telles que la surveillance des communications par l'État[314], la sauvegarde des droits des citoyens pendant les élections[315], les discours de haine en ligne[316], le cryptage et l'anonymat[317], le droit des enfants à s'exprimer[318], le rôle du secteur privé[319] et des fournisseurs d'accès numérique,[320] l'impact de l'intelligence artificielle sur les droits des citoyens,[321] la protection de la liberté d'expression des journalistes[322] et la prévention de la censure, tout en luttant contre les abus sexistes en ligne[323].

310 ONU (2011). General Comment No 34 (para 15), disponible sur : https://www2.ohchr.org/english/bodies/hrc/docs/gc34.pdf
311 Conseil des droits de l'homme des Nations Unies (2021). The promotion, protection and enjoyment of human rights on the Internet, disponible sur : https://ap.ohchr.org/documents/dpage_e.aspx?si=a/hrc/res/20/8
312 ONU (AGNU) (2012). The promotion, protection and enjoyment of human rights on the Internet, 16 July.2012, A/HRC/RES/20/8, disponible sur : http://ap.ohchr.org/documents/dpage_e.aspx?si=A/HRC/RES/20/8
313 Conseil des droits de l'homme des Nations Unies (2018). The Promotion, Protection and Enjoyment of Human Rights on the Internet, disponible sur : https://digitallibrary.un.org/record/1639840
314 Conseil des droits de l'homme des Nations Unies (2013). Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, disponible sur : https://www.ohchr.org/Documents/HRBodies/HRCouncil/RegularSession/Session23/A.HRC.23.40_EN.pdf
315 Conseil des droits de l'homme des Nations Unies (2014). Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, disponible sur : https://documents-dds-ny.un.org/doc/UNDOC/GEN/G14/071/50/PDF/G1407150.pdf?OpenElement
316 Assemblée générale des Nations Unies (2019). Promotion and protection of the right to freedom of opinion and expression, disponible sur : https://www.ohchr.org/Documents/Issues/Opinion/A_74_486.pdf
317 Conseil des droits de l'homme des Nations Unies (2015). Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, disponible sur : https://documents-dds-ny.un.org/doc/UNDOC/GEN/G15/095/85/PDF/G1509585.pdf?OpenElement
318 Assemblée générale des Nations Unies (2014). Promotion and protection of the right to freedom of opinion and expression, disponible sur : https://documents-dds-ny.un.org/doc/UNDOC/GEN/N14/512/72/PDF/N1451272.pdf?OpenElement
319 Conseil des droits de l'homme des Nations Unies (2016). Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, disponible sur : https://documents-dds-ny.un.org/doc/UNDOC/GEN/G16/095/12/PDF/G1609512.pdf?OpenElement
320 Conseil des droits de l'homme des Nations Unies (2017). Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, disponible sur : https://documents-dds-ny.un.org/doc/UNDOC/GEN/G17/077/46/PDF/G1707746.pdf?OpenElement
321 Assemblée générale des Nations Unies (2018). Promotion and protection of the right to freedom of opinion and expression, disponible sur : https://documents-dds-ny.un.org/doc/UNDOC/GEN/N18/270/42/PDF/N1827042.pdf?OpenElement
322 Conseil des droits de l'homme des Nations Unies (2012). Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, disponible sur : https://documents-dds-ny.un.org/doc/UNDOC/GEN/G12/137/87/PDF/G1213787.pdf?OpenElement7
323 Voir : http://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=21317&LangID=E ; DigWatch (2023). Freedom of expression online in 2023, disponible sur : https://dig.watch/topics/freedom-expression

Les plateformes numériques sont pilotées par des algorithmes qui déterminent comment gérer, hiérarchiser, distribuer et retirer ou supprimer des informations tierces en ligne. Il est possible que ces activités ne répondent pas aux normes de légalité, de légitimité et de proportionnalité pour des restrictions raisonnables à la liberté d'expression. En outre, les violations des informations personnelles ont un effet dissuasif sur la liberté d'expression. Les gens s'autocensurent et modifient leur comportement lorsqu'ils craignent d'être observés ou de manquer d'anonymat. Cet effet sera amplifié par la surveillance alimentée par l'IA, qui peut avoir un impact négatif sur la liberté d'expression.

Activité :

IA et liberté d'expression

Les participants regardent la vidéo et discutent des implications possibles de l'IA sur la liberté d'expression.

Source : UNESCO, https://www.youtube.com/watch?v=j0Oz54A68qo

Dans le monde numérique d'aujourd'hui, la jouissance de la liberté d'expression est régie dans des domaines privés, hybrides et publics façonnés par des entreprises privées, des autorités gouvernementales et des utilisateurs, dans des relations de pouvoir variées et hautement asymétriques. En outre, ces écosystèmes numériques ont ouvert la voie à de nouveaux types de gouvernance de l'expression, tels que ceux modérés par des systèmes d'IA sur les plateformes de médias sociaux pour organiser le contenu des flux d'actualités des utilisateurs.

IA, modération de contenu et liberté d'expression

Les intermédiaires Internet modèrent les contenus sur leurs plateformes. Cette modération de contenu est souvent menée à l'insu du public et est fréquemment effectuée par des systèmes d'IA opaques à grande échelle, sans garantie de conformité au cadre international des droits humains. Ces instruments automatisés peuvent imposer des restrictions au droit à la liberté d'expression et à l'accès à l'information, quelle que soit la méthode technologique employée.[324] Ils peuvent exclure du discours public des individus, des organisations, des idées ou des formes d'expression spécifiques.

Alors que la quantité d'informations en ligne nécessitant une modération augmente inévitablement et de manière exponentielle, les principales plateformes en ligne investissent massivement dans les systèmes d'IA pour automatiser la modération du contenu. Les lois de modération de contenu promulguées dans le monde entier imposent de lourdes amendes en cas de non-conformité, si les plateformes en ligne ne parviennent pas à supprimer rapidement les informations qui enfreignent les lois nationales sur la propriété intellectuelle, ainsi que les lois contre les discours haineux et la pédopornographie.[325]

L'un des principaux problèmes associés à l'automatisation de la modération de contenu est que les technologies d'IA utilisées pour cela sont construites sur une technologie de TAL spécifique au domaine, c'est-à-dire que la technologie n'identifiera que les types de contenu sur lesquels elle a été formée. Par exemple, un système de TAL formé pour identifier les discours racistes est incapable d'identifier les contenus violents. De plus, même au sein d'un sujet particulier, les algorithmes de TAL peuvent ne pas être en mesure de comprendre les nuances détaillées de la parole humaine, telles que le sarcasme et la parodie.[326] Un système capable de détecter un contenu raciste dans un article de blog peut ne pas reconnaître de manière fiable un contenu similaire dans un tweet, ce qui entraîne un taux d'erreur très élevé pour ces technologies.[327]

> Un guide ingénieux sur les questions de liberté d'expression et d'accès à l'information dans l'environnement numérique est le document de l'UNESCO intitulé **« Préserver la liberté d'expression et l'accès à l'information : principes pour une approche multipartite dans le contexte de la régulation des plateformes numériques »**

Pour illustrer davantage ce point, lors de l'épidémie de coronavirus, YouTube a remplacé bon nombre de ses examinateurs de contenu humains par des algorithmes d'IA chargés d'identifier et de supprimer les vidéos contenant de la désinformation et des discours de haine. L'expérience de modération de contenu sur la plateforme a échoué. Les algorithmes d'IA censuraient excessivement les utilisateurs, triplant ainsi le taux de suppression de contenu inexact. Après quelques mois, YouTube a réembauché certains de ses modérateurs humains.[328] Un autre exemple serait le cas de Kate Klonick, spécialiste de la modération de contenu, qui a été bannie de Twitter pour avoir publié un tweet contenant la phrase « Je vais t'assassiner », que l'algorithme

324 OSCE (2022). Spotlight on Artificial Intelligence and Freedom of Expression: A Policy Manual, disponible sur : https://www.osce.org/representative-on-freedom-of-media/510332

325 Raso F., Hilligoss H., Krishnamurthy V., Bavitz C., Kim L. (2018). Artificial Intelligence & Human Rights: Opportunities & Risks, disponible sur : https://cyber.harvard.edu/publication/2018/artificial-intelligence-human-rights

326 Mindmatters (2021). Can the machine know you are just being sarcastic, disponible sur : https://mindmatters.ai/2021/05/can-the-machine-know-you-are-just-being-sarcastic/.

327 *Ibid.* Voir aussi : Gaumond E., Régis C. (2023). Assessing Impacts of AI on Human Rights: It's Not Solely About Privacy and Nondiscrimination, disponiblesur :https://www.lawfareblog.com/assessing-impacts-ai-human-rights-its-not-solely-about-privacy-and-nondiscrimination.

328 *Ibid.* Voir aussi : Vincent J. (2020). YouTube brings back more human moderators after AI systems over-censor, disponible sur : https://www.theverge.com/2020/9/21/21448916/youtube-automated-moderation-ai-machine-learning-increased-errors-takedowns

de Twitter considérait comme une incitation à la violence.[329] Pourtant, Klonick ne préconisait en aucune façon la violence. En réalité, elle faisait référence à un échange humoristique entre Molly Jong-Fast et son mari, qui allait lui prendre son repas.

Il convient de noter que les outils de TAL ne sont pas encore aussi efficaces dans les langues autres que l'anglais. Par conséquent, les outils automatisés peuvent ne pas être aussi précis dans l'évaluation des non-anglophones, ce qui peut limiter injustement leur liberté d'expression. Cela est particulièrement vrai pour les outils de traduction linguistique, qui ont parfois du mal avec des significations et un contexte nuancés. Par exemple, un homme israélo-palestinien a été arrêté après avoir publié une photo sur Facebook avec la légende « bonjour », en arabe. Cependant, l'outil de traduction alimenté par l'IA de Facebook a incorrectement traduit la légende par « attaquez-les » en hébreu, ou « faites-les souffrir » en anglais. Facebook a reconnu plus tard son erreur et s'est excusé auprès de l'homme et de sa famille pour la gêne occasionnée.[330]

Activité :

Les participants à la formation lisent l'affaire Gonzalez contre Google et discutent des lois qui seraient applicables dans leurs juridictions dans ces circonstances. Quelle aurait été l'issue de l'affaire ?

En 2023, la Cour suprême des États-Unis a été saisie d'une affaire intéressante, Gonzalez c. Google. L'affaire a débuté après la mort tragique de Nohemi Gonzalez, 23 ans, dans les attentats terroristes de Paris, en 2015. La famille de Nohemi Gonzalez a cherché à tenir Google responsable pour son rôle dans les attaques, en vertu de la loi antiterroriste, qui permet aux familles des personnes tuées par des terroristes d'intenter une action en justice contre ceux qui « aident et encouragent » ces groupes. Initialement, la Cour suprême a refusé de se prononcer sur l'affaire, en particulier sur la question de savoir si les recommandations ciblées par les algorithmes des médias sociaux seraient exclues de la protection de l'article 230 de la loi sur la décence des communications. Cette décision a des implications pour l'avenir de la responsabilité, dans des cas similaires.

Source : https://www.supremecourt.gov/opinions/22pdf/21-1333_6j7a.pdf

329 Kloncik K. (2020). What I Learned in Twitter Purgatory, disponible sur : https://www.theatlantic.com/ideas/archive/2020/09/what-i-learned-twitter-purgatory/616144/; Gaumond E., Régis C. (2023). Assessing Impacts of AI on Human Rights: It's Not Solely About Privacy and Nondiscrimination, disponible sur : https://www.lawfareblog.com/assessing-impacts-ai-human-rights-its-not-solely-about-privacy-and-nondiscrimination

330 yHu X., Neupane B., Flores Echaiz L., Sibal P., Rivera Lam M. (2019). Rapport UNESCO Steering AI and advanced ICTs for knowledge societies: a Rights, Openness, Access, and Multi-stakeholder Perspective, disponible sur : https://unesdoc.unesco.org/ark:/48223/pf0000372132

Dans un environnement où les plateformes de médias sociaux utilisent des algorithmes pour décider quelles voix nous entendons, le droit à la liberté d'expression revêt une importance particulière. En 2014, des chercheurs de l'Université Cornell ont mené une étude sur la contagion émotionnelle, en collaboration avec Facebook, sur le réseau social.[331] Les chercheurs ont modifié les expériences de plus de 700 000 utilisateurs de Facebook en utilisant une technique d'analyse des sentiments, pour déterminer si des amis avaient contribué à des commentaires ou des messages désagréables. Ces éléments négatifs ont ensuite été supprimés des flux d'actualités des utilisateurs, dans le cadre d'une expérience visant à déterminer si le biais algorithmique du flux vers un contenu positif permettrait aux utilisateurs de rester plus longtemps sur le site. Cette étude met en évidence la façon dont les plateformes sont susceptibles de prendre des décisions basées sur des expressions d'utilisateurs qui encouragent une réalité et en dévaluent une autre.[332]

Activité :

Cas de modération de contenu « The Napalm Girl »

L'affaire de modération de contenu « Napalm Girl » fait référence à un incident controversé impliquant la modération d'images historiques et emblématiques liées à la guerre, sur les plateformes de médias sociaux. L'affaire tourne autour de la suppression ou de la censure d'une photographie lauréate du prix Pulitzer, intitulée « La terreur de la guerre », qui représente une jeune fille, Kim Phúc, fuyant une attaque au napalm pendant la guerre du Vietnam. Les participants à la formation lisent l'aperçu du cas et discutent de ses implications pour la liberté d'expression dans l'environnement numérique.

Historique :

- La photographie a été prise par Nick Ut, photographe de l'Associated Press (AP), le 8 juin 1972, pendant la guerre du Vietnam. Elle illustre les conséquences immédiates d'un attentat au napalm à Trang Bang, au Sud-Vietnam.
- L'image montre une fillette de neuf ans, nue et gravement brûlée, Kim Phúc, courant sur une route, à l'agonie.
- La photographie est devenue un symbole emblématique des horreurs de la guerre et a joué un rôle important dans la sensibilisation au coût humain de la guerre du Vietnam.

Incident de modération de contenu :

- En septembre 2016, Facebook a temporairement retiré la photo publiée par l'écrivain norvégien Tom Egeland, dans le cadre d'une série de photographies de guerre emblématiques.
- La politique de Facebook contre l'affichage de la nudité sur la plateforme était à l'origine de cette suppression.
- La décision a suscité l'indignation et la controverse, beaucoup faisant valoir que la signification historique et journalistique de la photographie devrait l'emporter sur les préoccupations concernant la nudité.
- Après d'importantes réactions et critiques du public, Facebook a infirmé sa décision et rétabli la photo.

331 Hu X., Neupane B., Flores Echaiz L., Sibal P., Rivera Lam M. (2019). Rapport UNESCO Steering AI and advanced ICTs for knowledge societies: a Rights, Openness, Access, and Multi-stakeholder Perspective, disponible sur : https://unesdoc.unesco.org/ark:/48223/pf0000372132.

332 Meyer R. (2014). Everything We Know About Facebook's Secret Mood-Manipulation Experiment, disponible sur : https://www.theatlantic.com/technology/archive/2014/06/everything-we-know-about-facebooks-secret-mood-manipulation-experiment/373648/

Questions et débats clés :

1. Liberté d'expression vs modération de contenu : L'affaire soulève des questions sur l'équilibre entre la liberté d'expression, le partage de contenu historique et digne d'intérêt, et la nécessité de modérer le contenu pour empêcher la diffusion de matériel inapproprié ou offensant.
2. Modération algorithmique : De nombreuses plateformes de médias sociaux utilisent des algorithmes pour détecter et supprimer automatiquement les contenus qui enfreignent leurs politiques. En l'espèce, les algorithmes n'ont pas réussi à faire la distinction entre une photographie historique lauréate du prix Pulitzer et un contenu inapproprié.
3. Sensibilité culturelle et contexte : Les critiques soutiennent que les algorithmes de modération de contenu n'ont pas la capacité de comprendre la signification historique, culturelle et contextuelle de certaines images, ce qui conduit à des suppressions erronées.
4. Responsabilité des entreprises technologiques : L'incident remet également en question la responsabilité des entreprises technologiques dans la prise de décisions nuancées concernant la modération du contenu et l'impact potentiel de ces décisions sur la liberté d'expression et la documentation historique.

En fin de compte, le cas de la modération de contenu « Napalm Girl » met en évidence les défis auxquels sont confrontées les plateformes de médias sociaux et les entreprises technologiques, pour trouver un équilibre entre la modération de contenu afin de respecter les normes de la communauté, et la reconnaissance de l'importance du contenu historique et journalistique, en particulier lorsqu'il dépeint des sujets sensibles ou pénibles. Il souligne la nécessité de politiques et de décisions de modération de contenu réfléchies et tenant compte du contexte.

Source : Content Moderation Case Study: Facebook Attracts International Attention When It Removes A Historic Vietnam War Photo Posted By The Editor-in-Chief Of Norway's Biggest Newspaper (2016), disponible sur : https://www.techdirt.com/2020/11/20/content-moderation-case-study-facebook-attracts-international-attention-when-it-removes-historic-vietnam-war-photo-posted/

Désinformation et IA

Les technologies de l'IA peuvent contribuer à un accès inégal à l'information et exacerber les fractures numériques existantes. Par exemple, l'IA peut être utilisée pour développer et diffuser une propagande ciblée, et ce problème est exacerbé par les algorithmes de médias sociaux alimentés par l'IA, animés par un « engagement » qui promeut les informations les plus susceptibles d'être cliquées. L'analyse des données utilisée par les entreprises de médias sociaux afin d'élaborer des profils d'utilisateurs pour la publicité ciblée est alimentée par des algorithmes de ML. En outre, des robots se faisant passer pour des utilisateurs authentiques propagent du contenu en dehors de groupes de médias sociaux étroitement ciblés, en distribuant des liens vers de fausses sources et en communiquant activement avec les gens en tant que chatbots, via un traitement du langage naturel.[333]

Les entités déployant des algorithmes de dépistage et de notation de l'IA n'offrent souvent pas de notification appropriée, le cas échéant, à celles qui sont notées et évaluées. Parce que les consommateurs ne savent pas comment ces outils prennent des décisions et à quels types de données ils recourent, leur utilisation peut éroder les restrictions relatives à l'accès à l'information. Parce que les individus ne comprennent pas comment ces outils fonctionnent, ils sont incapables de contester les décisions d'éligibilité affectant leur accès aux services, aux emplois, au logement ou aux avantages.[334]

333 *Ibid.*
334 Voir : https://epic.org/issues/ai/screening-scoring/

En outre, la menace des deepfakes, systèmes d'IA capables de réaliser des enregistrements vidéo et audio réalistes de personnes réelles, a conduit de nombreux individus à croire qu'à l'avenir, la technologie servira à réaliser de fausses images de dirigeants mondiaux, à des fins néfastes. Bien qu'il semble que les deepfakes n'aient pas encore été utilisés dans le cadre de campagnes de propagande ou de désinformation réelles, et que l'audio et la vidéo falsifiés ne soient pas encore convaincants sur l'aspect humain, l'IA qui les alimente progresse et le potentiel de propagation du chaos, d'incitation au conflit et de poursuite de la crise de la vérité ne doit pas être écarté.[335]

Dans les pays où la liberté religieuse est menacée, l'IA pourrait aider les responsables gouvernementaux à surveiller et à cibler les membres des organisations religieuses persécutées. Non seulement cela peut augmenter le secret de ces rassemblements par crainte d'être détectés, mais cela pourrait également entraîner des conséquences physiques allant de l'arrestation à la mort. En outre, l'IA peut être utilisée pour identifier et supprimer du contenu religieux. Si les gens étaient dans l'impossibilité de montrer des symboles religieux, de prier ou d'enseigner leur foi en ligne, cela constituerait une violation flagrante de la liberté de religion.[336]

L'ONG AccessNow souligne que le harcèlement en ligne activé par les robots constitue une menace claire et imminente pour la liberté d'expression. Ces comptes robots se présentent comme des utilisateurs humains et fournissent des réponses automatiques aux comptes désignés ou à toute personne partageant un point de vue particulier. Ce type de harcèlement en ligne incessant a un impact paralysant sur la liberté d'expression, en particulier pour les groupes défavorisés qui sont ciblés de manière disproportionnée. Les développeurs de robots appliquent plus fréquemment le traitement du langage naturel, ce qui exacerbe les menaces de harcèlement en ligne par les robots. Cela va compliquer l'identification, le signalement et l'élimination de ces comptes robots.[337]

Restrictions légitimes à la liberté d'expression et à l'accès à l'information

Dans le cadre international des droits humains et dans de nombreuses constitutions, il existe des conditions strictes pour justifier des limites préalables à la liberté d'expression et à l'accès à l'information. De ce point de vue, les outils d'IA sont particulièrement inquiétants, car ces systèmes sont cachés à l'examen du public, ne tiennent pas compte du contexte et fonctionnent de manière très opaque, ce qui empêche toute correction ou réponse efficaces. Bien que la présélection du contenu pour restreindre la transmission en ligne de logiciels malveillants et d'abus sexuels sur des enfants ait été largement considérée comme une application utile de l'automatisation, il faut faire preuve de prudence lors de l'application du même raisonnement à d'autres types de discours qui relèvent de la catégorie plus large du règlement sur le contenu.[338] Le droit international permet de restreindre les droits numériques (le droit à la vie privée, la liberté d'expression et l'accès à l'information) à la fois hors ligne et en ligne, mais

335 AccessNow (2018). AI and human rights, disponible sur : https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf
336 *Ibid.*
337 *Ibid.*
338 *Ibid.*

seulement dans des circonstances très limitées et spécifiques, et conformément à l'article 19 du PIDCP (liberté d'expression et accès à l'information), au moyen du test en trois parties décrit ci-dessous.[339]

Tableau 6. Test en trois parties des limites légitimes à la liberté d'expression

Principe	Explication
Les restrictions doivent être prévues par la loi	• Les lois sur les TIC doivent clairement stipuler toute restriction à la liberté d'expression, sans ambiguïté. Les citoyens doivent être en mesure de comprendre et de respecter les lois, ce qui les rend légitimes. Des dispositions vagues et trop larges ne satisferaient pas à cette norme. • Le Comité des droits de l'homme des Nations Unies a déclaré, dans son Observation générale n° 34, que les restrictions aux droits numériques doivent être spécifiques au contenu. Les interdictions générales sur certains sites et systèmes ne sont pas conformes au droit international. En outre, interdire la publication de matériel uniquement basé sur sa critique du gouvernement ou de son système politique et social va à l'encontre du droit international.[340]
La restriction doit poursuivre un but légitime	• Selon l'article (3)19 du PIDCP, des limitations ne doivent être imposées que pour des raisons légitimes, telles que la protection des droits et de la réputation d'autrui, la garantie de la sécurité nationale, le maintien de l'ordre public et la promotion de la santé ou de la moralité publiques.
La restriction doit être nécessaire à des fins légitimes	• Toute limitation du droit à la liberté d'expression doit être nécessaire et proportionnée. Bien que la surveillance publique puisse être autorisée, les États doivent démontrer que les mesures sont à la fois nécessaires et proportionnées. La surveillance numérique est un acte très intrusif qui viole les droits numériques. L'approbation préalable d'une autorité judiciaire compétente est nécessaire à une surveillance numérique proportionnée. Cela signifie également que les méthodes de surveillance les moins intrusives doivent être utilisées.[341] • Par exemple, la détection automatisée des menaces, un système couramment utilisé par les forces de police pour détecter les coups de feu et identifier les scènes de crime possibles, s'est montrée capable d'identifier de manière inexacte les sons comme des coups de feu dans 89 % des cas. De nombreux services de police qui utilisaient auparavant des services de police prédictifs ont abandonné ces systèmes en raison de leur utilité et de leur précision limitées.[342]

339 Voir : UNESCO (2021). Manuel de formation mondial pour les acteurs du judiciaire : normes juridiques internationales relatives à la liberté d'expression, l'accès à l'information et la sécurité des journalistes, Module 5, p. 164, disponible sur : https://unesdoc.unesco.org/ark:/48223/pf0000378755

340 UN (2011). General Comment No 34, disponible sur : https://www2.ohchr.org/english/bodies/hrc/docs/gc34.pdf

341 Commission internationale de juristes, Regulation of Communications Surveillance and Access to Internet in Selected African States, disponible sur : https://www.kas.de/documents/275350/0/Report-on-Regulation-of-Communications-Surveillance-and-Access-to-Internet-in-Selected-African-States.pdf/66dbd47d-4d7d-2779-a595-a34e9f93cfbb?t=1639140695434

342 *Ibid.*

La vidéo suivante de l'UNESCO explique le test en trois parties concernant les limites légitimes à la liberté d'expression :

Source : https://www.youtube.com/watch?v=Wg8fVtHPDag

Droit à la vie privée et à la protection des données

« Nul ne sera l'objet d'immixtions arbitraires ou illégales dans sa vie privée, sa famille, son domicile ou sa correspondance, ni d'atteintes illégales à son honneur et à sa réputation. Toute personne a droit à la protection de la loi contre de telles immixtions ou de telles atteintes ».

- Article 17 du PIDCP

Le droit à la vie privée est essentiel pour garantir d'autres droits humains, notamment la liberté d'expression, d'opinion, d'affiliation et de réunion. Sans protection de la vie privée, il n'est souvent ni pratique ni sûr d'organiser une opposition politique, de rivaliser commercialement ou de développer des alternatives aux politiques existantes, aux récits dominants ou à l'injustice vécue. La Déclaration universelle des droits de l'homme (DUDH, article 12), le Pacte international relatif aux droits civils et politiques (PIDCP, article 17) et plusieurs autres traités internationaux et régionaux relatifs aux droits de l'homme reconnaissent le droit à la vie privée en tant que droit humain.[343] Dans un monde centré sur les données, l'importance du droit à la vie privée pour l'exercice en ligne et hors ligne d'autres droits humains, tels que la liberté d'expression et l'accès à l'information, ne cesse de croître.[344]

Depuis l'entrée en vigueur du PIDCP en 1976, les nouvelles technologies numériques ont évolué et les gouvernements et les organisations

343 Par exemple, la Convention relative aux droits de l'enfant (article 16), la Convention internationale sur la protection des droits de tous les travailleurs migrants et des membres de leur famille (article 14), la Convention relative aux droits des personnes handicapées (article 22), la Charte africaine des droits et du bien-être de l'enfant (article 10), la Convention américaine relative aux droits de l'homme (article 11) et la Convention de sauvegarde des droits de l'homme et des libertés fondamentales (Convention européenne des droits de l'homme, article 8).

344 Conseil des droits de l'homme des Nations Unies (2021). The right to privacy in the digital age. Report of the United Nations High Commissioner for Human Rights, disponible sur : https://www.ohchr.org/en/HRBodies/HRC/RegularSessions/Session48/Documents/A_HRC_48_31_AdvanceEditedVersion.docx

privées les ont le plus souvent exploitées en dehors du cadre juridique et au mépris de la vie privée. Alors que la surveillance numérique et les technologies numériques ont progressé rapidement, la loi sur la vie privée n'a pas emboîté le pas. Bien que la législation sur la protection de la vie privée au niveau des droits humains internationaux soit basée sur des principes solides et bien établis, elle n'a pas évolué ni n'a été modifiée pour répondre aux exigences de la société du XXIe siècle. L'observation générale originale de 1988 du Comité des droits de l'homme des Nations Unies sur la vie privée ne prévoyait pas le développement de nouvelles formes de communication telles que le courrier électronique et le SMS, l'émergence de capacités gouvernementales à intercepter et traiter de grandes quantités de données électroniques, ou l'explosion des sites Web de médias sociaux, pour ne citer que quelques exemples.[345]

La résolution de l'Assemblée générale des Nations Unies sur le droit à la vie privée à l'ère numérique (2020) a fait référence au « piratage et à l'utilisation illégale des technologies biométriques », comme « des actes hautement intrusifs qui violent le droit à la vie privée », interfèrent avec la liberté d'expression et d'opinion, la liberté de réunion et d'association pacifiques et la liberté de religion ou de conviction, et « peuvent contredire les principes d'une société démocratique, y compris lorsqu'ils sont entrepris de manière extraterritoriale ou à grande échelle ».[346] Un rapport du Haut-Commissaire des Nations Unies aux droits de l'homme de 2021, intitulé « Le droit à la vie privée à l'ère numérique », a appelé à un moratoire sur l'utilisation des technologies de reconnaissance faciale dans les espaces publics, jusqu'à ce que les gouvernements puissent montrer qu'il n'y a pas de problèmes substantiels liés à l'exactitude ou aux impacts discriminatoires, et que ces technologies respectent des normes strictes en matière de confidentialité et de protection des données.[347]

Confidentialité et protection des données dans le domaine numérique

La compréhension de la loi sur la protection des données et de la vie privée dans le domaine numérique nécessite une compréhension globale de la définition, de la classification et de l'émergence de la vie privée en tant que préoccupation sociale. Dans une société démocratique, le droit à la vie privée est un principe fondamental et joue un rôle crucial dans l'équilibre des pouvoirs entre le gouvernement, les entités du secteur privé qui collectent, traitent et stockent des données personnelles, et les personnes dont les données personnelles sont collectées, traitées et stockées. Dans un monde centré sur les données, l'importance du droit à la vie privée pour l'exercice en ligne et hors ligne d'autres droits humains, tels que la liberté d'expression et l'accès à l'information, augmente.[348]

345 American Civil Liberties Union (2015). Information Privacy in the Digital Age, disponible sur : https://www.aclu.org/other/human-right-privacy-digital-age

346 Conseil des droits de l'homme des Nations Unies (2021). The right to privacy in the digital age. Report of the United Nations High Commissioner for Human Rights, available at: https://www.ohchr.org/en/HRBodies/HRC/RegularSessions/Session48/Documents/A_HRC_48_31_AdvanceEditedVersion.docx

347 *Ibid.*

348 Conseil des droits de l'homme des Nations Unies (2021). The right to privacy in the digital age. Report of the United Nations High Commissioner for Human Rights, disponible sur : https://www.ohchr.org/en/HRBodies/HRC/RegularSessions/Session48/Documents/A_HRC_48_31_AdvanceEditedVersion.docx

Depuis le début de l'ère de l'information, le droit à la vie privée et la nécessité de protéger les informations ou données personnelles ont fait l'objet d'une grande attention. Nous vivons à une époque où les technologie numériques permettent un traitement de masse rentables des données personnelles en ligne, ainsi qu'un suivi des individus où qu'ils se trouvent (y compris le suivi de leurs activités en ligne). Alors qu'Internet, le partage d'informations et la collecte de données en ligne augmentent à un rythme exponentiel, les développements législatifs n'ont pas réussi à suivre le rythme et à protéger adéquatement les informations personnelles. Les gouvernements du monde entier ont commencé à adopter des instruments et des réglementations liés à la protection des données, pour protéger les droits à la vie privée de leurs citoyens.[349]

Le concept de vie privée est une constellation de principes. Le droit à la vie privée garantit un espace réservé à l'expression de soi. Ainsi, ce droit est fortement lié à la liberté d'expression. Il est de plus en plus reconnu que le droit à la vie privée joue un rôle essentiel dans la facilitation du droit à la liberté d'expression et à l'accès à l'information. Par exemple, la protection du droit à la vie privée permet aux individus de partager des points de vue anonymement, dans des circonstances où ils peuvent craindre d'être censurés pour ces points de vue, aux lanceurs d'alerte de divulguer des informations en étant protégés, et aux membres des médias et militants de communiquer en toute sécurité au-delà de la portée de la surveillance gouvernementale.[350] En outre, le droit à la vie privée protège l'intimité et la dignité. Il implique également le droit de décider de son mode de vie et le droit à l'autonomie en général. Le droit à la vie privée inclut la vie privée informationnelle, ainsi que le droit à l'accès et au contrôle de ses informations personnelles, quel que soit leur format. Ces sous-composantes de la protection de la vie privée ne sont pas exhaustives, elles servent plutôt de feuille de route pour le développement futur de mesures de protection de la vie privée dans l'environnement numérique.[351]

La frontière entre le monde en ligne et hors ligne se brouille de plus en plus. De fait, il semble que les gens vivent en permanence en ligne et hors ligne, ce qui rend plus difficile la définition de limites claires. Avec l'aide de l'IA, les organisations (privées et gouvernementales) peuvent facilement collecter, traiter et réutiliser de grandes quantités de données et d'images, y compris des données utilisateur sensibles. Les algorithmes d'IA permettent de prédire la vie personnelle des gens, comme leurs habitudes de sommeil et même leur lieu de résidence.

Les entreprises de médias sociaux prospèrent grâce à la collecte et à la commercialisation de grands volumes de données sur les utilisateurs d'Internet,

349 Media Defence (2022). Module 4: Data Privacy and Data Protection, disponible sur : https://www.mediadefence.org/ereader/publications/modules-on-litigating-freedom-of-expression-and-digital-rights-in-south-and-southeast-asia/module-4-data-privacy-and-data-protection/introduction/

350 *Ibid.*

351 American Civil Liberties Union (2015). Information Privacy in the Digital Age, disponible sur : https://www.aclu.org/other/human-right-privacy-digital-age

ce qui souligne d'autant plus la nécessité de protéger la vie privée des utilisateurs dans le monde en ligne et hors ligne, de plus en plus floue. En effet, « les gens semblent vivre dans un continuum en/hors ligne, de sorte qu'il est difficile de tracer des lignes nettes et significatives entre les deux ». L'IA facilite la collecte, le traitement et la réutilisation de quantités massives de données et d'images, encourageant les organisations (du secteur privé et du gouvernement) à collecter, conserver et traiter des données sensibles sur les utilisateurs. Les algorithmes d'IA établissent des prédictions sur la vie personnelle des gens, notamment l'endroit où ils vivent et leurs habitudes de sommeil.

Au quotidien, les traceurs GPS de nos smartphones peuvent collecter une multitude de données sur nos mouvements, même si nous n'utilisons pas activement Internet. Lorsque nous nous rendons dans des cafés, des écoles ou des établissements médicaux, ces informations peuvent être utilisées pour tirer des conclusions sur notre identité personnelle, nos intérêts, nos aspirations, nos problèmes et nos réseaux sociaux, en fonction de la durée de notre séjour et des mouvements des autres autour de nous. Ces données peuvent être assez révélatrices et avoir des implications importantes sur notre vie privée et notre sécurité. Par exemple, lorsque nous nous déplaçons en ville et que nous nous rendons dans un café, une école ou un établissement médical, le traceur GPS de notre smartphone est capable de détecter où nous sommes et combien de temps nous restons sur place, et ainsi de collecter ces données (et de les corréler avec les mouvements des autres), même si nous n'avons pas accès à Internet sur nos appareils. Des déductions significatives peuvent être faites à partir de ces données, en ce qui concerne notre identité, nos intérêts, nos aspirations, nos problèmes et nos réseaux.

De nouvelles formes peu coûteuses d'analyse et de stockage de données, associées à une connectivité numérique et en ligne améliorées (des appareils intelligents aux nanobots à l'intérieur du corps humain) et à des technologies émergentes telles que l'IA et l'IdO, ont permis aux gouvernements et aux grandes sociétés de devenir des explorateurs de données, collectant des informations sur tous les aspects des activités, du comportement et du mode de vie humains.

Les réglementations en matière de protection de la vie privée se sont pas adaptées aux nouveaux défis posés par l'environnement numérique et en ligne. De nombreux pays à travers le monde ont mis en œuvre des réglementations exigeant le consentement des personnes concernées pour utiliser et traiter leurs données personnelles en ligne, garantissant l'accès aux données personnelles par les personnes concernées et donnant le droit de faire supprimer, corriger ou transférer ces données personnelles à une autre entité.

Les lois préservant la vie privée dans l'environnement de l'IA visent à donner aux individus le droit de consulter le contenu des bases de données contenant des informations à leur sujet. Ces lois visent également à restreindre l'utilisation des informations personnelles sans le consentement de la personne concernée, sauf dans des circonstances limitées définies par la loi. En vertu de ces lois, les individus ont le droit d'accepter les conditions d'utilisation avant de télécharger une application sur leur téléphone portable ou de commencer à utiliser des logiciels gratuits, c'est-à-dire des produits et services dont le modèle économique repose sur la commercialisation de données personnelles.[352]

Les données personnelles, stockées en ligne, sont souvent traitées de nombreuses manières et à de nombreuses fins, dont certaines ne peuvent être anticipées au moment où le consentement est accordé par la personne concernée. En outre, peu d'entre nous passent en revue les conditions d'utilisation, même lorsqu'elles sont concises et affichées en gros caractères.[353] Par exemple, sur une année, il nous faudrait 76 jours pour lire toutes les politiques de confidentialité qui nous concernent.[354]

Un autre aspect de la vie privée dans l'environnement de l'IA est de l'envisager comme le « droit à la tranquillité ».[355] Il s'agit du droit à garder un espace sûr et protégé autour de notre corps, de nos pensées intimes, de nos sentiments et de notre mode de vie, lorsque nous sommes en ligne. La surveillance constante en ligne de nos actions par des capteurs, des caméras de surveillance, des assistants numériques, tels que Siri, Alexa et d'autres outils d'IA et numériques, peut avoir un impact profond sur le droit à la vie privée en tant que droit humain.[356]

352 Altshuler T. S. (2019). Privacy in a digital world, disponible sur : https://techcrunch.com/2019/09/26/privacy-queen-of-human-rights-in-a-digital-world
353 *Ibid.*
354 Popkin H. A. S. (2012). Life is too short to read privacy policies - here's statistical proof!, disponible sur : https://www.nbcnews.com/tech/tech-news/life-too-short-read-privacy-policies-heres-statistical-proof-flna297399
355 Altshuler T. S. (2019). Privacy in a digital world, disponible sur : https://techcrunch.com/2019/09/26/privacy-queen-of-human-rights-in-a-digital-world
356 *Ibid.*

Étude de cas : Enregistrement Amazon Alexa et envoi de conversations privées

Une famille de l'Oregon, aux États-Unis, a signalé que son appareil Amazon Echo avait enregistré une conversation privée qu'elle avait chez elle. Plus inquiétant encore, la conversation enregistrée a ensuite été envoyée à un contact du carnet d'adresses de la famille, un collègue de l'un des membres, à son insu et sans son consentement. L'incident a été révélé lorsque le destinataire de la conversation enregistrée a contacté la famille pour l'informer du message inhabituel. Amazon a enquêté sur l'incident et l'a attribué à une combinaison de circonstances extrêmement rare. Selon Amazon, l'appareil Echo avait interprété par erreur des parties de la conversation comme des commandes pour envoyer un message. Il s'agissait d'un cas de détection de mot de réveil « faux positif », où l'appareil a pensé à tort qu'il avait entendu le mot de réveil (probablement « Alexa ») et a commencé à enregistrer et à envoyer la conversation. Amazon a pris l'incident au sérieux et entamé des mesures pour améliorer la technologie de reconnaissance des mots de réveil, afin d'éviter de tels faux positifs. La société a également introduit une fonctionnalité qui permet aux utilisateurs d'ajouter un code PIN aux achats vocaux, afin d'éviter les commandes accidentelles via des commandes vocales. Cet incident a suscité des discussions sur la confidentialité et la sécurité des appareils à commande vocale, ce qui accentué la sensibilisation et les préoccupations des utilisateurs relatives au potentiel d'écoute clandestine. Dans l'ensemble, l'incident a mis en évidence la nécessité pour les entreprises technologiques d'améliorer continuellement les fonctionnalités de confidentialité et de sécurité des appareils à commande vocale comme Amazon Alexa. Il a également souligné l'importance de l'éducation des utilisateurs concernant les paramètres de l'appareil et les contrôles de confidentialité, pour assurer une expérience utilisateur plus sûre et sécurisée.

Source : Wolfson S. (2018). Amazon's Alexa recorded private conversation and sent it to random contact, disponible sur : https://www.theguardian.com/technology/2018/may/24/amazon-alexa-recorded-conversation

Il est important de noter que les entreprises technologiques ont pris des mesures pour répondre à ces préoccupations et améliorer la confidentialité des utilisateurs, en offrant plus de transparence, en améliorant les paramètres de confidentialité et en permettant aux utilisateurs de supprimer les enregistrements vocaux. Cependant, ces incidents soulignent la nécessité pour les utilisateurs d'être vigilants quant à leurs paramètres de confidentialité et aux risques potentiels associés aux appareils à commande vocale. Les utilisateurs doivent également connaître les pratiques de collecte et de stockage des données des assistants virtuels qu'ils utilisent et prendre des décisions éclairées quant à leur utilisation.

Profilage de l'IA

Un troisième aspect de la vie privée dans l'environnement de l'IA est le droit de s'opposer au profilage automatique, en limitant la capacité des entités commerciales ou gouvernementales à combiner des données personnelles avec des données de masse recueillies auprès d'autres personnes, pour construire des profils comportementaux à l'aide de l'IA et de l'apprentissage automatique.[357] Les outils d'IA sont utilisés pour identifier des modèles dans le comportement humain. Avoir accès aux ensembles de données corrects peut servir à déduire des aspects quotidiens profondément privés et personnels, tels que le nombre d'habitants d'un quartier susceptibles de visiter un lieu de culte spécifique, les programmes de télévision qu'ils pourraient apprécier et même, prosaïquement, leurs habitudes de sommeil.

L'utilisation de techniques d'IA peut identifier des groupes, tels que ceux qui partagent une position politique ou personnelle spécifique, et tirer des conclusions générales sur les individus, y compris sur leur santé mentale et physique. Malgré leur caractère probabiliste, les jugements et les prédictions fournis par l'IA peuvent souvent servir de base à des décisions qui ont un impact sur les droits fondamentaux des personnes. Ces problèmes sont exacerbés dans le contexte du pouvoir judiciaire, par exemple lorsque les juges s'appuient sur des décisions prises à l'aide de systèmes d'IA.[358]

L'histoire de la façon dont Target a utilisé l'analyse de données pour prédire qu'une adolescente était enceinte avant que sa famille ne le sache est un exemple bien connu de la puissance de l'analyse de données et de la modélisation prédictive dans le commerce de détail. Voici un résumé de l'affaire :

En 2012, un article du New York Times a révélé que Target, un géant américain de la vente au détail, avait développé un algorithme pour prédire les habitudes et les préférences d'achat des clients. Ils ont utilisé ces données pour envoyer des publicités ciblées et des bons de réduction aux clients. L'un des exemples les plus célèbres de cet article concernait une adolescente.

L'algorithme de Target avait identifié qu'une adolescente achetait de la lotion non parfumée, des compléments alimentaires et des boules de coton. Bien que ces achats puissent sembler sans rapport, l'algorithme a reconnu que cette combinaison de produits indiquait souvent une grossesse. L'algorithme a attribué une note de « prédiction de grossesse » à chaque cliente, en fonction de son historique d'achats. Une fois que le système obtenait une note de prédiction élevée pour une cliente, il commençait à lui envoyer des publicités et des bons de réduction liés à la grossesse et aux produits pour bébés. Dans ce cas précis, Target a commencé à envoyer à l'adolescente des bons pour des produits pour bébés tels que des couches, des lits et des vêtements pour bébés.

357 *Ibid.*

358 Conseil des droits de l'homme des Nations Unies (2021). The right to privacy in the digital age. Report of the United Nations High Commissioner for Human Rights, disponible sur : https://www.ohchr.org/en/HRBodies/HRC/RegularSessions/Session48/Documents/A_HRC_48_31_AdvanceEditedVersion.docx

Le père de la jeune fille, stupéfait de trouver ces publicités liées à la grossesse adressées à sa fille, a appelé le magasin pour s'en plaindre, car il les jugeait inappropriées. Cependant, quelques jours plus tard, il a découvert que sa fille était bien enceinte.

L'algorithme avait prédit avec précision la grossesse de la fille, en fonction de ses habitudes d'achats, avant même que sa famille ne s'en rende compte. La combinaison de produits apparemment sans rapport dans son historique d'achats, tels que la lotion non parfumée et les boules de coton, indiquait une forte probabilité de grossesse.

Ce cas illustre la manière dont l'analyse de données avancée et la modélisation prédictive peuvent être utilisées par les détaillants pour comprendre le comportement des clients et envoyer des publicités très ciblées. Bien que cela puisse être efficace à des fins de marketing, cela soulève également d'importantes questions sur la confidentialité et l'éthique de la collecte et de l'utilisation des données des clients. Il est essentiel pour les entreprises de gérer les données des clients de manière responsable et transparente, afin de maintenir la confiance avec leurs clients.

Source : Duhigg C. (2012). How Companies Learn Your Secrets, disponible sur : https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html

Pour les faits de l'affaire Nubian Rights Forum et autres c. The Attorney General, Kenya, 2021 et pour discuter de ses implications sur la surveillance numérique et la vie privée au Kenya, veuillez lire l'article de Privacy International **« Data Protection Impact Assessments and ID systems: the 2021 Kenyan ruling on Huduma Namba »**

Les outils d'IA peuvent également être utilisés pour le profilage des juges. On trouve une initiative réglementaire intéressante qui vise à préserver l'intégrité du juge et à prévenir le profilage par l'IA dans la loi française sur la programmation et la réforme de la justice (2019-2022). Dans son article 33, ce règlement vise à empêcher quiconque - mais surtout les entreprises de technologie juridique axées sur la prédiction et l'analyse des litiges - de divulguer publiquement le comportement des juges par rapport aux décisions judiciaires. Il stipule que « les données d'identité des juges et des membres du pouvoir judiciaire ne peuvent être réutilisées dans le but ou afin d'évaluer, d'analyser, de comparer ou de prédire leurs pratiques professionnelles réelles ou présumées ».[359]

De nombreux gouvernements ont commencé à numériser leurs services publics, à les mettre en ligne et à offrir des systèmes nationaux d'identification numérique (ID). En amassant de gros volumes de données personnelles, ces systèmes et bases de données numériques menacent le droit à la vie privée des citoyens. Les programmes nationaux d'identité numérique ne sont que l'un des nombreux exemples de la façon dont les droits numériques peuvent être violés par les gouvernements. Ces programmes nécessitent la collecte et le stockage de données personnelles sensibles et d'identifiants biométriques pour créer un identifiant numérique unique, afin d'améliorer la prestation des services gouvernementaux. Cependant,

359 French Law on Programming and Reform of Justice (2019-2022), disponible sur : https://www.wipo.int/wipolex/en/legislation/details/18789

il est important que les gouvernements comprennent les risques potentiels pour les utilisateurs, avant de créer des bases de données centralisées de données personnelles et biométriques. Pour prévenir les violations des droits humains et la cybersécurité, les lois doivent inclure des protections appropriées avant de déployer de tels programmes. De nombreux tribunaux nationaux et régionaux ont donné suite à des poursuites contre ces systèmes numériques intentées par des citoyens et des organisations de la société civile (OSC).

L'un de ces cas est Nubian Rights Forum et autres c. le Procureur général, Kenya, 2021, où la Haute Cour du Kenya a déclaré inconstitutionnel le Système national intégré de gestion de l'identité (NIIMS) du pays, un système d'identification numérique.[360] La Cour a déclaré qu'une évaluation de l'impact sur la protection des données aurait dû précéder le programme et qu'un cadre juridique approprié pour atténuer les risques pour la vie privée et la protection des données aurait dû être en place avant la mise en œuvre du NIIMS.[361] Ce passage met en évidence les pièges courants que les décisions de justice dans divers pays ont identifiés lors de la décision des contestations des OSC et d'autres parties prenantes contre les systèmes d'identification numérique. Dans une autre affaire, la Cour suprême mauricienne a souligné l'absence de défense adéquate contre les risques de sécurité associés à la biométrie. L'arrêt Aadhaar, en Inde, a exprimé des préoccupations concernant les bases de données centralisées, tandis que la Cour suprême des Philippines a identifié le risque de suivi individuel par le biais d'un système d'identité national. Enfin, la Haute Cour kenyane a identifié le risque d'exclusion dû à des échecs d'enregistrement biométrique et d'autres systèmes d'identité.[362]

Activité : Publicité ciblée et discrimination par les prix propulsées par des algorithmes d'IA. Les participants à la formation discutent des principaux problèmes juridiques et des droits humains impactés par la publicité ciblée et la tarification personnalisée. Quelles lois sont applicables dans ces circonstances ?

Publicité ciblée

À l'ère numérique actuelle, les algorithmes d'auto-apprentissage sont devenus partie intégrante de l'analyse des données de masse. Avec l'aide de l'IA, les entreprises privées peuvent collecter une pléthore d'informations personnelles, telles que vos habitudes de navigation, vos goûts sur les réseaux sociaux, vos données de santé et vos habitudes d'achats. Ces informations peuvent ensuite être utilisées pour créer le profil détaillé d'un individu, qui peut ensuite servir au suivi et au profilage en ligne. Cela aide les entreprises à adapter leurs publicités, leurs prix et leurs conditions contractuelles au profil spécifique du client, et à tirer parti des préjugés et de la volonté de payer du consommateur, le tout grâce aux résultats de l'économie comportementale. En outre, les informations basées sur l'IA peuvent également être utilisées pour les systèmes de notation, qui peuvent décider si un consommateur spécifique est éligible à l'achat d'un produit ou la prestation d'un service particulier. L'utilisation d'algorithmes d'auto-apprentissage dans l'analyse de données de masse permet aux entreprises privées d'obtenir un aperçu détaillé

360 Voir : https://globalfreedomofexpression.columbia.edu/cases/nubian-rights-forum-v-attorney-general.
361 UNESCO (2022). Lignes directrices à l'intention des acteurs judiciaires sur la vie privée et la protection des données, disponible sur : https://unesdoc.unesco.org/ark:/48223/pf0000381298
362 Privacy International (2022). Data Protection Impact Assessments and ID systems: the 2021 Kenyan ruling on Huduma Namba, disponible sur : https://privacyinternational.org/news-analysis/4778/data-protection-impact-assessments-and-id-systems-2021-kenyan-ruling-huduma

de sa situation personnelle, de ses comportements et de sa personnalité (achats, sites visités, likes sur les réseaux sociaux, données de santé). L'IA est utilisée dans le suivi et le profilage en ligne des personnes dont les habitudes de navigation sont collectées par des « cookies » et des empreintes numériques, puis combinées à des requêtes via des moteurs de recherche ou des assistants virtuels. Les entreprises peuvent adapter leur publicité, leurs prix et leurs conditions contractuelles au profil du client concerné et – en s'appuyant sur les conclusions de l'économie comportementale – exploiter les préjugés du consommateur et/ou sa volonté de payer. Les informations basées sur l'IA peuvent également être utilisées pour les systèmes de notation, afin de décider si un consommateur spécifique peut acheter un produit ou solliciter un service.

L'utilisation croissante de la publicité ciblée, qui repose sur le suivi et le profilage sur Internet, a soulevé des préoccupations concernant la confidentialité et la protection des données. Tout étant automatisé, les utilisateurs sont souvent incapables de donner un consentement valable. L'utilisation de l'IA pour un traitement intensif des données peut exacerber d'autres violations des droits, en particulier dans les cas où les données personnelles sont utilisées pour cibler des personnes dans des contextes tels que les demandes d'assurance ou d'emploi. Dans certains cas, les algorithmes peuvent même constituer une menace à la fois pour le droit à la vie privée et la liberté d'expression. Cela crée des problèmes croissants pour la confidentialité et la protection des données. La publicité ciblée utilise le suivi et le profilage sur Internet en fonction des intérêts attendus de la personne. Toutes ces méthodes ont empêché les utilisateurs de donner un consentement valable, car tout est automatisé. Le traitement intensif des données à l'aide de l'IA peut exacerber d'autres violations des droits, lorsque les données personnelles sont utilisées pour cibler des personnes, par exemple dans le cadre de demandes d'assurance ou d'emploi, ou lorsque les algorithmes menacent à la fois le droit à la vie privée et la liberté d'expression.[363] Par exemple, les algorithmes des médias sociaux décident du contenu du fil d'actualité d'un utilisateur et influencent le nombre de personnes qui voient et partagent des informations. Les algorithmes des moteurs de recherche indexent le contenu et déterminent ce qui apparaît en haut des résultats de recherche. Ces algorithmes menacent le pluralisme des médias et suppriment la diversité des points de vue.[364] Pour illustrer cela, en 2023, Meta a été condamnée à une amende de 390 millions d'euros par le Comité irlandais de protection des données, pour violation du RGPD. Le régulateur a allégué que l'utilisation par Meta des données personnelles sur Facebook et Instagram, en particulier pour la publicité personnalisée, n'était pas conforme au RGPD.[365]

Discrimination par le prix

À l'ère numérique, l'IA joue un rôle important en aidant les entreprises à adapter leurs offres aux clients individuels. En analysant le comportement et les préférences des consommateurs, les algorithmes d'IA peuvent estimer le prix le plus élevé qu'un client particulier est disposé ou apte à payer. Cette approche est particulièrement pertinente pour des industries telles que le crédit et l'assurance, qui fonctionnent sur des structures de coûts basées sur le risque et qui prennent en compte les caractéristiques uniques de chaque consommateur. Cependant, la question de savoir si les régulateurs devraient autoriser la discrimination par le prix dans d'autres secteurs en fonction de la capacité de paiement d'un client est complexe et controversée, et nécessite une

363 Council of Europe (2017). Study on the human rights dimensions of automated data processing techniques (in particular algorithms) and possible regulatory implications, disponible sur : https://rm.coe.int/study-hr-dimension-of-automated-data-processing-incl-algorithms/168075b94a

364 Access Now (2018). Human rights in the age of artificial intelligence, disponible sur : https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf

365 La Commission de protection des données (2023). Data Protection Commission announces conclusion of two inquiries into Meta Ireland, disponible sur https://www.dataprotection.ie/en/news-media/data-protection-commission-announces-conclusion-two-inquiries-meta-ireland

exploration et un débat plus approfondis. L'IA soutient les entreprises numériques en présentant aux consommateurs des prix individualisés et en offrant à chacun une approximation du prix le plus élevé qu'il est en mesure ou désireux de payer. Certains marchés, tels que le crédit ou l'assurance, fonctionnent sur des structures de coûts basées sur des profils de risque corrélés à des caractéristiques propres aux consommateurs individuels, ce qui suggère qu'il peut être raisonnable d'offrir des prix différents (par exemple, des taux d'intérêt) à différents consommateurs. Les régulateurs devraient-ils également autoriser la discrimination par le prix dans d'autres cas, en fonction de la capacité des différents consommateurs à payer ?[366]

Il est préoccupant de constater que les consommateurs ne savent généralement pas quand la publicité, les informations, les prix ou les conditions contractuelles ont été personnalisés en fonction de leur profil. Si un algorithme calcule une certaine note qui se traduit par un contrat qui n'est pas proposé ou seulement dans des conditions défavorables, les consommateurs ont souvent du mal à comprendre comment cette note a été générée. De plus, la complexité, l'imprévisibilité et le comportement semi-autonome des systèmes d'IA peuvent représenter des défis pour l'application de la législation sur la consommation, car il est difficile de remonter aux décisions d'un seul acteur et d'assurer la conformité juridique. Les consommateurs ne sont généralement pas conscients que la publicité, les informations, les prix ou les conditions contractuelles ont été personnalisés en fonction de leur profil. Supposons qu'un certain contrat ne soit pas conclu ou seulement proposé à des conditions défavorables en raison d'une certaine note calculée par un algorithme. Dans ce cas, les consommateurs sont souvent incapables de comprendre comment cette note a été obtenue. La complexité, l'imprévisibilité et le comportement semi-autonome des systèmes d'IA peuvent également rendre difficile l'application efficace de la législation sur la consommation, car la décision ne peut pas être attribuée à un acteur unique et sa conformité juridique ne peut être vérifiée.

Toutes ces pratiques de profilage automatisé permises par l'IA ont eu de graves implications pour la jouissance du droit à la vie privée et familiale. Les traces d'informations personnelles, telles que l'épuisement numérique sciemment ou inconsciemment produit par les téléphones portables, les ordinateurs et autres technologies, laissées dans le domaine numérique sont sans fin. La manière dont ces informations personnelles sont collectées et utilisées par des tiers constitue une énorme préoccupation pour les régulateurs.[367]

L'IA est utilisée dans le suivi et le profilage en ligne des personnes dont les habitudes de navigation sont collectées par des « cookies » et des empreintes numériques, puis combinées à des requêtes via des moteurs de recherche ou des assistants virtuels. Les applications mobiles traitent les données comportementales (telles que les données de localisation et de santé) des appareils intelligents. Cela crée des problèmes croissants pour la confidentialité et la protection des données. La publicité ciblée utilise le suivi et le profilage sur Internet en fonction des intérêts

366 Parlement européen (2019). Artificial Intelligence: Challenges for EU Citizens and Consumers, disponible sur : https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/631043/IPOL_BRI(2019)631043_EN.pdf

367 Perry W. L., McInnis B., Price C. C., Smith S., Hollywood J. S. (2013). Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations, RAND Corporation: Santa Monica, disponible sur : https://www.rand.org/pubs/research_reports/RR233.html

Pour une expérience directe du suivi en ligne, les participants à la formation doivent se rendre en ligne sur le gestionnaire de préférences publicitaires de Google à l'adresse : http://www.google.com/ads/preferences/ et examiner les marqueurs utilisés par l'entreprise pour les définir et évaluer leur précision.
Les informations suivies sont utilisées pour créer des profils numériques des utilisateurs auxquels l'accès est vendu sur le marché, y compris des échanges spécialisés, afin d'aider les annonceurs à mieux commercialiser leurs produits

attendus de la personne. L'utilisation de toutes ces méthodes a empêché les utilisateurs de donner un consentement valable, car tout est automatisé. Même si l'on peut demander le consentement des utilisateurs comme l'exige la loi, (a) ils ne comprennent pas toujours nécessairement ce qui leur est demandé ; (b) la terminologie et les conditions générales peuvent être déroutantes et s'étendre sur de nombreuses pages ; et (c) avec autant de contenu en ligne, les utilisateurs souffrent d'une surcharge d'informations.

Étude de cas : Jurisprudence sur le profilage des personnes par le biais de l'ADM

En 2018, l'autorité italienne de protection des données (Garante) a découvert qu'un contrôleur de données violait la loi nationale sur la protection des données, en offrant des tarifs personnalisés aux clients de son service d'autopartage, en fonction de leurs habitudes et caractéristiques observées. Dans la procédure administrative, le défendeur a contesté, affirmant qu'il n'y avait pas de « catégorisation » des utilisateurs du service, car les informations utilisées pour déterminer les frais n'étaient pas liées aux sujets. La Garante a rejeté les objections du défendeur, estimant qu'il était évident qu'il y avait eu traitement de données à caractère personnel en l'espèce, qu'il s'agissait d'un traitement exclusivement automatisé, et qu'il était destiné à définir le profil ou la personnalité d'une personne ou à analyser ses habitudes ou ses choix de consommation. La Cour suprême italienne (Corte Suprema di Cassazione) a confirmé cette décision en novembre 2021, ce qui a entraîné une amende administrative de 60 000 €. Dans la procédure d'appel, la Cour suprême s'est rangée du côté de la Garante, car elle a jugé que le traitement de données à caractère personnel à l'aide d'un algorithme pour déterminer un taux individuel constitue du profilage, même si les données ne sont ni stockées par le responsable du traitement ni imputables à la personne concernée.

Source : Future of Privacy Forum (2022). GDPR and the AI Act interplay: Lessons from FPF's ADM Case Law Report, disponible sur : https://fpf.org/blog/gdpr-and-the-ai-act-interplay-lessons-from-fpfs-adm-case-law-report

L'anonymisation des données ne mène pas toujours à la protection de la vie privée

La confidentialité des données est généralement protégée par l'anonymisation. Les aspects identifiables tels que les noms, les numéros de téléphone et les adresses e-mail sont supprimés. Les ensembles de données sont modifiés pour être moins précis, et du « bruit » est introduit dans les données. Cependant, une étude publiée par Nature Communications suggère que l'anonymisation ne protège pas toujours la vie privée. Les chercheurs ont développé un modèle de ML qui estime la manière dont les individus peuvent être réidentifiés à partir d'un ensemble de données anonymisées en entrant leur code postal, leur genre et leur date de naissance.

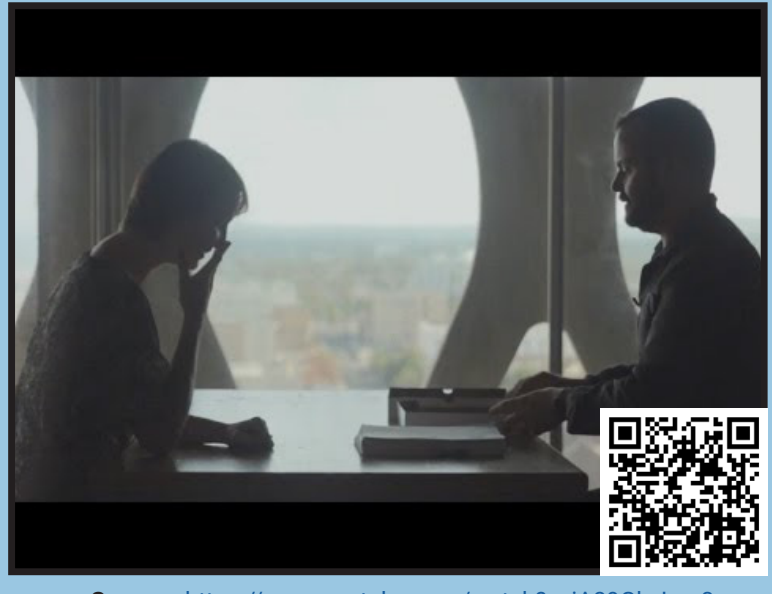
Source : Rocher L., Hendrickx J. M., de Montjoye Y.-A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models, Nature Communications, 10 (3069), disponible sur : https://www.nature.com/articles/s41467-019-10933-3

Nouveaux enjeux liés à la protection de la vie privée

La création de nouvelles données est un défi unique dans le traitement automatisé des données personnelles. Il est souvent possible de combiner des données personnelles, ce qui conduit à la création d'une deuxième, voire d'une troisième génération de données sur une personne en particulier. Par rapport à un ensemble de données beaucoup plus important, deux informations apparemment sans rapport pourraient « se reproduire » et engendrer de nouvelles données, à l'insu de la personne concernée. Des questions importantes sont soulevées concernant les concepts de consentement, d'ouverture et d'autonomie personnelle.[368] Questions qui méritent une attention particulière : Dans quelle mesure les personnes auront-elles un contrôle sur les informations collectées à leur sujet ? Compte tenu de leur intérêt dans la fourniture de données personnelles à des fins de formation au ML, les individus devraient-ils avoir le droit d'utiliser le modèle ou, au moins, savoir à quoi il sert ? Les systèmes de ML à la recherche de données pourraient-ils violer par inadvertance la vie privée des personnes si, par exemple, l'analyse du génome d'un membre d'une famille révélait des données sur la santé d'autres membres de la famille ?[369]

Point de discussion (10-15 minutes) : « Le pouvoir de la vie privée (1/5) : Internet sait-il où vous vivez ? »

Les participants à la formation regardent la vidéo produite par The Guardian, « The power of privacy (1/5) : Does the internet know where you live ? » et discutent de la façon dont la notion de vie privée a changé dans le domaine numérique et de l'impact que cela a eu sur leur travail. Ils discutent également des exemples de leurs juridictions respectives.



Source : https://www.youtube.com/watch?v=iA89GhyLao8

368 Committee of Experts on Internet Intermediaries (MSI-NET) (2018). Algorithms and Human Rights: Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications, Étude du Conseil de l'Europe, DGI/2017/12, disponible sur : https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5

369 Parlement européen (2020). The ethics of artificial intelligence: Issues and initiatives, disponible sur : https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2020)634452

Tous ces défis ont été exacerbés dans le secteur public. Selon l'ONG Access Now, avec l'expansion d'Internet et la croissance des nouvelles technologies, la surveillance gouvernementale a augmenté et l'IA permet des capacités de surveillance plus intrusives que jamais. Même s'il n'existe actuellement aucun système de reconnaissance faciale gouvernemental complètement centralisé, certains pays ont tenter de déployer davantage de caméras de vidéosurveillance dans les espaces publics et de centraliser leurs systèmes de reconnaissance faciale.[370] La moitié des adultes américains sont maintenant dans des bases de données de reconnaissance faciale des forces de l'ordre.[371] L'utilisation de ces technologies constitue une menace pour l'anonymat, et la crainte d'être observé peut empêcher l'exercice d'autres droits, tels que la liberté d'association. Les groupes démographiques défavorisés, qui sont déjà sous le contrôle fréquent des forces de sécurité, seraient les plus directement menacées par les effets négatifs de la surveillance alimentée par l'IA. En outre, étant donné que la surveillance de l'ensemble de la population 24 heures sur 24, sept jours sur sept n'est ni essentielle ni proportionnelle à l'objectif de sécurité publique ou de prévention de la criminalité, elle violerait certainement le droit à la vie privée.[372]

370 AccessNow (2018) AI and human rights, disponible sur : https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf
371 *Ibid.*
372 *Ibid.*

Études de cas

Le cas du système de reconnaissance faciale SARI Real Time, Italie

L'autorité italienne de protection des données (Garante) a publié un avis sur le système Sari Real Time présenté pour examen au ministère de l'Intérieur du pays, affirmant que si elle était utilisée comme prévu, la technologie « établirait un type de surveillance de masse ». Sari, qui n'est pas encore opérationnel, est un système de reconnaissance faciale qui, à l'aide de plusieurs caméras installées dans des zones géographiques spécifiées, analyserait les visages des individus filmés en temps réel et les comparerait à une base de données déjà prête contenant jusqu'à 10 000 visages. Sari serait mise en œuvre « lorsqu'il y a un besoin de technologie de reconnaissance faciale pour aider les forces de police dans la gestion de l'ordre et de la sécurité publique, ou en réponse aux seules exigences de la police judiciaire ».

La Garante a déclaré que Sari « effectuerait un traitement automatisé à grande échelle qui pourrait inclure les personnes présentes aux manifestations politiques et sociales qui ne font pas l'objet d'une "attention particulière" de la police ». En outre, le fait que « l'identification d'une personne serait accomplie par le traitement des données biométriques de toutes les personnes présentes dans l'espace surveillé » entraînerait une « transition de la surveillance ciblée d'individus spécifiques à la perspective d'une surveillance universelle ». La Garante a déterminé que le ministère n'avait pas clarifié le fondement juridique sur lequel il mènerait de telles actions. Il a été déclaré qu '« un cadre réglementaire efficace doit prendre en compte tous les droits et libertés en jeu, et identifier les scénarios dans lesquels l'utilisation de tels systèmes est autorisée, sans laisser une grande marge de manœuvre aux utilisateurs ».

Source : DigWatch, Italian data protection authority: Sari facial recognition system proposed by Ministry of Interior could lead to mass surveillance, disponible sur : https://dig.watch/updates/italian-data-protection-authority-sari-facial-recognition-system-proposed-ministry-interior

Utilisation de la technologie de reconnaissance faciale en direct à Buenos Aires, Argentine

Entre 2019 et 2022, une technologie de reconnaissance faciale en direct a été mise en œuvre à Buenos Aires, la capitale argentine, pour aider les forces de sécurité à identifier les criminels potentiels recherchés dans la base de données nationale des fugitifs du pays. Le système s'appuyait sur des images en direct des systèmes de vidéosurveillance disposés dans toute la ville, notamment dans les trois principales gares ferroviaires et le réseau de transport souterrain, utilisé par plus de 1,3 million de passagers chaque jour. Cependant, en avril 2022, une ordonnance du tribunal a été adoptée pour suspendre temporairement l'utilisation de la technologie, en raison d'allégations de perquisitions non autorisées. En septembre 2022, un tribunal municipal a statué que les conditions actuelles de fonctionnement du système étaient inconstitutionnelles, ce qui devrait encore prolonger la suspension du système de reconnaissance faciale. Selon l'Association argentine pour les droits civils (Asociación por los Derechos Civiles - ADC), la technologie de reconnaissance faciale a été mise en œuvre non seulement dans la capitale, mais aussi dans d'autres régions, notamment les provinces de Cordoue, Salta et Mendoza, ainsi que dans le comté de Tigre, à Buenos Aires. Il a été signalé qu'il est également prévu de déployer la technologie dans la province de Santa Fe. Ces informations étaient exactes au début de l'année 2021.

Utilisation de la technologie de reconnaissance faciale au Brésil

L'utilisation de la technologie de reconnaissance faciale est assez répandue au Brésil, avec des déploiements identifiés dans 30 villes, depuis 2019. Cette technologie est utilisée à diverses fins, notamment la prévention de la fraude dans la distribution des avantages sociaux. Elle a été utilisée pour vérifier l'identité des bénéficiaires de subventions de transport public dans de nombreuses villes brésiliennes, et suivre les exigences de fréquentation scolaire pour les programmes de transferts monétaires, dans l'État de Pernambuco. Cependant, la technologie de reconnaissance faciale a également été déployée à des fins de marketing, telles que le placement de publicités devant les passagers dans le métro de São Paulo, à l'aide de techniques de détection des émotions très controversées. Ce projet a finalement été annulé après qu'un tribunal local a déclaré que la collecte de données sur les passagers du métro ne répondait pas aux exigences minimales de consentement.

L'Argentine et le Brésil sont deux systèmes fédéraux soumis à la coexistence complexe de lois municipales, fédérales et étatiques. Cela conduit souvent à une mosaïque de réglementations, avec des normes et des garanties variables qui peuvent être assez déroutantes. Cette complexité a conduit à des difficultés à justifier la légalité des déploiements de reconnaissance faciale. En Argentine et au Brésil, les gouvernements locaux ont mis en œuvre un mélange de législation municipale et de propositions réglementaires au niveau de l'État qui sont souvent en deçà des normes énoncées dans leurs constitutions respectives, les traités internationaux relatifs aux droits humains et les lois fédérales.

Source : Chatam House (2022). Regulating facial recognition in Latin America, Policy lessons from police surveillance in Buenos Aires and São Paulo, disponible sur : https://www.chathamhouse.org/2022/11/regulating-facial-recognition-latin-america/03-facial-recognition-rollouts-trends-buenos

Classification de la protection des données en tant que droit indépendant

La classification de la protection des données en tant que droit indépendant a été un point de discorde dans les tribunaux internationaux et le monde universitaire. Elle découle du fait que la protection des données, en tant que question réglementaire, émane en partie des réglementations, normes et préoccupations en matière de protection de la vie privée, et a évolué en de nouveaux ensembles d'obligations imposées aux autorités publiques et aux entités commerciales, pour fournir aux individus un contrôle sur les informations qui les concernent, ainsi que les moyens d'exercer ce contrôle — accès à ces informations, confirmation de leur existence, correction de données incorrectes, etc.

Cependant, la protection des données s'étend au-delà des préoccupations en matière de confidentialité. Il peut y avoir d'importantes préoccupations en matière de protection des données, lorsque les considérations de confidentialité ne sont pas pertinentes ou secondaires, comme illustré ci-dessous, dans la section qui traite des principes de protection des données.[373] La protection des données s'appuie sur le droit à la vie privée, mais englobe également d'autres droits des personnes concernées vis-à-vis du gouvernement et des grandes entreprises qui collectent, traitent et stockent des données personnelles, tels que le droit d'être informé, le droit d'accès aux données personnelles, le droit à l'oubli, le droit à la rectification, le droit à la portabilité des données, le droit de s'opposer au traitement et les droits liés à la prise de décision automatisée et au profilage.[374]

De nombreux pays à travers le monde reconnaissent la protection des données comme un droit fondamental. La protection des données personnelles est incorporée en tant que droit indépendant dans divers statuts, notamment la Charte des droits fondamentaux de l'Union européenne (article 8). En outre, il a été récemment reconnu comme tel par la Cour suprême brésilienne. De même, dans une affaire récente (Justice K. S. Puttaswamy (Retd.) c. Union of India [375]), la Cour suprême indienne a affirmé que la vie privée était un droit fondamental.[376]

Droits à la protection des données liés à la prise de décision automatisée et au profilage

Dans de nombreuses juridictions, les personnes concernées ont des droits liés à la prise de décision automatisée et au profilage. Cela couvre diverses techniques de profilage, qui peuvent impliquer l'évaluation de caractéristiques personnelles spécifiques liées à une personne qui évaluent ou prévoient un comportement lié à la performance au travail, à la situation financière, à la santé, aux préférences personnelles, aux loisirs, à la fiabilité, à la conduite ou à la localisation. Le droit d'être exempté de la prise de décision automatisée est généralement garanti aux personnes concernées lorsque ces décisions ont un impact important sur leur vie. Toutefois, ces droits ne s'appliquent pas aux décisions partiellement automatisées. Ils ne garantissent pas non plus nécessairement que, dans la pratique, une personne affectée peut facilement détecter si elle a été traitée de manière inégale par rapport aux autres et, le cas échéant, si un tel traitement différencié constituait une discrimination et était donc illégal. La personne concernée a la liberté de renoncer à certains de ses droits en consentant à des pratiques spécifiques qui constitueraient autrement une violation des droits, renonçant ainsi aux protections que ceux-ci offrent.

373 *Ibid.*
374 *Ibid.*
375 Status as Fundamental Right (2017). Justice K.S. Puttaswamy (Retd.) v. Union of India, disponible sur : https://privacylibrary.ccgnlud.org/case/justice-ks-puttaswamy-ors-vs-union-of-india-ors
376 UNESCO (2018). Legal Standards on Freedom of Expression, Toolkit for the Judiciary in Africa, disponible sur : https://unesdoc.unesco.org/ark:/48223/pf0000366340.

Par exemple, il existe un risque important que les droits à la protection des données soient trop facilement abandonnés par les titulaires de droits individuels, à une époque de mise en réseau fondée sur un modèle commercial de « services gratuits » : en échange d'un accès « gratuit » aux services numériques et de l'efficacité et de la commodité qu'ils offrent, les individus échangeront volontiers leurs données personnelles.[377] Par ailleurs, les principes fondamentaux de la protection des données comprennent des obligations incontournables imposées aux responsables du traitement des données, qui ne peuvent être levées par les titulaires de droits individuels, notamment les principes de légalité du traitement, de spécification de la finalité et de minimisation des données. Cela offre une protection plus systématique et plus robuste des valeurs fondamentales sous-jacentes et des intérêts collectifs que les régimes de protection des données cherchent à protéger.[378]

377 Committee of Experts on Internet Intermediaries (MSI-AUT) (2019). A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework, Council of Europe Study, DGI/2019/05, disponible sur : https://rm.coe.int/a-study-of-the-implications-of-advanced-digital-technologies-including/168096bdab

378 *Ibid.*

Activité : Les participants à la formation lisent les études de cas ci-dessous et discutent de la manière dont les lois sur la protection des données sont appliquées dans leurs juridictions, en notant les cas renommés et en les comparant aux cas du RGPD ci-dessous. Comment une affaire similaire serait-elle jugée et tranchée dans votre juridiction ? Quelles seraient les lois applicables ?

Violations de la vie privée par Meta dans l'UE

Après avoir découvert que les informations personnelles des utilisateurs de Facebook ont été publiées sur un forum de pirates en ligne, Meta, le propriétaire de Facebook, a été condamné à une amende de 265 millions d'euros par le régulateur irlandais, la Commission de protection des données, pour violation des lois sur la protection des données. Les informations divulguées comprenaient les noms complets, les coordonnées, les dates de naissance et les villes de résidence des utilisateurs de Facebook, en 2018 et 2019.

Meta a reconnu que les informations avaient été récupérées à l'aide de technologies destinées à aider les individus à identifier des amis via des numéros de téléphone. Facebook a été pénalisé pour « non-application de la protection des données dès la conception et par défaut », conformément au RGPD. L'amende aurait pu être évitée si cette fonctionnalité avait été conçue de manière plus sécurisée.

Source : Satariano A. (2022). Meta Fined $275 Million for Breaking E.U. Data Privacy Law, disponible sur : https://www.nytimes.com/2022/11/28/business/meta-fine-eu-privacy.html

Violations de la vie privée par Google, dans l'UE

Le 6 janvier 2022, l'autorité française de protection des données (CNIL) a infligé à Google Ireland une amende de 90 millions d'euros. L'amende concerne la manière dont les processus de consentement aux cookies de YouTube sont mis en œuvre par Google Europe. L'amende de Google Ireland était l'une des deux sanctions prononcées dans la même affaire ; l'autre a été prononcée contre Google LLC de Californie (qui exploite Google Search).

Selon la CNIL, Google aurait dû permettre aux utilisateurs de YouTube de rejeter facilement les cookies. YouTube place des cookies sur les appareils à des fins de marketing, pour suivre les activités en ligne. Il est simple d'accepter les cookies sur YouTube, mais plus difficile de les rejeter. La CNIL a observé que le rejet des cookies nécessitait de nombreux clics, mais que l'acceptation des cookies n'en nécessitait qu'un seul. En vertu du RGPD, le consentement doit être « volontaire » : si une offre peut être acceptée en un seul clic, il doit être possible de la rejeter de la même manière.

La CNIL a justifié la sanction relativement lourde en citant le grand nombre d'utilisateurs de YouTube et les énormes revenus de Google sur le site.

Source : Lomas N. (2022). France spanks Google $170M, Facebook $68M over cookie consent dark patterns, disponible sur https://techcrunch.com/2022/01/06/cnil-facebook-google-cookie-consent-eprivacy-breaches/.

Restrictions légitimes au droit à la vie privée

Le PIDCP (article 2) oblige les États parties au PIDCP à « respecter et garantir » sans discrimination les droits énumérés dans le Pacte, pour tous les individus se trouvant sur leur territoire et relevant de leur juridiction. Les droits à la protection de la vie privée ne sont pas absolus. Dans de nombreuses juridictions, les organismes d'application de la loi sont exemptés de la législation sur la confidentialité des données.[379] Les gouvernements peuvent légitimement perturber la vie privée d'une personne dans certaines circonstances spécifiées par la loi, telles que des situations d'urgence ou des menaces à la sécurité nationale. Toute limitation des droits énumérés dans le PIDCP doit être autorisée en vertu des dispositions pertinentes du PIDCP. Les gouvernements doivent justifier leurs actions de surveillance et démontrer que toute atteinte à la vie privée est établie dans des lois et des règlements clairs et précis, nécessaires[380] pour atteindre les objectifs légitimes du gouvernement et proportionnels à la réalisation de ces objectifs limités. Une institution judiciaire ou administrative indépendante, impartiale et compétente doit superviser les actions de surveillance des organismes d'application de la loi. De plus, les responsables gouvernementaux et autres doivent être tenus responsables des fautes et des erreurs.[381]

Selon le Haut-Commissariat des Nations Unies aux droits de l'homme, les activités de surveillance de l'État doivent respecter la loi. Les exceptions à la surveillance numérique doivent être limitées et fondées sur les principes de nécessité et de proportionnalité, pour assurer une confidentialité adéquate des données dans toutes les branches du gouvernement.[382] Les exigences minimales suivantes doivent régir l'adoption de lois spécifiques à la surveillance :

- La loi doit être accessible au public et suffisamment spécifique. Elle doit définir précisément la portée du pouvoir discrétionnaire de surveillance accordé à l'organisme gouvernemental et le mode de surveillance. La loi doit également décrire la nature de l'infraction et la catégorie de personnes pouvant faire l'objet d'une surveillance. Les références non spécifiques à la « sécurité nationale » ou à la « santé publique » ne sont pas considérées comme des justifications spécifiques et légitimes, car elles sont vagues et générales. La surveillance doit être fondée sur des soupçons raisonnables, et toute décision autorisant la surveillance doit être suffisamment ciblée. La loi doit définir précisément les compétences de l'institution habilitée à effectuer la surveillance numérique.
- En ce qui concerne son champ d'application, le cadre juridique de la surveillance doit également inclure les demandes de surveillance du gouvernement aux entreprises. Le cadre juridique doit également inclure l'accès aux informations détenues de manière extraterritoriale et l'échange d'informations avec d'autres États. La loi doit explicitement

379 UNESCO (2022). Guidelines for Judicial Actors on Privacy and Data Protection, disponible sur : https://unesdoc.unesco.org/ark:/48223/pf0000381298

380 La composante de nécessité du critère des restrictions est la plus difficile et la plus litigieuse. Elle implique divers facteurs dans diverses juridictions internationales. Deux facteurs clés pour déterminer la nécessité sont (i) la restriction doit répondre à un besoin social urgent, et (ii) les justifications de la restriction doivent être suffisantes et pertinentes. Voir : Icelandic Human Rights Centre, https://www.humanrights.is/en/human-rights-education-project/comparative-analysis-of-selected-case-law-achpr-iachr-echr-hrc/the-right-to-freedom-of-opinion-and-expression/permissible-limitations. Voir aussi : Australian Human Rights Commission, Permissible Limitations on Rights, https://humanrights.gov.au/our-work/rights-and-freedoms/permissible-limitations-rights

381 Icelandic Human Rights Centre, https://www.humanrights.is/en/human-rights-education-project/comparative-analysis-of-selected-case-law-achpr-iachr-echr-hrc/the-right-to-freedom-of-opinion-and-expression/permissible-limitations. Voir aussi : ONU (2018). The right to privacy in the digital age. Report of the United Nations High Commissioner for Human Rights, disponible sur : https://documents-dds-ny.un.org/doc/UNDOC/GEN/G18/239/58/PDF/G1823958.pdf?OpenElement

382 Conseil des droits de l'homme des Nations Unies (2018). The right to privacy in the digital age. Report of the United Nations High Commissioner for Human Rights, disponible sur : https://documents-dds-ny.un.org/doc/UNDOC/GEN/G18/239/58/PDF/G1823958.pdf?OpenElement

établir une structure pour assurer la responsabilité et la transparence au sein des organisations gouvernementales chargées de la surveillance.

- Les pouvoirs de surveillance ne peuvent être justifiés que s'ils sont strictement nécessaires pour atteindre un objectif légitime et s'ils satisfont à l'exigence de proportionnalité. La portée de la surveillance doit être limitée à la prévention ou à l'enquête sur les infractions ou les menaces les plus graves. La durée de la surveillance doit être maintenue au minimum absolu requis pour atteindre l'objectif spécifié. Sur la base de la stricte nécessité et de la proportionnalité, la loi doit contenir des règles strictes pour l'utilisation et le stockage des données collectées, et définir précisément les circonstances dans lesquelles les données collectées et stockées doivent être effacées. Les mêmes règles de légalité, de stricte nécessité et de proportionnalité doivent s'appliquer à l'échange de renseignements.[383]
- Lorsque les gouvernements envisagent le piratage ciblé, ils doivent procéder avec une extrême prudence, en recourant à de telles mesures uniquement dans des circonstances exceptionnelles, pour enquêter ou prévenir les infractions ou les menaces les plus graves, et avec la participation du pouvoir judiciaire. La conception des opérations de piratage doit être limitée, en restreignant l'accès à des cibles et à des catégories d'informations spécifiques. Les États ne doivent pas obliger des entités privées à participer à des opérations de piratage, car cela compromettrait la sécurité de leurs propres produits et services. Le décryptage obligatoire ne peut être autorisé qu'au cas par cas, avec un mandat et la préservation des droits à une procédure particulière.[384]

Les mesures de surveillance, telles que les demandes de données de communication des entreprises et le partage de renseignements, doivent être autorisées, examinées et supervisées par des organismes indépendants à tous les stades, y compris lorsqu'elles sont initialement ordonnées, pendant leur exécution et lorsqu'elles sont résiliées.[385] L'organisme indépendant autorisant des mesures de surveillance particulières, de préférence une autorité judiciaire, doit s'assurer qu'il existe des preuves suffisantes d'une menace et que la surveillance proposée est ciblée, strictement nécessaire et proportionnée, avant d'autoriser (ou de rejeter) les mesures de surveillance ex ante.

L'organisme indépendant autorisant des mesures de surveillance particulières, de préférence une autorité judiciaire, doit s'assurer qu'il existe des preuves claires d'une menace suffisante et que la surveillance proposée est ciblée, strictement nécessaire et proportionnée, avant d'autoriser (ou de rejeter) ex ante les mesures de surveillance.[386]

Les cadres de surveillance comprennent les organismes administratifs, judiciaires ou parlementaires. Les organes de contrôle doivent être indépendants des autorités de surveillance et dotés de l'expertise, des compétences et des ressources nécessaires.

383 Conseil des droits de l'homme (2013). Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, disponible sur : https://www.ohchr.org/Documents/HRBodies/HRCouncil/RegularSession/Session23/A.HRC.23.40_EN.pdf

384 Conseil des droits de l'homme des Nations Unies (2015). Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, disponible sur : https://documents-dds-ny.un.org/doc/UNDOC/GEN/G15/095/85/PDF/G1509585.pdf?OpenElement

385 Pacte international relatif aux droits civils et politiques (2015). Concluding observations on the fifth periodic report of France, disponible sur : https://tbinternet.ohchr.org/_layouts/15/TreatyBodyExternal/Download.aspx?symbolno=CCPR%2FC%2FFRA%2FCO%2F5&Lang=en

386 Agence européenne des droits fondamentaux (2017). Surveillance by Intelligence Services: Fundamental Rights Safeguards and Remedies in the EU. Volume II: Field Perspectives and Legal Update, disponible sur : https://fra.europa.eu/sites/default/files/fra_uploads/fra-2017-surveillance-intelligence-services-vol-2_en.pdf

Sur le plan institutionnel, les règles doivent différencier et séparer les fonctions d'autorisation et de surveillance. En plus des évaluations périodiques des capacités de surveillance et des progrès technologiques, les organismes de surveillance indépendants doivent enquêter et surveiller les activités de ceux qui effectuent la surveillance et accèdent à ses produits.[387] Les organismes effectuant la surveillance doivent être tenus de fournir toutes les informations nécessaires à une surveillance efficace sur demande, de soumettre des rapports réguliers aux organismes de surveillance et de tenir des registres de toutes les mesures de surveillance. En outre, les processus de surveillance doivent être ouverts et soumis à un examen public approprié, et les décisions des organismes de surveillance doivent faire l'objet d'un appel ou d'un examen indépendant.[388]

Principe de transparence : Une discussion et un examen ouverts sont essentiels pour comprendre les avantages et les contraintes des techniques de surveillance. Par conséquent, les autorités de l'État et les organes de surveillance doivent également s'engager dans l'information publique sur les lois, politiques et pratiques existantes en matière de surveillance et d'interception des communications, ainsi que d'autres formes de traitement des données personnelles.[389] L'agence de surveillance doit expliquer la limitation du droit à la vie privée à ceux qui ont été la cible de la surveillance. De plus, les personnes soumises à la surveillance doivent avoir le droit de modifier et de supprimer les informations personnelles inutiles, si elles ne sont plus nécessaires pour les enquêtes en cours ou à venir.[390]

En principe, pour être légales, les restrictions au droit à la vie privée par le biais du cadre national des droits humains, de la protection des données, de la cybersécurité, de la cybercriminalité et de la surveillance numérique ou des lois et politiques relatives aux TIC doivent respecter certaines normes minimales du droit international des droits humains. Ces normes figurent dans la résolution de l'Assemblée générale des Nations Unies sur le droit à la vie privée à l'ère numérique de 2014[391], dans le rapport de 2014 du Rapporteur spécial sur la promotion et la protection des droits de l'homme et des libertés fondamentales dans la lutte antiterrorriste[392] et dans le rapport du Haut-Commissariat des Nations Unies aux droits de l'homme à la vie privée à l'ère numérique.[393] Selon ces normes[394], pour être légales, les restrictions au droit à la vie privée imposées par les gouvernements doivent être les suivantes :

- **Imposées uniquement à des fins légitimes :** En ce qui concerne le droit à la vie privée, la surveillance numérique ne doit être autorisée que dans la poursuite des objectifs nationaux les plus vitaux. La restriction doit être essentielle pour atteindre un but légitime, proportionnel à l'objectif et au choix le moins invasif possible. En outre, il doit être démontré que la restriction apportée au droit (telle qu'une atteinte à la vie privée pour sauvegarder la sécurité nationale

387 Voir Cour européenne des droits de l'homme, Kennedy c. Royaume-Uni, requête n 26839/05, arrêt du 18 mai 2010.
388 https://www.cipil.law.cam.ac.uk/projects/human-rights-big-data-and-technology-hrbdt-project
389 Conseil des droits de l'homme des Nations Unies (2009). Report of the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism, disponible sur : https://daccess-ods.un.org/tmp/9699321.3891983.html
390 Conseil des droits de l'homme des Nations Unies (2017). Report of the Special Rapporteur on the right to privacy, disponible sur : https://daccess-ods.un.org/tmp/2525206.50625229.html
391 Conseil des droits de l'homme des Nations Unies (2014). The right to privacy in the digital age: report of the Office of the United Nations High Commissioner for Human Rights, disponible sur : https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Fwww.ohchr.org%2Fsites%2Fdefault%2Ffiles%2FDocuments%2FIssues%2FDigitalAge%2FA-HRC-27-37_en.doc&wdOrigin=BROWSELINK%20
392 Voir : https://www.ohchr.org/en/special-procedures/sr-terrorism
393 Conseil des droits de l'homme des Nations Unies (2021). The right to privacy in the digital age, Report of the United Nations High Commissioner for Human Rights, disponible sur : https://www.ohchr.org/en/HRBodies/HRC/RegularSessions/Session48/Documents/A_HRC_48_31_AdvanceEditedVersion.docx
394 Il convient de noter que ces normes ne sont pas universellement acceptées par tous les gouvernements. Nombre d'entre eux interprètent différemment les dispositions du PIDCP. Par exemple, les États-Unis ont historiquement noté (voir page 235, disponible sur : https://2017-2021.state.gov/wp-content/uploads/2019/10/2018-Digest-Final-Draft.pdf#page=235) que l'article 19 du PIDCP n'impose pas de norme de légalité, de nécessité et de proportionnalité - seulement que la surveillance ne peut être illégale ou arbitraire.

ou le droit à la vie d'autrui) peut raisonnablement atteindre l'objectif visé. La charge incombe aux autorités qui tentent de restreindre le droit de démontrer que la restriction sert un but légitime.

➔ **Légales :** Les limites du droit à la vie privée doivent être énoncées clairement et sans ambiguïté dans la loi et doivent être fréquemment revues pour s'assurer que les protections et les garanties de la vie privée suivent le rythme des développements rapides de la technologie numérique.[395] Selon le rapport du Haut-Commissariat des Nations Unies aux droits de l'homme, « Le droit à la vie privée à l'ère numérique » : « L'ingérence autorisée par le droit national peut néanmoins être "illégale", si ce droit national est en conflit avec les dispositions du Pacte international relatif aux droits civils et politiques ».[396]

➔ **Respectueuses du principe de non-discrimination dans leur conception et leur application :** Les limites du droit à la vie privée ne doivent pas discriminer les groupes vulnérables.

➔ **Nécessaires et proportionnées :** La surveillance numérique est un acte très intrusif qui viole le droit à la vie privée. L'approbation préalable d'une autorité judiciaire compétente est nécessaire à une surveillance numérique proportionnée. Cela signifie également que les méthodes de surveillance les moins intrusives doivent être utilisées.[397]

Les gouvernements limitent le droit à la vie privée pour les raisons suivantes :

- Sécurité nationale
- Sécurité publique
- Bien-être économique national
- Protection des droits et libertés d'autrui
- Prévention de l'atteinte à l'ordre public ou de la criminalité
- Protection de la santé ou de la moralité[398]

395 MISA Zimbabwe, Konrad Adenauer Stiftung (2021). Cybersecurity and Cybercrime Laws in the SADC Region: Implications on Human Rights, disponible sur : https://fdocuments.net/document/cybersecurity-and-cybercrime-laws-in-the-sadc-region.html?page=3

396 Conseil des droits de l'homme des Nations Unies (2014). The right to privacy in the digital age: report of the Office of the United Nations High Commissioner for Human Rights, disponible sur : https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Fwww.ohchr.org%2Fsites%2Fdefault%2Ffiles%2FDocuments%2FIssues%2FDigitalAge%2FA-HRC-27-37_en.doc&wdOrigin=BROWSELINK%20

397 Commission internationale de juristes, Regulation of Communications Surveillance and Access to Internet in Selected African States, disponible sur : https://www.kas.de/documents/275350/0/Report-on-Regulation-of-Communications-Surveillance-and-Access-to-Internet-in-Selected-African-States.pdf/66dbd47d-4d7d-2779-a595-a34e9f93cfbb?t=1639140695434

398 Voir : https://africaninternetrights.org/en/node/2558#:~:text=This%20advocacy%20toolkit%20provides%20an%20overview%20of%20the,the%20formulation%20and%20implementation%20of%20data%20protection%20frameworks.

Approfondissement : Approche fondée sur les droits de l'homme (HRBA) pour évaluer l'impact de la réglementation sur le droit à la vie privée dans l'environnement numérique

Les pays qui adoptent des lois sur la cybercriminalité, la cybersécurité et la protection des données doivent suivre la HRBA pour rédiger la réglementation numérique. Une HRBA est basée sur les principes dérivés des traités internationaux et régionaux, et place les droits humains au centre de la formulation des politiques et de la rédaction législative. Les éléments fondamentaux de cette approche sont la participation, la responsabilité et la transparence, la non-discrimination et l'égalité, l'autonomisation des titulaires de droits et la légalité. La réglementation de la surveillance numérique doit être sans ambiguïté quant aux agences habilitées à effectuer la surveillance, juger les demandes de surveillance, aux tests juridiques qu'un tribunal doit appliquer aux demandes et aux sanctions légales qui s'appliquent à la surveillance non autorisée.[399] Les avocats et groupes de défense et les OSC qui travaillent dans le domaine de la vie privée numérique doivent recourir à la HRBA comme un outil, pour évaluer si les restrictions imposées au droit à la vie privée par le gouvernement sont légitimes, légales, conformes au principe de non-discrimination dans leur conception et leur application, et nécessaires et proportionnées.[400]

Une HRBA doit impliquer une évaluation de la réglementation numérique nationale par rapport aux Principes internationaux sur l'application des droits humains à la surveillance des communications.[401] L'illustration ci-dessous décrit les principaux domaines sur lesquels ces principes se concentrent :

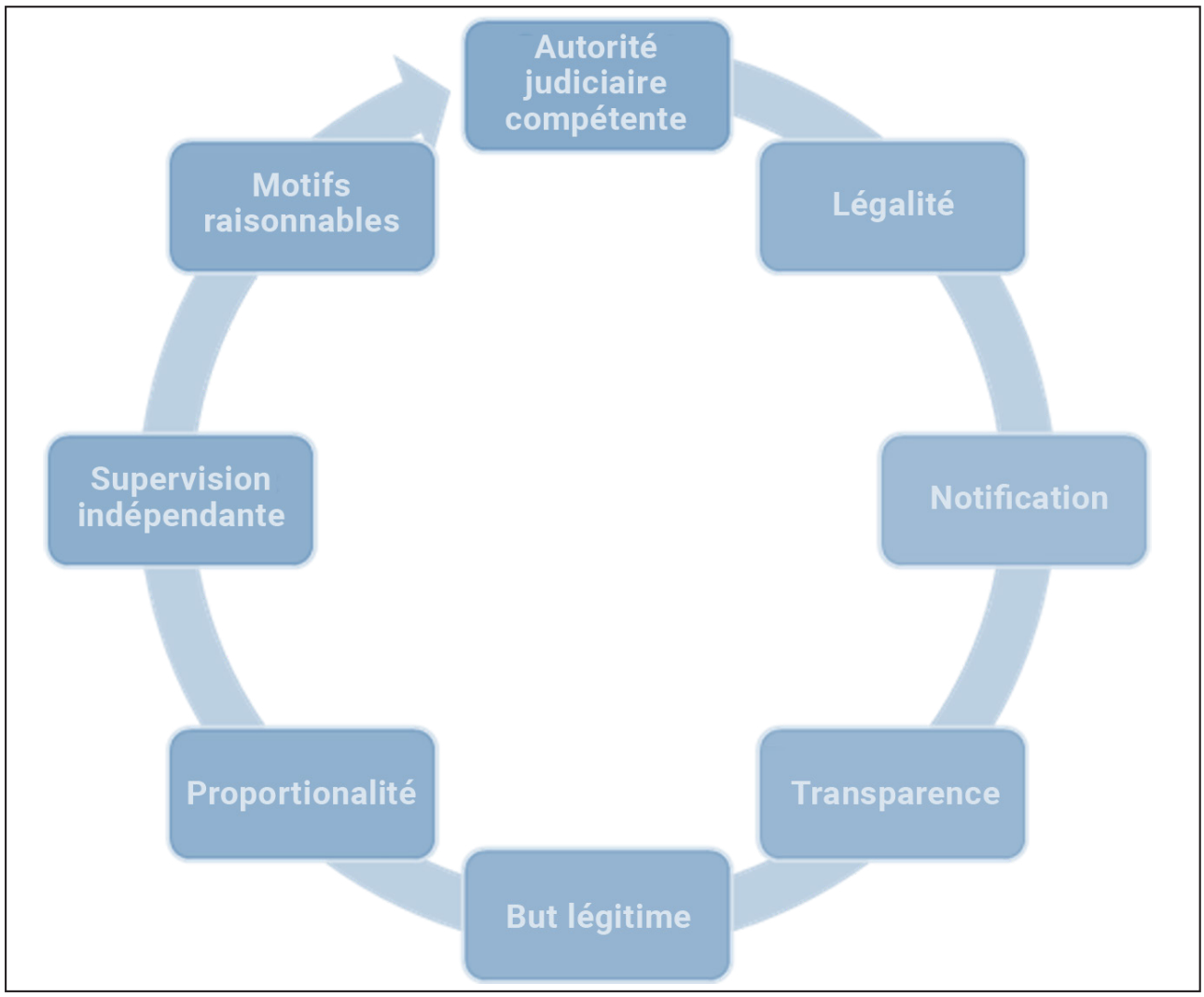
- ➔ Autorisation préalable de surveillance par une autorité judiciaire compétente : Existe-t-il un juge ayant une expertise en matière de technologie numérique et de droits humains capable d'évaluer et d'autoriser les demandes de surveillance des agences gouvernementales d'enquête ?
- ➔ But légitime : La loi établit-elle certains objectifs légaux de surveillance, tels que la prévention du terrorisme ou des crimes graves, avec une peine légale de 10 ans ou plus d'emprisonnement ?
- ➔ Motifs raisonnables : Les juges sont-ils habilités à déterminer s'il existe un niveau élevé de menace pour un objectif légitime et une forte probabilité que la surveillance génère des preuves qui éliminent la menace ?
- ➔ Légalité : La surveillance est-elle effectuée exclusivement dans les limites et par les agences spécifiées par la loi ? La loi rend-elle illégale toute autre surveillance et prévoit-elle des sanctions ?
- ➔ Nécessité : Les juges sont-ils autorisés à déterminer si une surveillance est nécessaire pour obtenir les preuves et qu'il n'existe pas de méthode moins intrusive pour atteindre le but légitime ?

399 Voir : https://unsdg.un.org/2030-agenda/universal-values/human-rights-based-approach

400 Roberts T., Mohamed A., Farahat, M., Oloyede R., Mutung'u G. (2021). Surveillance Law in Africa: a Review of Six Countries, Institute of Development Studies: Brighton, disponible sur : https://opendocs.ids.ac.uk/opendocs/bitstream/handle/20.500.12413/16893/Roberts_Surveillance_Law_in_Africa.pdf

401 Plus de 600 groupes, dont Privacy International, l'Open Rights Group, l'Electronic Frontier Foundation et l'Association for Progressive Communications, ont coordonné la rédaction des Principes internationaux, disponible sur : https://www.eff.org/files/necessaryandproportionatefinal.pdf

- ➔ Proportionnalité : Les juges sont-ils habilités à déterminer si la surveillance proposée a une portée limitée et si sa durée est proportionnelle aux éléments de preuve requis pour éliminer la menace ?
- ➔ Notification du sujet : La loi exige-t-elle que le sujet de la surveillance soit informé de la surveillance dès que possible, afin de lui donner la possibilité de faire appel et de bénéficier d'une procédure régulière ?
- ➔ Rapports de transparence : Les rapports annuels sur l'ouverture rendent-ils public le nombre de demandes de surveillance, de justifications et d'autorisations ?
- ➔ Surveillance indépendante : Les pratiques de surveillance disposent-elles de mécanismes de surveillance publics pour assurer leur responsabilité et leur transparence ?[402]



Source : Adapté de Roberts T., Mohamed A., Farahat, M., Oloyede R., Mutung'u G. (2021). Surveillance Law in Africa: a Review of Six Countries, Brighton: Institute of Development Studies, available at: DOI: 10.19088/IDS.2021.059

402 Roberts T., Mohamed A., Farahat, M., Oloyede R., Mutung'u G. (2021). Surveillance Law in Africa: a Review of Six Countries, Institute of Development Studies: Brighton, disponibles sur : https://opendocs.ids.ac.uk/opendocs/bitstream/handle/20.500.12413/16893/Roberts_Surveillance_Law_in_Africa.pdf

Comme l'IA s'intègre rapidement dans tous les secteurs, il est important que les opérateurs judiciaires prennent en compte les avantages et les risques uniques associés aux différents systèmes d'IA. Les assistants virtuels, les véhicules autonomes et les recommandations vidéo pour les enfants présentent tous différents niveaux d'avantages et de risques. Par conséquent, l'élaboration des politiques et la gouvernance doivent être abordées différemment pour chaque système d'IA, en fonction des risques encourus, de leur gravité et de leur impact sur les droits humains. Le tableau 7 ci-dessous donne un aperçu des principes directeurs régissant l'IA.

Tableau 7. Sélectionner les principes directeurs dans la gouvernance de l'IA

Principes	Questions clés dans la mise en œuvre des principes
Plus le risque pour les droits humains est élevé, plus les normes juridiques doivent être strictes pour l'utilisation de la technologie de l'IA.	Les secteurs où les enjeux d'empiètement sur les droits fondamentaux individuels sont élevés, tels que la sécurité nationale, la justice pénale, l'application de la loi, la santé et la protection sociale, doivent avoir la priorité. Une approche proportionnée aux risques de la réglementation de l'IA nécessite l'interdiction de technologies, d'applications et de cas d'utilisation spécifiques de l'IA qui produisent des impacts potentiels ou réels qui violent les droits humains internationaux, notamment ceux qui ne respectent pas les exigences de nécessité et de proportionnalité.[403]
Les applications d'IA discriminantes ne doivent pas être autorisées.	La notation sociale des individus par les gouvernements[404] ou l'utilisation de systèmes d'IA qui classent les individus en groupes en fonction de facteurs discriminatoires interdits doivent être proscrits. Les gouvernements devront contrôler l'utilisation et l'achat de technologies d'IA dont le déploiement dans le système judiciaire constitue un danger pour les droits humains. Lorsque des violations des droits humains sont susceptibles de se produire, l'exigence d'une surveillance humaine (incluant l'intervention humaine) doit être obligatoire. Les gouvernements doivent reporter le déploiement de technologies potentiellement à haut risque, telles que la reconnaissance faciale à distance en temps réel, jusqu'à ce qu'il soit garanti que leur mise en œuvre ne viole pas les droits humains.[405]
Si un système d'IA est utilisé pour dialoguer avec des humains dans le cadre de services publics, en particulier la justice, le bien-être et les soins de santé, l'utilisateur doit être informé et averti de la possibilité de consulter un professionnel, sur demande et sans délai.	Ceux pour qui une décision, uniquement ou substantiellement basée sur le résultat d'un système d'IA, a été prise par une autorité publique doivent être alertés et recevoir les informations susmentionnées dès que possible.[406] Cela peut prendre la forme soit d'une divulgation publique d'informations sur le système en question, ses processus, ses effets directs et indirects sur les droits humains et les mesures prises pour identifier et atténuer les conséquences négatives du système sur les droits humains, soit d'un audit impartial, approfondi et efficace. Dans tous les cas, les informations fournies doivent permettre une évaluation significative du système d'IA. Aucun système d'IA ne doit être si compliqué que l'évaluation et l'inspection humaines soient impossibles. Les systèmes ADM qui ne peuvent se soumettre à des normes de transparence et de responsabilité adéquates ne doivent pas être utilisés dans la prestation des services publics.[407]

403 Le projet de loi sur l'IA de l'Union européenne adopte une telle approche fondée sur les risques.

404 CAHAI (2020). L'impact de l'intelligence artificielle sur les droits de l'homme, la démocratie et l'État de droit, par. 75, disponible sur : https://rm.coe.int/cahai-2020-06-fin-c-muller-the-impact-of-ai-on-human-rights-democracy-/16809ed6da ; voir aussi : UNESCO (2021). Recommendation on the Ethics of Artificial Intelligence, disponible sur : https://unesdoc.unesco.org/ark:/48223/pf0000380455

405 Parlement européen (2019). A governance framework for algorithmic accountability and transparency, disponible sur : https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2019)624262 voir aussi : CAHAI (2020). The Impact of Artificial Intelligence on Human Rights, Democracy and the Rule of Law, disponible sur : https://rm.coe.int/cahai-2020-06-fin-c-muller-the-impact-of-ai-on-human-rights-democracy-/16809ed6da

406 CAHAI (2020). The Impact of Artificial Intelligence on Human Rights, Democracy and the Rule of Law, disponible sur : https://rm.coe.int/cahai-2020-06-fin-c-muller-the-impact-of-ai-on-human-rights-democracy-/16809ed6da

407 *Ibid.*

Le projet de loi de l'UE sur l'IA en tant qu'exemple de réglementation de l'IA fondée sur les risques

En 2019, à la suite de la publication des lignes directrices éthiques pour une IA digne de confiance[408], la Commission européenne a lancé une approche à plusieurs volets pour réglementer l'IA et traiter les risques associés. En plus du projet de loi sur l'IA, les règles de responsabilité civile nouvelles et modifiées[409] agissent conjointement avec d'autres politiques actuelles et prévues relatives aux données, telles que le RGPD[410], la loi sur les services numériques[411], la loi proposée sur les données[412] et la loi proposée sur la cyber-résilience[413].

Le projet de loi de l'UE sur l'IA établit des normes horizontales pour le développement, la commercialisation et l'utilisation de produits, de services et de systèmes alimentés par l'IA, au sein de l'UE. Il fournit des lignes directrices fondamentales basées sur les risques de l'IA applicables à tous les secteurs, et comprend un « cadre de sécurité des produits » avec quatre catégories de risques, spécifiant les règles d'entrée sur le marché et la certification des systèmes d'IA à haut risque par le biais d'un processus de marquage CE obligatoire. Ce régime de conformité couvre également les ensembles de données utilisés pour la formation, les tests et la validation de l'apprentissage automatique, afin de garantir des résultats équitables.

Le projet de loi de l'UE sur l'IA utilise une stratégie basée sur les risques, avec de multiples mécanismes d'application. Les applications d'IA à faible risque seraient soumises à un cadre réglementaire plus indulgent, tandis que celles présentant des risques inacceptables seraient interdites. À mesure que le risque augmente, des réglementations plus strictes s'appliquent. Celles-ci vont d'exigences de certification externes plus légères tout au long du cycle de vie de l'application, à des évaluations d'impact de la législation non contraignantes combinées à des codes de conduite.

Le cadre réglementaire définit quatre niveaux de risque, dans l'IA :

(i) Risque inacceptable. Les systèmes d'IA préjudiciables aux droits, à la sécurité et aux moyens de subsistance des personnes doivent être interdits, notamment les systèmes de notation sociale utilisés par les gouvernements et les jouets à commande vocale qui favorisent les comportements à risque.[414]

(ii) Risque élevé. La proposition initiale (2021) comprenait (i) les infrastructures critiques (par exemple, les transports), qui pourraient mettre en danger la vie et la santé des citoyens ; (ii) la formation scolaire ou professionnelle qui peut déterminer l'accès à l'éducation et au cursus professionnel de la vie de quelqu'un (par exemple, les notes aux examens) ; (iii) les composants de sécurité des produits (par exemple, les applications d'IA dans la chirurgie

408 Commission européenne (2019). Ethics guidelines for trustworthy AI, disponibles sur : https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
409 Commission européenne (2022). New liability rules on products and AI to protect consumers and foster innovation, disponible sur : https://ec.europa.eu/commission/presscorner/detail/en/ip_22_5807
410 Commission européenne (2021). Data Protection, disponible sur : https://commission.europa.eu/law/law-topic/data-protection_en
411 Commission européenne (2022). Digital Services Act, disponible sur : https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package
412 Commission européenne (2023). Data Act: Commission welcomes political agreement on rules for a fair and innovative data economy, disponible sur : https://ec.europa.eu/commission/presscorner/detail/en/ip_23_3491
413 Commission européenne (2022). Cyber Resilience Act , disponible sur : https://digital-strategy.ec.europa.eu/en/library/cyber-resilience-act
414 Voir : https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

assistée par robot) ; (iv) l'emploi, la gestion des travailleurs et l'accès au travail indépendant (par exemple, les services de tri de curriculum vitae à des fins de recrutement) ; (v) les services privés et publics essentiels (par exemple, la notation de crédit refusant aux citoyens la possibilité d'obtenir un prêt) ; (vi) les activités d'application de la loi interférant avec les droits humains (par exemple, l'évaluation de l'admissibilité des preuves) ; (vii) la gestion de la migration, de l'asile et du contrôle aux frontières (par exemple, la vérification de l'authenticité des documents de voyage) ; (viii) l'administration de la justice et des processus démocratiques (par exemple, l'application de la loi à un ensemble concret de faits).

La proposition de décembre 2022 a supprimé la détection de deepfakes par les forces de l'ordre, l'analyse de la criminalité et la vérification de l'authenticité des documents de voyage de la liste des systèmes d'IA à haut risque. Les dernières modifications précisent que le champ d'application du projet de loi n'englobe pas l'IA à des fins de sécurité nationale, de défense et militaires.

Toutes les technologies d'identification biométrique à distance sont soumises à des réglementations strictes et sont considérées comme à haut risque. En général, il est interdit d'utiliser l'identification biométrique à distance pour l'application de la loi, dans les zones ouvertes au public. Seules quelques situations peuvent être autorisées à titre exceptionnel, par exemple lorsqu'il est impératif de retrouver un enfant disparu, de mettre fin à une menace terroriste spécifique et imminente, ou de trouver, identifier ou poursuivre un auteur ou un suspect d'un crime grave. Une telle utilisation est soumise à des limites de temps, d'emplacement et de recherche dans la base de données, ainsi qu'à l'approbation d'un organe judiciaire ou d'un autre organe impartial.[415]

(iii) Risque limité. Les systèmes d'IA à risque limité doivent respecter des exigences de divulgation spécifiques. Les utilisateurs doivent être conscients qu'ils interagissent avec une machine lorsqu'ils utilisent des systèmes d'IA comme les chatbots, afin qu'ils puissent décider eux-mêmes s'ils doivent poursuivre ou rétrocéder.[416]

(iv) Risque minime ou absence de risque. Des applications telles que des filtres anti-spam ou des jeux vidéo avec IA sont concernées.

Les utilisateurs assurent le contrôle et la surveillance humaines, une fois qu'un système d'IA est mis sur le marché, tandis que les fournisseurs disposent d'une structure de surveillance post-commercialisation. Les autorités sont responsables de la surveillance du marché. Les événements graves et les dysfonctionnements seront signalés à la fois par les fournisseurs et les utilisateurs.[417]

415 *Ibid.*
416 *Ibid.*
417 *Ibid*

Approches de la gouvernance de l'IA fondées sur les droits humains

Une approche fondée sur les droits humains est essentielle pour construire des systèmes d'IA fiables dans la prestation des services publics. Pour garantir une approche fondée sur les droits dans les opérations du secteur public, les gouvernements des pays en développement doivent disposer d'un cadre analytique facilement accessible pour les aider à identifier à quel moment les composantes de l'IA pourraient avoir un impact sur les droits humains et la manière dont la responsabilité algorithmique pourrait atténuer ces risques. Lorsque les systèmes d'IA menacent les droits fondamentaux, les pays doivent protéger et promouvoir ces droits et veiller à ce que les acteurs du secteur privé procèdent à une diligence raisonnable et à des évaluations d'impact sur les droits humains (EIDH) conformément à leur responsabilité. Les résultats des évaluations des risques pour la santé doivent conduire à différentes garanties attribuées aux risques et impacts spécifiques établis dans le processus.[418]

Les gouvernements du monde entier, tels que celui des États-Unis (Blueprint for an AI Bill of Rights)[419], ont tenté de résoudre les problèmes de responsabilité et de transparence de l'IA sous l'aspect des droits humains. Un cadre précieux pour la réalisation d'évaluations d'impact algorithmiques basées sur l'approche des droits humains est fourni par l'outil d'évaluation d'impact des droits fondamentaux et des algorithmes (FRAIA) développé par le ministère néerlandais de l'Intérieur et des Relations du Royaume.[420]

418 Agence des droits fondamentaux de l'Union européenne (2022). Bias in Algorithms – Artificial Intelligence and Discrimination, disponible sur : https://fra.europa.eu/sites/default/files/fra_uploads/fra-2022-bias-in-algorithms_en.pdf

419 Maison-Blanche (2022). Blueprint for an AI Bill of Rights, disponible sur : https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf

420 Ministère néerlandais de l'Intérieur et des Relations du Royaume (2022). Impact Assessment Fundamental rights and algorithms, disponible sur https://www.government.nl/documents/reports/2022/03/31/impact-assessment-fundamental-rights-and-algorithms.

Évaluations d'impact sur les droits humain (EIDH)

Les EIDH peuvent aider à identifier les risques que les opérateurs judiciaires pourraient autrement ne pas prévoir dans le développement et le déploiement de l'IA. Pour ce faire, les EIDH accordent la priorité aux implications en matière de droits humains, plutôt qu'à l'optimisation de la technologie ou de ses résultats. Des évaluations des risques pour la santé ou des processus comparables pourraient assurer le respect des droits humains dès la conception, tout au long du cycle de vie de la technologie.

Les EIDH évaluent la technologie en fonction d'une grande variété d'impacts potentiels sur les droits humains. Lorsque l'ADM est utilisé dans les contextes judiciaires, les parties prenantes doivent mener des EIDH transparentes, impartiales et inclusives qui consistent en un examen des produits, des services et des systèmes d'intermédiaires entourant le développement et le déploiement de l'IA, et leurs effets sur les droits humains. Ces EIDH doivent intégrer les contributions des communautés touchées et des organisations de parties prenantes, notamment la société civile et les groupes marginalisés. Les résultats des EIDH doivent être rendus publics et librement accessibles et compréhensibles.[421]

Les EIDH pour l'IA doivent étudier le fonctionnement interne des algorithmes, c'est-à-dire qu'ils doivent analyser leurs composants techniques. Les EIDH pour les algorithmes doivent également être entreprises tout au long du cycle de vie d'un système d'IA, dès les premières étapes de sa conception et lors des phases importantes de son développement et de son déploiement. Elles ne doivent pas intervenir ex ante ou ex post uniquement. L'évaluation de l'impact sur les droits fondamentaux et les algorithmes (FRAIA), développée par le gouvernement néerlandais, et l'évaluation de l'impact sur les droits humains, l'éthique et la société en IA, développée par Alessandro Manterelo à l'Université de Turin, sont des évaluations récentes des droits humains qui répondent à ces exigences. Ces deux EIDH donnent des recommandations pour aider les développeurs et les déployeurs d'IA à identifier l'impact des systèmes d'IA sur un large éventail de droits fondamentaux. En outre, elles fournissent divers exemples de stratégies d'atténuation potentielles, pour prévenir les effets indésirables. Tout cela minimise la probabilité de violations injustifiables des droits humains. La FRAIA examine l'impact des systèmes d'IA sur plus d'une centaine de droits et sous-droits fondamentaux — par exemple, la liberté d'expression est subdivisée en de nombreux sous-droits, tels que la liberté de la presse, la liberté académique et la dénonciation — et propose une liste complète de mesures préventives et d'atténuation pour limiter les violations de ces droits.

Voici un aperçu du processus FRAIA :

Cette évaluation de l'impact des droits fondamentaux et des algorithmes (‹FRAIA›) est un outil de discussion et de prise de décision pour les organisations gouvernementales. L'outil facilite un dialogue interdisciplinaire en étant responsable du développement et/ou de l'utilisation du système algorithmique Tan. Le client de mise en service est principalement responsable de la mise en œuvre (déléguée) de la FRAIA.

La FRAIA comprend un grand nombre de questions sur les sujets qui doivent être discutés et auxquelles une réponse doit être formulée dans tous les cas où une organisation gouvernementale envisage de déléguer le développement, l'achat, l'ajustement et/ou l'utilisation d'un algorithme (ci-après par souci de brièveté, l'utilisation usi anr). Lorsqu'un ritm ss ald est utilisé, la FRAIA peut servir d'outil de réflexion. La discussion sur les différentes questions devrait avoir lieu dans une équipe pluridisciplinaire composée de personnes ayant un large éventail de spécialisations et d'antécédents. Par question, la FRAIA indique qui doit être impliqué dans la discussion. Cet outil porte une attention particulière à tous les rôles au sein d'une équipe pluridisciplinaire, qui sont inclus dans le diagramme ci-dessous. La liste n'en est cependant pas exhaustive. De même, les noms des rôles ou des fonctions peuvent différer d'une organisation à l'autre.

Rôle	FRAIA Partie 1	FRAIA Partie 2	FRAIA Partie 3	FRAIA Partie 7
Groupe interset	•			
Gestion	•			
Panel citoyen	•			
CISO ou CIO	•	•		
Spécialiste en communication		•	•	
Expert en mégadonnées		•	•	
Responsable du traitement ou propriétaire de la source de données		•		
Expert de domaine (Employé qui a des connaissances domani redg le champ d'application de l'algorithme)	•	•	•	•
Délégué à la protection des données		•		
Membre du personnel RH			•	
Conseiller juridique	•	•	•	•
Développeur d'algorithmes		•		
Client de mise en service	•	•	•	
Autres membres de l'équipe de projet	•			
Chef du projet	•	•	•	•
Consultant en éthique stratégique		•	•	

Source : OCDE, AI in Society, disponible sur : https://www.oecd-ilibrary.org/sites/969ff07f-en/index.html?itemId=/content/component/969ff07f-en; Gaumond E., Régis C. (2023). Assessing Impacts of AI on Human Rights: It' s Not Only About Privacy and Nondiscrimination, disponible sur : https://www.lawfareblog.com/assessing-impacts-ai-human-rights-its-not-solely-about-privacy-and-nondiscrimination.

421 OSCE (2022). Spotlight on Artificial Intelligence and Freedom of Expression: A Policy Manual, disponible sur : https://www.osce.org/representative-on-freedom-of-media/510332

Le cadre d'assurance des droits de l'homme, de la démocratie et de l'état de droit (HUDERAF) pour les systèmes d'IA

L'HUDERAF, proposé par l'institut Alan Turing (qui a conseillé le CAHAI – le Comité ad hoc du Conseil de l'Europe sur l'intelligence artificielle) vise à présenter une méthode cohérente et intégrée pour évaluer les effets négatifs potentiels sur les droits humains, la démocratie et l'état de droit causés par les systèmes d'IA, ainsi que pour s'assurer que les risques identifiés posés par l'IA aux opérateurs judiciaires sont atténués et gérés de manière adéquate. Le cadre est spécifiquement composé de plusieurs procédures et outils bien articulés, mais logiquement connectés. Il combine des approches transparentes de gestion des risques, d'atténuation des impacts et d'assurance de l'innovation, avec des évaluations des risques basées sur le contexte et une implication appropriée des parties prenantes. Les opérateurs judiciaires pourraient utiliser le cadre HUDERAF pour évaluer les impacts négatifs potentiels de l'IA sur les droits humains.

L'HUDERAF comprend quatre composantes :

(1) L'analyse préliminaire des risques basée sur le contexte (PCRA) donne une première indication des risques contextuels qu'un système d'IA peut poser pour les droits humains, la démocratie et l'état de droit. L'objectif principal de la PCRA est d'aider les équipes de projet d'IA à élaborer une stratégie raisonnable pour les procédures de gestion et d'assurance des risques ainsi que le degré d'implication des parties prenantes requis tout au long du cycle de vie du projet.

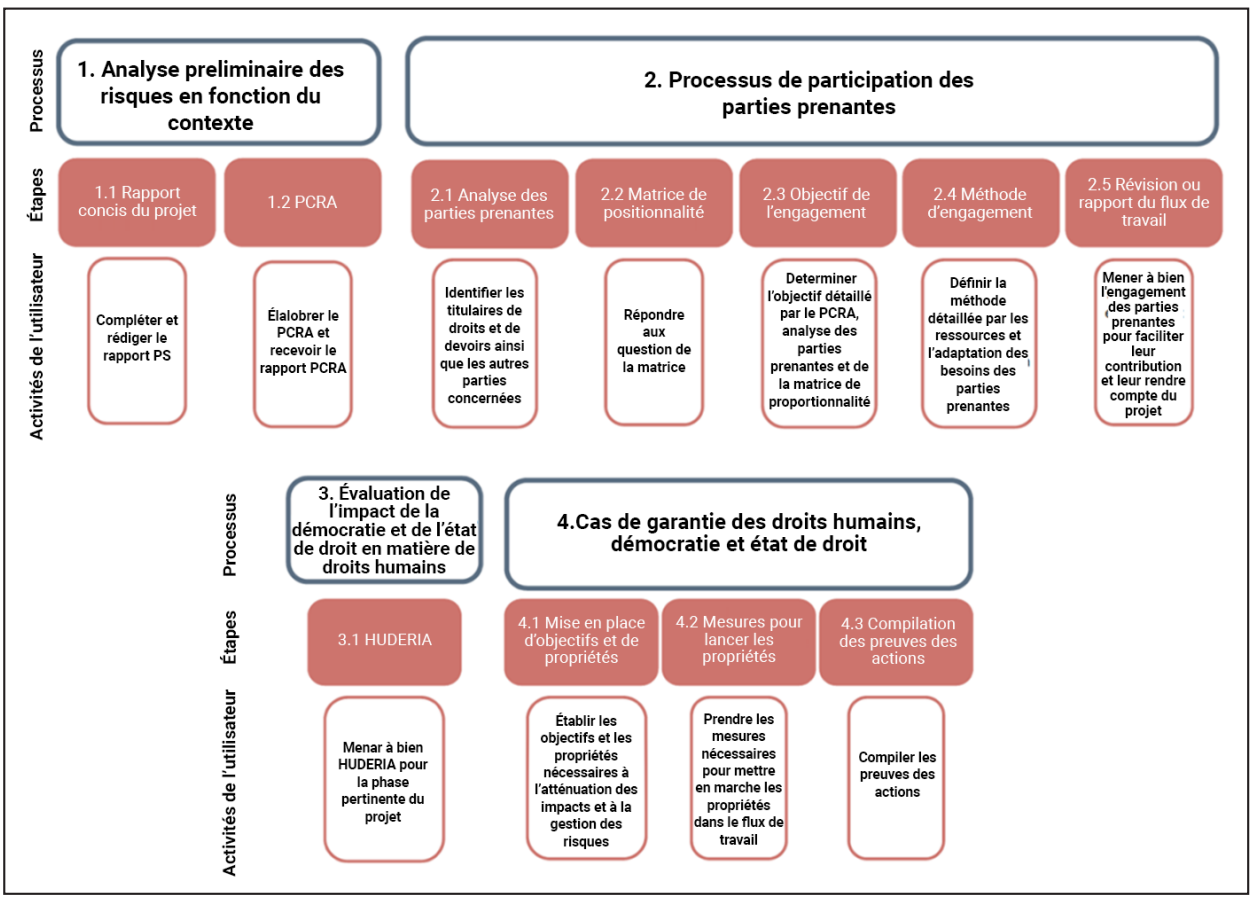
(2) Le processus d'engagement des parties prenantes (SEP) soutient la participation et la contribution appropriées des parties prenantes tout au long du processus de projet, en aidant les équipes de projet à identifier la prépondérance des parties prenantes. Grâce à la participation, à la révision et à l'examen des parties prenantes, cette méthode protège l'égalité et la précision contextuelle des processus de gouvernance HUDERAF.

(3) L'évaluation d'impact sur les droits humains, la démocratie et l'état de droit (HUDERIA) donne aux équipes de projet et aux parties prenantes impliquées la possibilité de travailler ensemble pour créer des évaluations approfondies des effets possibles et réels que la conception, le développement et l'utilisation d'un système d'IA pourraient avoir sur les droits humains, la démocratie et l'état de droit. Grâce à l'intégration des points de vue des parties prenantes, ce processus contextualise et valide les préjudices potentiels précédemment identifiés, permet la découverte de préjudices supplémentaires, l'évaluation collaborative de la gravité des impacts négatifs potentiels identifiés, facilite la co-conception d'un plan d'atténuation des impacts, établit l'accès aux recours et définit des protocoles pour le suivi et la réévaluation des impacts.

(4) L'affaire de l'assurance des droits humains, de la démocratie et de l'état de droit (HUDERAC) permet aux équipes de projet d'IA de construire une justification structurée qui donne aux parties prenantes une assurance démontrable que les allégations concernant la réalisation des objectifs énoncés dans l'HUDERIA et

d'autres processus de gouvernance de l'HUDERAF sont justifiées à la lumière des preuves disponibles. La création d'un cas d'assurance facilite la réflexion et la discussion internes, en encourageant l'adoption des meilleures pratiques et en les intégrant dans les cycles de vie de la conception, du développement et du déploiement. En outre, cela offre un moyen clair d'informer les parties prenantes concernées des mesures prises tout au long du flux de travail du projet, afin de réduire les risques et de garantir l'identification d'objectifs normatifs pertinents. Un dossier d'assurance soigneusement élaboré fournit un cadre transparent et facile à comprendre pour la gestion des risques et l'atténuation de leurs effets, en soutenant les bons niveaux d'acceptation sociale, de responsabilité et d'ouverture.[422]

Figure 14 : Démocratie des droits humains et cadre d'assurance de l'état de droit (HUDERAF)



Source : Leslie D., Burr C., Aitken M., Cowls J., Katell M., Briggs M. (2021). Artificial intelligence, human rights, democracy, and the rule of law: a primer. Le Conseil de l'Europe, disponible sur : https://www.turing.ac.uk/sites/default/files/2021-03/cahai_feasibility_study_primer_final.pdf

422 Voir : https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592

Ces activités de groupe visent à encourager les participants à la formation à discuter et à débattre des cas d'éventuelles atteintes aux droits humains, en utilisant l'ADM et l'IA dans les opérations judiciaires, et des cas de délibération judiciaire sur les droits humains violés lors de l'utilisation de l'IA par des tiers.

Activité 1

Veuillez consulter l'annexe B de la Directive canadienne sur la prise de décision automatisée et examiner les quatre niveaux d'impact qu'une décision assistée par l'IA peut avoir sur les droits fondamentaux.[423]

Envisagez le scénario suivant : L'agence pour l'emploi du pays X a l'intention de calculer la probabilité que les demandeurs d'emploi enregistrés trouvent un emploi dans un certain délai à l'avenir, en tenant compte de plusieurs facteurs : le groupe d'âge des demandeurs d'emploi, le genre, l'éducation, les conditions de santé, les tâches de soins, la performance de leur marché du travail régional et la durée de leur enregistrement auprès de l'agence pour l'emploi. Sur la base de la probabilité calculée, les demandeurs d'emploi seront répartis en différents groupes : le premier groupe couvre les demandeurs d'emploi ayant des opportunités de marché élevées, le deuxième bénéficie d'opportunités moyennes et le dernier d'opportunités faibles. Le système d'IA aidera les conseillers des agences pour l'emploi à évaluer les opportunités des demandeurs d'emploi et permettra une utilisation plus efficace des ressources. Sur la base de ce scénario, les participants à la formation examinent les quatre niveaux d'impact qu'une décision prise par le système d'IA peut avoir sur les droits des demandeurs d'emploi.[424]

Annexe B : Niveaux d'évaluation d'impact	
Niveau	Description
I	La décision aura probablement peu ou pas d'impact sur : • les droits des individus ou des communautés ; • la santé ou le bien-être des individus ou des communautés ; • les intérêts économiques des individus, des entités ou des communautés ; • la durabilité continue d'un écosystème. Les décisions de niveau I conduiront souvent à des impacts réversibles et brefs.
II	La décision aura probablement des impacts modérés sur : • les droits des individus ou des communautés ; • la santé ou le bien-être des individus ou des communautés ; • les intérêts économiques des individus, des entités ou des communautés ; • la durabilité continue d'un écosystème. Les décisions de niveau II entraîneront souvent des impacts susceptibles d'être réversibles et à court terme.
III	La décision aura probablement des impacts élevés sur : • les droits des individus ou des communautés ; • la santé ou le bien-être des individus ou des communautés ; • les intérêts économiques des individus, des entités ou des communautés ; • la durabilité continue d'un écosystème. Les décisions de niveau III conduiront souvent à des impacts qui peuvent être continus et difficiles à inverser. Compte tenu de l'impact élevé sur les libertés et les droits des individus et des communautés précédemment mis en évidence, le niveau III au minimum serait probablement atteint pour les activités de police prédictive.
IV	La décision aura probablement des impacts très élevés sur : • les droits des individus ou des communautés ; • la santé ou le bien-être des individus ou des communautés ; • les intérêts économiques des individus, des entités ou des communautés ; • la durabilité continue d'un écosystème. Les décisions de niveau IV entraîneront souvent des impacts irréversibles et perpétuels.

423 Voir : https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592

424 Barros Vale S., Zanfir-Fortuna G. (2022). Automated Decision-Making Under the GDPR: Practical Cases from Courts and Data Protection Authorities, disponible sur: https://fpf.org/blog/fpf-report-automated-decision-making-under-the-gdpr-a-comprehensive-case-law-analysis/

Activité 2

Modèle de fiche d'information sur l'évaluation des risques – examinez le modèle d'évaluation des risques suivant et voyez si vous imaginez d'autres questions que vous pourriez poser pour évaluer l'outil d'évaluation des risques.

- Qui a créé l'évaluation des risques ? S'agit-il d'une organisation publique ou privée ?
- Quelle était la taille de l'ensemble de données de formation ?
- Comment l'ensemble de données de formation a-t-il été recueilli et assemblé (c.-à-d. de quelle(s) juridiction(s) provient-il) ?
- Sur quelle période les données ont-elles été collectées ?
- Quels facteurs (c.-à-d. les caractéristiques du défendeur) ont été inclus dans l'ensemble de données ? Cette question porte sur tous les facteurs disponibles au sujet des défendeurs, pas nécessairement tous les facteurs qui ont été utilisés pour former ou développer le modèle.
- L'ensemble de données comprend-il des cas de défendeurs qui ont été détenus ? Si oui, les données incluent-elles les résultats pour ces personnes (c.-à-d., les données ont-elles pris en compte l'estimation erronée ; si oui, comment) ?
- Y a-t-il des problèmes ou des erreurs connus avec les données ?
- En quelle année l'évaluation des risques a-t-elle été créée ?
- Quels facteurs, parmi tous les facteurs des données de formation, ont été pris en compte dans l'élaboration de l'évaluation des risques ? Si tous les facteurs n'ont pas été pris en compte, comment ceux qui ont été envisagés ont-ils été sélectionnés ?
- Comment les facteurs ont été finalement choisis pour l'exclusion ou l'inclusion dans le modèle final (l'évaluation des risques elle-même) ?
- Le modèle final inclut-il comme facteur les arrestations qui n'ont pas conduit à des condamnations ? Le modèle final inclut-il des facteurs socio-économiques tels que le logement et la situation professionnelle ? Le modèle final inclut-il des facteurs de santé personnelle tels que la santé mentale ou la toxicomanie ? [diviser en plusieurs questions, si des informations pertinentes sont disponibles]
- Comment les pondérations ont-elles été attribuées à chaque facteur inclus dans le modèle final (coefficients de corrélation arrondis, méthode de Burgess, etc.) ?
- Comment le modèle final définit-il les résultats (c.-à-d., au cours du processus d'élaboration du modèle, y a-t-il eu un résultat distinct défini pour chaque type d'échec - défaut de comparution, nouveau crime, nouveau crime violent, etc. - ou les résultats ont-ils été aggravés) ?
- À quoi ressemble le résultat du modèle (c.-à-d., une note sur une échelle de 1 à 10, etc.) ?
- Le modèle produit-il des désignations de niveau de risque ou convertit-il les notes brutes en désignations de niveau de risque telles que « risque faible », « risque modéré » et « risque élevé » ?

- Quelle proportion d'échantillons dans l'ensemble de données de formation a échoué à chaque note et/ou niveau de risque (par exemple, quel pourcentage de personnes ayant une note de 5 ou une étiquette de « risque modéré » n'est pas apparu) ?
- Les développeurs du modèle ont-ils évalué la validité prédictive du modèle ? Si oui, comment ?
- Où l'évaluation des risques est-elle utilisée ?
- Les facteurs et les pondérations de l'évaluation des risques sont-ils accessibles au public ?
- L'adoption d'une évaluation des risques coûte-t-elle de l'argent à une juridiction ?
- L'adoption de l'évaluation des risques nécessite-t-elle une formation ? Si oui, laquelle ?
- L'évaluation des risques s'accompagne-t-elle d'un logiciel ou d'un progiciel ?
- L'évaluation des risques implique-t-elle ou nécessite-t-elle un entretien en personne ?
- Comment l'évaluation des risques tient-elle compte des informations manquantes ?
- L'évaluation des risques a-t-elle été analysée sur des données non liées à la formation pour une validité prédictive ? L'évaluation des risques a-t-elle été analysée avec des données d'entraînement ou des données non liées à l'entraînement sur les performances pour différents groupes ethniques ? L'évaluation des risques a-t-elle été analysée avec des données de formation ou des données non liées à la formation sur les performances pour différents genres ? Si oui, par qui, quand et à partir de quelles données ?[425]

Activité 3

Veuillez discuter des questions suivantes avec les autres participants à la formation :

- Qu'implique la vie privée, à une époque où la collecte de données en temps réel est monnaie courante et où il existe un risque de violation de données, de vol d'identité ou de fraude en ligne ?
- Pouvons-nous nous exprimer librement sur tous les outils et plateformes numériques sans nous soucier de la censure de l'IA ou du profilage ?
- Tout le monde peut-il avoir un accès égal à des informations fiables, compte tenu de la diffusion généralisée de matériel préjudiciable et de mensonges en ligne ?
- Comment pouvons-nous nous assurer que les technologies de l'IA aident à combler la fracture numérique plutôt que d'élargir les disparités déjà existantes ?

426 Voir : https://law.stanford.edu/pretrial-risk-assessment-tools-factsheet-project/

5. Ressources

1. Access Now (2018). Mapping artificial intelligence strategies in Europe: a new report by Access Now, disponible sur : https://www.accessnow.org/mapping-artificial-intelligence-strategies-in-europe/
2. Article 19, The Danish Institute for Human Rights (2017). Sample ccTLD Human Rights Impact Assessment Tool, disponible sur : https://www.article19.org/wp-content/uploads/2017/12/Sample-ccTLD-HRIA-Dec-2017.pdf
3. Auditing machine learning algorithms. A white paper for public auditors. by the Supreme Audit Institutions of Finland, Germany, the Netherlands, Norway and the UK (2023), https://www.auditingalgorithms.net/
4. Australian Human Rights Commission (2018). Final Report: Human Rights and Technology, disponible sur : https://tech.humanrights.gov.au/sites/default/files/2018-07/Human%20Rights%20and%20Technology%20Issues%20Paper%20FINAL.pdf
5. CAHAI (2020). Legal Framework for AI Systems. Feasibility study of a legal framework for the development, design and application of artificial intelligence, based on Council of Europe's standards on human rights, democracy and the rule of law, Council of Europe Study, DGI/2021/04, disponible sur : https://edoc.coe.int/en/artificial-intelligence/9648-a-legal-framework-for-ai-systems.html
6. Council of Europe (2020). Recommendation CM/Rec (2020) of the Committee of Ministers to member States on the human rights impacts of algorithmic systems, disponible sur : https://rm.coe.int/09000016809e1154
7. Dearden L. (2018). New Technology Can Detect ISIS Videos before They Are Uploaded, disponible sur : http://www.independent.co.uk/news/uk/home-news/isis-videos-artificial-intelligence-propaganda-ai-home-office-islamic-state-radicalisation-asi-data-a8207246.html
8. Duarte N., Llanso E., Loup A. (2017), Mixed Messages? The Limits of Automated Social Media Content Analysis, disponible sur : https://cdt.org/insight/mixed-messages-the-limits-of-automated-social-media-content-analysis/
9. Elsayed-Ali S. (2017). Artificial Intelligence and the Future of Human Rights, disponible sur : https://medium.com/amnesty-insights/artificial-intelligence-and-the-future-of-human-rights-b58996964df5
10. Edwards L. (2022). Expert opinion: Regulating AI in Europe. Four problems and four solutions, disponible sur : https://www.adalovelaceinstitute.org/report/regulating-ai-in-europe/
11. EUR-Lex (2021). Draft EU AI Regulation, disponible sur : https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206
12. European Commission (2021). Study to Support an Impact Assessment of Regulatory Requirements for Artificial Intelligence in Europe, disponible sur : https://artificialintelligenceact.eu/wp-content/uploads/2022/06/AIA-COM-Impact-Assessment-3-21-April.pdf
13. European Data Protection Supervisor. Necessity & Proportionality, disponible sur : https://edps.europa.eu/data-protection/our-work/subjects/necessity-proportionality_en
14. ICO, AI and data protection risk toolkit, disponible sur : https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/ai-and-data-protection-risk-toolkit/
15. Jones K. (2023), disponible sur : https://www.chathamhouse.org/2023/01/ai-governance-and-human-rights

16. Latonero M (2018) Artificial Intelligence & Human Rights: A Workshop at Data & Society, A Workshop at Data & Society, disponible sur : https://medium.com/datasociety-points/artificial-intelligence-human-rights-a-workshop-at-data-society-fd6358d72149

17. Latonero M. (2019). Governing Artificial Intelligence: Upholding Human Rights and Human Dignity, disponible sur : https://datasociety.net/wpcontent/uploads/2018/10/DataSociety_Governing_Artificial_Intelligence_Upholding_Human_Rights.pdf

18. L.iberty (2020). Liberty wins ground-breaking victory against facial recognition tech, disponible sur :https://www.libertyhumanrights.org.uk/issue/liberty-wins-ground-breaking-victory-against-facial-recognition-tech/

19. Mounk Y. (2018). Verboten. Germany's risky law for stopping hate speech on Facebook and Twitter, disponible sur : https://newrepublic.com/article/147364/verboten-germany-law-stopping-hate-speech-facebook-twitter

20. OECD AI Policy Observatory. Live data, disponible sur : https://oecd.ai/en/data?selectedArea=ai-research&selectedVisualization=top-countries-in-ai-scientific-publications-in-time-from-scopus

21. OECD AI Policy Observatory. Principles Overview, disponible sur : https://oecd.ai/en/ai-principles

22. Ortiz Freuler J., Iglesias C. (2018). Algorithms and Artificial Intelligence in Latin America: A Study of Implementation by Governments in Argentina and Uruguay, World Wide Web Foundation, disponible sur : http://webfoundation.org/docs/2018/09/WF_AI-in-LA_Report_Screen_AW.pdf

23. Peralta Gutiérrez (2022). Marco normativo de la Inteligencia Artificial en el ámbito comparado. In: Herrera Triguero F., Peralta Gutiérrez A., Torres López L.S., El derecho y la inteligencia artificial, EUG: Granada 189–222.

24. Pielemeier J. (2018). The Advantages and Limitations of Applying the International Human Rights Framework to Artificial Intelligence, disponible sur : https://points.datasociety.net/the-advantages-and-limitations-of-applying-the-international-human-rights-framework-to-artificial-291a2dfe1d8a

25. Reiling D., Contini F. (2022). E-Justice Platforms: Challenges for Judicial Governance, International Journal for Court Administration, 13 (1), disponible sur : https://iacajournal.org/articles/10.36745/ijca.445

26. Reisman D., Schultz J., Crawford K., Whittaker M. (2018). Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability, disponible sur : https://ainowinstitute.org/publication/algorithmic-impact-assessments-report-2

27. Reitman R. (2022). Podcast Episode: Algorithms for a Just Future, disponible sur : https://www.eff.org/deeplinks/2022/01/podcast-episode-algorithms-just-future

28. Stankovich M. (2021). Regulating AI and Big Data Deployment in Healthcare: Proposing Robust and Sustainable Solutions for Developing Countries' Governments, disponible sur : https://www.dai.com/uploads/regulating-ai-cda.pdf

29. Vincent J. (2019). AI won't relieve the misery of Facebook's human moderators, disponible sur : https://www.theverge.com/2019/2/27/18242724/facebook-moderation-ai-artificial-intelligence-platforms

30. YouTubeHelp. How Content ID Works, disponible sur : https://support.google.com/youtube/answer/2797370?hl=en

Ressources suggérées par l'UNESCO

Documents de l'UNESCO

Publications

Manuel de formation mondial pour les acteurs du judiciaire : normes juridiques internationales relatives à la liberté d'expression, l'accès à l'information et la sécurité des journalistes

- ☐ Disponible en : anglais, arabe, chinois, espagnol, français, portugais et russe

Normes juridiques sur la liberté d'expression : manuel de formation pour les acteurs du judiciaire en Afrique

- ☐ Disponible en : anglais, français et portugais

Lignes directrices destinées aux procureurs relatives aux crimes commis contre les journalistes

- ☐ Disponible en : anglais, amharique, arabe, chinois, dari, français, indonésien, italien, khmer, portugais, pachtoune, russe, somalien, espagnol, swahili, thaï, ukrainien, ouzbek

Lignes directrices destinées aux acteurs judiciaires relatives au respect de la vie privée et à la protection des données

- ☐ Disponible en : anglais, arabe, chinois, espagnol, français, portugais et russe

COVID-19 : le rôle des acteurs du judiciaire pour la protection et la promotion du droit à la liberté d'expression : lignes directrices

- ☐ Disponible en : anglais, arabe, birman, chinois, espagnol, français, khmer, portugais et russe

Sécurité des journalistes couvrant les manifestations : préserver la liberté de la presse en période de troubles

- ☐ Disponible en : anglais, arabe, birman, chinois, espagnol, français, portugais et russe

Manuel de formation mondial pour les agents des forces de l'ordre : Liberté d'expression, accès à l'information et sécurité des journalistes

- ☐ Disponible en : anglais, arabe, chinois, espagnol, français, portugais et russe

Freedom of expression and public order: fostering the relationship between security forces and journalists

- ☐ Disponible en : anglais, portugais, russe, somalien et ukrainien

L'« utilisation abusive » du système judiciaire pour attaquer la liberté d'expression : tendances, défis et réponses

- ☐ Disponible en : anglais, arabe, chinois, espagnol, français, italien, portugais et russe

Guide de l'UNESCO pour les interventions d'amicus curiae dans les affaires de liberté d'expression

- ☐ Disponible en : anglais, arabe, chinois, espagnol, français et russe

Série de vidéos et de webinaires

The Next Frontier: Intellectual Property in the Era of Generative Artificial Intelligence

- ☐ Disponible en : anglais et espagnol

The Admissibility Challenge: AI-Generated Evidence in the Courtroom

- ☐ Disponible en : anglais

Internet Governance Forum 2021 – Artificial Intelligence and the Rule of Law in the Digital Ecosystem

- ☐ Disponible en : anglais

Internet Governance Forum 2022 – Why Digital Transformation and Artificial Intelligence Matter for Justice

- ☐ Disponible en : anglais

UNESCO Video Explainers - Comment mettre fin à l'impunité pour les crimes contre les journalistes ?

- ☐ Disponible en : anglais, arable, chinois, espagnol, français et russe

Les limites légitimes à la liberté d'expression : le test en trois parties

☐ Disponible en : anglais, arabe, chinois, espagnol, français, portugais et russe

Les limites légitimes de la liberté d'expression : le Plan d'action de Rabat

☐ Disponible en : anglais, arabe, chinois, espagnol, français, portugais et russe

UNESCO Video Explainers - À quoi ressemblerait un monde sans médias indépendants ?

☐ Disponible en : anglais, arabe, chinois, espagnol, français et russe

UNESCO Video Explainers - Pourquoi la liberté d'expression et l'accès à l'info sont-ils essentiels pour des élections libres ?

☐ Disponible en : anglais, arabe, chinois, espagnol, français et russe

UNESCO Video Explainers - Les cours régionales en Afrique et la jurisprudence de référence en matière de liberté d'expression

☐ Disponible en : anglais, français et portugais

COVID-19 et liberté d'expression

☐ Disponible en : anglais, français et espagnol

Cours

MOOC - L'intelligence artificielle et l'état de droit

☐ Disponible en : anglais, arabe, chinois, espagnol, français, portugais et russe

MOOC - Les normes internationales en matière de liberté d'expression et de sécurité des journalistes

☐ Disponible en : anglais, arabe, chinois, espagnol, français, portugais et russe

COMMENT UTILISER CE MANUEL DE FORMATION ?

Ce manuel de formation encourage un modèle pédagogique expérientiel : il n'est pas destiné à être prescriptif, et les utilisateurs sont encouragés à s'inspirer de leurs propres expériences en tenant compte des contextes pertinents dans lesquels le manuel de formation est utilisé. Bien qu'il s'adresse principalement aux opérateurs judiciaires, il peut également être utile à d'autres, notamment les organisations de la société civile. Il existe plusieurs façons d'utiliser le manuel de formation en tant que ressource :

- Atelier complet en présentiel : Nous conseillons qu'un atelier complet en présentiel couvrant les quatre modules dure au moins trois jours. Dans des circonstances où les participants ne sont pas familiers des principes fondamentaux de la législation internationale des droits humains, nous conseillons que l'atelier se déroule sur au moins quatre jours.
- Atelier ciblé : Des ateliers pourraient également être organisés sur des modules sélectionnés dans le manuel de formation. Dans de telles circonstances, les formateurs doivent toujours s'assurer que les bases sont posées à partir des autres modules qui peuvent être nécessaires aux participants pour comprendre pleinement les concepts et accomplir les tâches.
- Cours en ligne (tel qu'un cours en ligne ouvert à tous) et atelier en présentiel combinés : Ce format donnerait plus de temps aux participants pour s'approprier le matériel et les exercices d'auto-évaluation, avant d'être réunis en présentiel. Idéalement, la composante en ligne doit être soutenue par des forums de discussion en ligne, entre autres méthodes.
- Auto-apprentissage : Le manuel de formation est de nature explicite et peut servir de ressource d'auto-apprentissage utile pour s'engager individuellement ou parmi un groupe de personnes travaillant dans une organisation particulière. Bien qu'il y ait souvent des avantages à avoir des discussions collaboratives et à partager des expériences, cela peut également être un point de départ et une référence utile pour qui cherche à améliorer sa compréhension des problèmes émergents en matière d'IA et de droits humains.

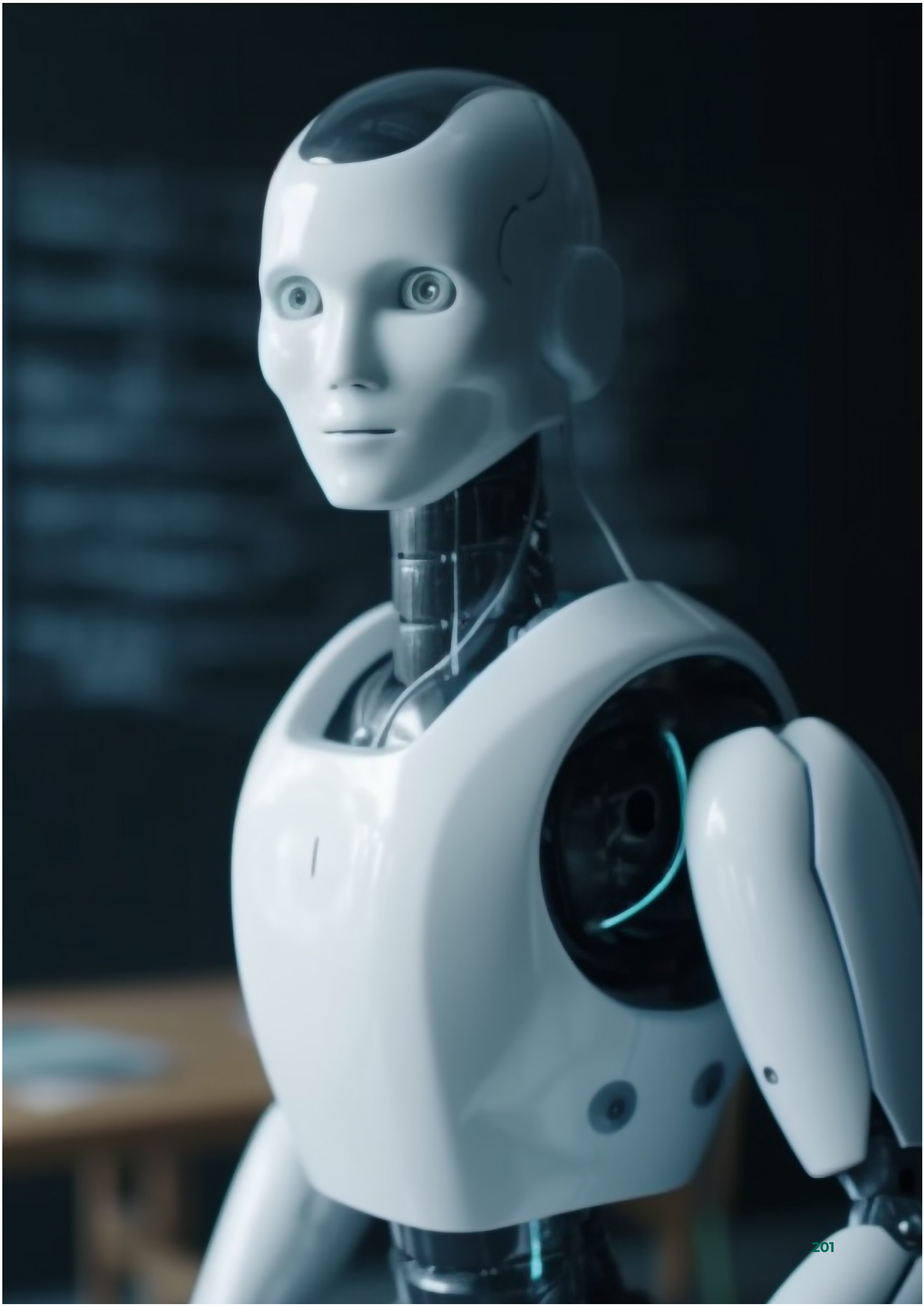
Bien que des ateliers plus longs permettraient de participer à davantage d'activités, il est peu probable qu'il y ait le temps de mener toutes les activités suggérées. Ceci se fait à la discrétion des formateurs. Les formateurs doivent chercher à évaluer, à partir des groupes, les aspects les plus pertinents pour les participants et qui peuvent être les mieux intégrés dans leur travail et leurs scénarios nationaux.

Il est conseillé aux formateurs de distribuer au préalable un questionnaire aux participants, pour s'assurer de leur expérience dans ce domaine du droit. Le modèle suivant pourrait être adapté en fonction des participants attendus à la formation :

Manuel de formation mondial : l'IA et l'état de droit pour le pouvoir judiciaire

Coordonnées des participants Nom : Organisation : Titre : Pays	Expérience des participants Avez-vous de l'expérience en droit ? Avez-vous de l'expérience dans la prise de décision automatisée, l'IA et les droits humains ? Veuillez détailler.
Quels modules seraient les plus utiles pour vous et votre travail ? Veuillez sélectionner. ☐ Module 1. Introduction à l'IA et à l'état de droit ☐ Module 2. Adoption de l'IA par le pouvoir judiciaire ☐ Module 3. Défis juridiques et éthiques du déploiement de l'IA dans le système judiciaire ☐ Module 4 Droits humains et IA : gouvernance, réglementation et politique	
Veuillez détailler.	Quels sont vos objectifs pour cette formation ?

« L'IA et les droits humains » est un domaine dynamique et évolutif du droit. En tant que tel, il est probable qu'il y ait de nouveaux développements fréquents. Les instructeurs doivent s'assurer de rester au courant de ces développements et de mettre à jour le matériel de formation en conséquence.



ANNEXE I
ÉVALUATION DE L'IMPACT ÉTHIQUE DE L'UNESCO POUR LES SYSTÈMES D'IA

Cet instrument a deux objectifs. Premièrement, évaluer si des algorithmes spécifiques sont alignés sur les valeurs, les principes et les orientations établis par la recommandation. Et deuxièmement, assurer la transparence en demandant que les informations sur les systèmes d'IA et la façon dont ils ont été développés soient accessibles au public. Ce n'est pas ainsi que cela fonctionne aujourd'hui, même pour les informations de base sur la sécurité et la fiabilité de l'IA.

Les outils d'évaluation d'impact gagnent du terrain, pour évaluer l'impact réel des systèmes d'IA. En fait, les évaluations d'impact sont mandatées par le projet de loi de l'UE sur l'IA pour les systèmes à haut risque, et elles sont proposées dans le cadre de la discussion du Conseil de l'Europe sur une convention pour l'IA.

La recommandation de l'UNESCO est unique, en ce sens qu'elle prend en compte l'ensemble du cycle de vie de l'IA. L'évaluation d'impact éthique comprend donc des exigences ex ante et ex post. À un stade précoce, elle établit l'importance d'assurer la qualité et la représentativité des données, la diversité des équipes développant les produits, la robustesse et la transparence des algorithmes, leur auditabilité, et la possibilité d'insérer des points de contrôle à différents moments du processus de développement.

L'EIE est proposée aux acheteurs de systèmes d'IA, car c'est l'un des principaux canaux utilisés par les algorithmes des domaines publics très sensibles. Mais les questions et la structure du document sont conçues pour que les outils puissent également être utilisés plus généralement par les développeurs de systèmes d'IA, dans les secteurs public ou privé, qui souhaitent développer l'IA de manière éthique et pleinement conforme aux normes internationales telles que la Recommandation.

Le document comprend deux parties principales qui, ensemble, établissent un équilibre entre la procédure et le fond. Dans la première partie, liée à la portée, l'objectif est de comprendre les bases du système, ainsi que de poser quelques questions préliminaires, telles que celle de savoir si l'automatisation est la meilleure solution pour le cas d'espèce. Cela soulève également des questions sur l'équipe du projet et le fait de savoir si des plans sont en place pour impliquer

les différentes parties prenantes. La deuxième partie est consacrée à la mise en œuvre des principes de la Recommandation de l'UNESCO.

Pour chaque principe, les questions viseront à évaluer :

a. si des garanties procédurales suffisantes ont été mises en place pour s'assurer que le système est conforme à la recommandation ;

b. les résultats positifs (potentiels) et les impacts négatifs qui peuvent découler de l'acquisition et du déploiement du système, spécifiques à son contexte d'utilisation.

L'outil d'évaluation est disponible sur : https://unesdoc.unesco.org/ark:/48223/pf0000386276

ANNEXE II
EXEMPLES D'ACTIVITÉS SUPPLÉMENTAIRES

- Case studies of AI systems in public services in Latin America - http://webfoundation.org/docs/2018/09/WF_AI-in-LA_Report_Screen_AW.pdf
- Interactive activity (courtroom algorithm game) of using Compas - https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm/
- Trustworthy AI Playbook - https://www.hhs.gov/sites/default/files/hhs-trustworthy-ai-playbook.pdf
- Algorithmic impact assessment activity, Government of Canada, Algorithmic Impact Assessment. https://canada-ca.github.io/aia-eia-js
- Assessment List for Trustworthy AI (ALTAI) exercise

Outil de cartographie d'IA[427]

#	Question	Réponses
1	Quel est le nom de l'outil d'intelligence artificielle évalué par ce questionnaire ?	
2	Décrivez brièvement les principales fonctionnalités de l'outil.	
3	Qu'est-ce qui motive l'utilisation d'outils d'IA dans ce cas ? (Cochez toutes les réponses appropriées)	1) Arriéré de travail ou de cas existant 2) Amélioration de la qualité globale des décisions 3) Réduction des coûts de transaction d'un programme existant 4) Réalisation de tâches que les humains ne pourraient pas accomplir dans un délai raisonnable 5) Utilisation d'approches innovantes 6) Autre
4	Comment cet outil a-t-il été développé ?	1) Entièrement développé par le personnel technique de votre établissement 2) Développé en collaboration avec une entité externe 3) Acheté, développé entièrement par une partie externe 4) Je ne sais pas 5) Autre
5	Pour quelle plateforme de justice électronique cet outil est-il développé ?	

427 Brehm K., Hirabayashi M., Langevin C., Munoscano B.R., Sekizawa K., Shu J. (2020). The future of ai in the brazilian judicial system - ai mapping, integration and governance. The Future of AI in the Brazilian judicial System. AI Mapping, Integration, and Governance, Technical report, ITS Rio, disponible sur : https://itsrio.org/wp-content/uploads/2020/06/SIPA-Capstone-The-Future-of-AI-in-the-Brazilian-Judicial-System-1.pdf

#	Question	Réponses
6	À quel stade de développement se trouve actuellement l'outil ?	1) En cours de développement/processus d'approvisionnement en cours 2) Prototype / Essais 3) Prêt pour le déploiement, ne fonctionne pas actuellement 4) Entièrement déployé
7	Sur quelles méthodes l'outil est-il basé ?	1) Régression logistique 2) Machines à vecteurs de support 3) Arbres de décision / Forêt aléatoire 4) Réseaux neuronaux / CNN 5) Méthodes de suréchantillonnage / rééchantillonnage 6) Méthodes de réduction de la dimensionnalité (PCA, Clustering, Manifold Learning) 7) Autre :
8	Veuillez vérifier lesquelles, le cas échéant, des fonctionnalités suivantes s'appliquent à l'outil. (Cochez toutes les réponses appropriées)	1) Modélisation et évaluation des risques : analyse des ensembles de données pour identifier les modèles et recommander des lignes de conduite et, dans certains cas, déclencher des actions spécifiques. 2) Organisation des données : analyse des données pour catégoriser, traiter, trier, personnaliser et servir un contenu spécifique pour des contextes spécifiques. 3) Reconnaissance d'image et d'objet : analyse des données pour automatiser la reconnaissance, la classification et le contexte associés à une image ou à un objet. 4) Analyse du texte et de la parole : analyse des données pour reconnaître, traiter et étiqueter le texte, la parole, la voix et faire des recommandations, des classifications ou d'autres types de résultats en fonction du marquage. 5) Optimisation des processus et automatisation des flux de travail : analyse des données pour identifier les anomalies, les modèles de cluster, prédire les résultats ou les moyens d'optimiser ; et automatiser des flux de travail spécifiques. 6) Aucun / Non applicable 7) Autre
9	L'outil effectue-t-il une analyse quelconque des données non structurées ?	1) Oui 2) Non 3) Je ne sais pas.
10	Les données qui ont été utilisées pour former l'outil sont-elles connues de l'équipe qui les utilise ?	1) Oui 2) Non 3) Je ne sais pas. 4) Non applicable

#	Question	Réponses
11	Le code de l'outil est-il accessible au public et révisable ?	1) Oui 2) Non 3) Je ne sais pas. 4) Non applicable
12	L'algorithme de l'outil et son code sont-ils...	1) Open source 2) Propriété du tribunal 3) Propriété d'un tiers
13	L'outil collecte-t-il et/ou analyse-t-il des données personnelles (telles que définies par la loi générale sur la protection des données) ?	1) Collecte 2) Analyse 3) Ni l'un ni l'autre
14	L'outil collecte-t-il et/ou analyse-t-il des informations personnellement identifiables ?	1) Collecte 2) Analyse 3) Ni l'un ni l'autre
15	Les données utilisées par l'outil... (Cochez toutes les réponses appropriées)	1) Ont été collectées par un tribunal ou une entité gouvernementale. 2) Sont accessibles au public et vérifiables. 3) Sont partagées avec une autre entité. 4) Ont été collectées par une entité externe. 5) Sont partagées avec une entité externe.
16	Le personnel technique de votre établissement peut-il expliquer :	1) Quelles sont les entrées de l'outil. 2) Quels sont les résultats de l'outil. 3) Le processus par lequel les entrées deviennent des résultats.
17	Le personnel non technique de votre établissement peut-il expliquer :	1) Quelles sont les entrées de l'outil. 2) Quels sont les résultats de l'outil. 3) Le processus par lequel les entrées deviennent des résultats.
18	L'outil est-il passé par :	1) Un suivi technique et des processus d'assurance qualité. 2) Un examen de ses données de formation pour détecter les biais. 3) Un examen juridique et/ou administratif. 4) Autre

ANNEXE III
PROGRAMME DE FORMATION - MODÈLE

Manuel de formation mondial : l'IA et l'état de droit pour le pouvoir judiciaire

Formation de 3 jours

Date :

Titre	Formation pour [insérer le public cible] sur le Manuel de formation mondial de l'UNESCO : l'IA et l'état de droit pour le pouvoir judiciaire
Modalité	Physique
Public cible	
Dates	
Durée	3 jours
Description	Le programme de formation est basé sur le manuel de formation mondial : l'IA et l'état de droit pour le pouvoir judiciaire.
Organisme	La formation sera organisée par l'UNESCO
Date limite d'inscription	
Frais de formation	
Langue	

1. OBJECTIFS D'APPRENTISSAGE

Ce programme de formation est destiné à fournir aux opérateurs judiciaires l'accès aux informations et aux outils nécessaires pour comprendre et prendre en compte les avantages de l'intelligence artificielle (« IA ») dans leurs opérations. Dans le même temps, le programme de formation aidera le pouvoir judiciaire à reconnaître les inconvénients et les risques de l'IA, notamment les préjugés, la discrimination, les boîtes noires et le manque de responsabilité et de transparence. Le programme de formation aidera les opérateurs judiciaires à mieux juger et à réduire les risques potentiels pour les droits humains, en offrant des conseils et des perspectives sur les principes, les réglementations et la jurisprudence pertinentes qui sous-tendent l'utilisation responsable de l'IA dans les contextes judiciaires et en général.

Pour équilibrer les opportunités et les défis que les technologies de l'IA peuvent représenter pour le secteur judiciaire, la Recommandation de l'UNESCO sur l'éthique de l'IA souligne que « les États membres doivent renforcer la capacité du pouvoir judiciaire à prendre des décisions relatives aux systèmes d'IA conformément à l'état de droit… ». D'où l'importance de ce programme de formation pour élaborer la manière dont la justice peut tirer parti des technologies de l'IA et s'assurer qu'elles sont utilisées de manière éthique, responsable et conformément au cadre du droit international relatif aux droits de l'homme.

2. RÉSULTATS D'APPRENTISSAGE

À la fin de ce programme de formation, les opérateurs judiciaires seront en mesure :

- D'acquérir une compréhension de l'IA et de la prise de décision algorithmique (ADM) et de son utilisation dans les processus et opérations judiciaires.
- De comprendre que l'IA n'est pas neutre et qu'il s'agit d'un système sociotechnique qui représente le monde qui nous entoure.
- De développer une capacité à examiner les affaires juridiques liées à l'utilisation de l'IA.
- De comprendre les problèmes clés liés aux préjugés algorithmiques (tels que les préjugés sexistes, les préjugés raciaux, les formes croisées de préjugés, etc.) et aux boîtes noires, et expliquer pourquoi ils sont importants dans les contextes judiciaires.
- De se familiariser avec les mesures réglementaires et la jurisprudence les plus récentes liées aux biais algorithmiques, à l'utilisation inappropriée des algorithmes dans la prise de décision, y compris en violation de la loi, et aux boîtes noires.
- De comprendre et d'expliquer l'impact de l'IA sur les droits fondamentaux suivants : vie privée, liberté d'expression, accès à l'information, protection contre la discrimination, droit d'accès au tribunal, procès et audiences équitables et impartiaux, et procédure régulière.

3.PUBLIC CIBLE

Le public cible principal de la formation est constitué d'opérateurs judiciaires, principalement des juges. La formation peut également inclure des procureurs, des avocats d'État, des avocats publics, d'autres parties prenantes du secteur judiciaire dans le monde entier et des entreprises de technologie juridique.

4. CONDITIONS D'ADMISSION

Lecture : Manuel de formation mondial : l'IA et l'état de droit pour le pouvoir judiciaire

5. FORMATEURS

NOM DES FORMATEURS	CONTACTS

6. CONTENU DE LA FORMATION

Le programme de formation est principalement basé sur le manuel de formation mondial sur l'IA et l'état de droit pour le pouvoir judiciaire et couvre les sujets suivants :

1. Module 1 : Introduction à l'IA et à l'état de droit
2. Module 2 : Adoption de l'IA dans le système judiciaire
3. Module 3 : Défis juridiques et éthiques du déploiement de l'IA
4. Module 4 : Droits humains et IA

7. CONTENU ET ORDRE DU JOUR DU PROGRAMME DE FORMATION

Jour 1 : Introduction à l'IA et à son utilisation dans le système judiciaire

Heure	Programme
8 h 30 – 9 h	Connexion et inscription des participants
9 h 00 – 9 h 30	Ouverture et introduction aux objectifs du programme de formation
9 h 30 – 11 h 00	Session 1 : Comprendre l'IA et ses éléments constitutifs Animateur : Cette session vise à fournir une compréhension globale de l'IA, en explorant sa définition et ses principaux éléments constitutifs. À travers des discussions impliquantes, des exemples illustratifs, des études de cas et des activités de groupe, nous examinerons les différentes composantes des systèmes d'IA, y compris les algorithmes, l'apprentissage automatique, les données et les modèles informatiques. Cette session abordera également les principaux risques liés au développement et au déploiement de l'IA, tels que les préjugés, les boîtes noires et la cybersécurité. À la fin de cette session, les participants auront acquis une solide compréhension des concepts clés liés à l'IA, ce qui leur permettra de naviguer dans le domaine avec confiance et clarté.
11 h 00 – 11 h 30	Pause-café

11 h 30 – 13 h 00	Session 2 : Quelles sont les applications de l'IA dans le secteur de la justice ? Animateur : Cette session décrira certaines des principales applications de l'IA dans le système judiciaire, telles que la découverte électronique et l'examen des documents, l'utilisation de l'IA générative pour aider à la rédaction de documents, l'analyse prédictive et le soutien ADM, les outils d'évaluation des risques, le règlement des différends, la reconnaissance et l'analyse linguistiques, les fichiers numériques et la gestion des cas.
13 h 00 – 14 h 30	Déjeuner
14 h 30 – 16 h 00	Session 3 : Études de cas sur l'utilisation de l'IA dans le système judiciaire Animateur : Cette session examinera certaines études de cas sur le déploiement de l'IA dans le système judiciaire, telles que VICTOR (Brésil), le système de transcription judiciaire intelligent de Singapour, Prometea (Argentine), PretorIA (Colombie), l'utilisation de l'IA dans le système judiciaire chinois, l'utilisation de l'IA dans le système judiciaire indien, HART (Harm Assessment Risk Tool, Royaume-Uni), PredPol et Palantir. La session invitera les participants à partager leur expérience des systèmes d'IA et à engager une conversation plus large sur les opportunités, les défis et les risques associés à l'utilisation de ces systèmes dans le secteur judiciaire.
16 h 00 – 16 h 30	Rétroaction et évaluation
16 h 30 – 16 h 45	Conclusion de la première journée et aperçu de l'ordre du jour de la deuxième journée

Jour 2 : Questions juridiques et éthiques liées aux systèmes d'IA

Heure	Programme
8 h 30 – 9 h	Connexion et inscription des participants
9 h – 11 h	Session 4 : Responsabilité algorithmique et transparence Animateur : À travers des discussions pertinentes et des études de cas du monde réel, cette session guidera les participants à travers les concepts de transparence algorithmique et les concepts de responsabilité, et mettra en évidence les questions juridiques clés dont les opérateurs judiciaires doivent être conscients. Une attention particulière sera accordée à l'identification biométrique, à la reconnaissance faciale et aux deepfakes.
11 h 00 – 11 h 30	Pause-café

11 h 30 – 13 h 00	Session 5 : Jurisprudence émergente sur les préjugés et les boîtes noires Animateur : La session présentera la jurisprudence existante qui traite des boîtes noires algorithmiques et des biais dans les ADM et les systèmes d'IA utilisés dans la prestation de services publics et privés. À l'aide d'études de cas réelles, les participants discuteront de la manière dont les préjugés et les boîtes noires ont entraîné une violation des droits humains ou tout autre préjudice, et de la manière dont les tribunaux de différentes juridictions ont traité ce problème. Les participants discuteront des questions de responsabilité pour les dommages causés par ces systèmes, ainsi que de l'utilisation de l'IA à des fins de preuve. Les affaires examinées comprendront : l'affaire Deliveroo (2021), l'affaire Foodinho (2021), l'affaire People c. Chubbs (2015), l'affaire État du New Jersey c. Francisco Arteaga, l'affaire État c. Loomis, l'affaire People c. Alvin Davis, l'affaire État du New Jersey c. Pickett, l'affaire Uber concernant l'utilisation du programme de détection des fraudes Mastermind États-Unis c. Ellis et l'affaire australienne Robodebt.
13 h 00 – 14 h 30	Déjeuner
14 h 30 – 16 h 00	Session 6 : Évaluation de l'impact éthique des systèmes d'IA Animateur : Cette session présentera aux participants les principales questions liées à l'éthique de l'IA, ainsi que les principaux cadres d'éthique de l'IA aux niveaux international, régional et national. À l'aide de l'évaluation de l'impact éthique des systèmes d'IA de l'UNESCO, les participants évalueront des scénarios hypothétiques dans le cadre de groupes de discussion.
16 h 00 – 16 h 30	Rétroaction et évaluation
16 h 30 – 16 h 45	Conclusion de la deuxième journée et aperçu de l'ordre du jour de la troisième journée

Jour 3 : IA et droits humains

Heure	Programme
8 h 30 – 9 h	Connexion et inscription des participants
9 h – 11 h	Session 7 : Droits humains et IA : droit à l'accès au tribunal, procès équitable et procédure régulière, recours effectif et droits à la protection contre la discrimination. Animateur : Les applications de l'IA peuvent affecter directement l'égalité d'accès aux droits fondamentaux, y compris le droit à la vie privée et à la protection des informations personnelles, le droit à l'accès à la justice et le droit à un procès équitable, en particulier en ce qui concerne la présomption d'innocence et la charge de la preuve, le droit à l'emploi, à l'éducation, au logement et à la santé, ainsi que le droit aux services publics et à la protection sociale. Si elles ne sont pas accompagnées de garanties adéquates contre les préjugés, les technologies de l'IA pourraient contribuer à refuser l'accès aux droits qui affectent de manière disproportionnée les femmes, les minorités et ceux qui sont déjà les plus vulnérables et les plus marginalisés.

11 h 00 – 11 h 30	Pause-café
11 h 30 – 13 h 00	• Session 8 : Droits humains et IA : (i) liberté d'expression, (ii) droit à la vie privée et à la protection des données, et (iii) accès à l'information. Animateur : Cette session présentera et abordera certains des droits humains impactés par les systèmes d'IA déployés par des tiers et jugés par les tribunaux, tels que la liberté d'expression, le droit à la vie privée et à la protection des données, et l'accès à l'information. La session abordera également la jurisprudence pertinente liée aux droits humains et aux applications de l'IA.
13 h 00 – 14 h 30	Déjeuner
14 h 30 – 16 h 00	Session 9 : Problèmes émergents à l'intersection de l'IA et du droit Animateur : - La session abordera brièvement les préoccupations concernant : - La cybersécurité - Les droits de propriété intellectuelle - Les preuves générées par l'IA devant les tribunaux - L'utilisation de la RA et de la RV devant les tribunaux
16 h 00 – 16 h 30	Rétroaction et évaluation
16 h 30 – 17 h 00	Synthèse et conclusion du programme

8. MÉTHODOLOGIE (approche didactique)

Le programme de formation est basé sur le manuel de formation mondial relatif à l'IA et l'état de droit pour le pouvoir judiciaire. Le manuel de formation comprend des activités, du contenu et des ressources pertinents pour l'IA, les droits humains et l'état de droit, pour les opérateurs judiciaires.

Cette formation sera dispensée physiquement et comprendra des conférences, des exercices interactifs et des discussions. La formation s'appuiera sur des diapositives PowerPoint, des documents de référence sélectionnés et des questionnaires d'auto-évaluation quotidiens. Les participants doivent réviser, étudier, participer aux activités prévues et entreprendre des auto-évaluations.

9. ÉVALUATION ET NOTATION

Les performances des participants à cette formation seront évaluées à l'aide d'une combinaison de notes sur la participation aux discussions des sessions et les questionnaires d'auto-évaluation.

- La participation aux sessions comptera pour 30 % de la note finale.
- Les questionnaires d'auto-évaluation représenteront 70 % de la note finale de la formation. Il y aura 6 questions par questionnaire.

À la fin, les participants recevront un certificat d'achèvement.

10. PRÉSENTATIONS DE FORMATION

Un certain nombre de présentations pour chaque module et dans différentes langues qui peuvent être utilisées pour les formations sont disponibles sur : Digital Innovation & Transformation (CI/DIT)