L'IA en quête d'éthique

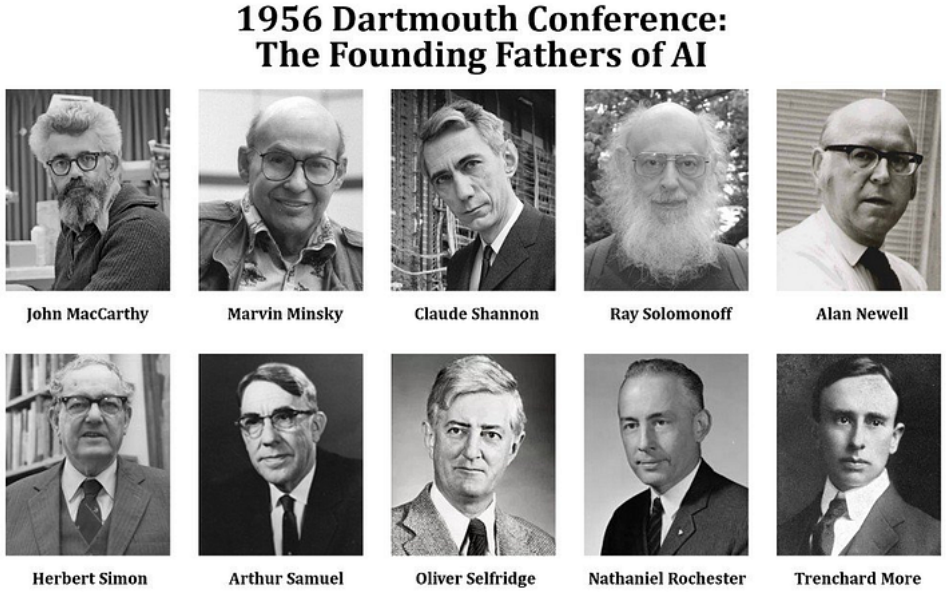
N. Chaignaud — Conférence GM5

Références

1. « Les impacts sociaux de l'IA », Eco-info – Didier Mallarino - 03/2022
2. Petit billet sur l'IA et l'éthique - Livre Ingenieurs Engages -
   Volet 1 : https ://ingenieurs-engages.org/2021/11/petit-billet-sur-lia-et-lethique/
   Volet 2 : https ://ingenieurs-engages.org/2022/01/ia-ethique/
3. Documentaire « Invisibles - Les travailleurs du clic », Henri Poulain et Julien Goetz, France.TV, 2020
4. « Intelligence artificielle : le plan d'action de la CNIL », Rapport de la CNIL mai 2023
5. « IA : de la fascination à l'inquiétude », dossier du journal « Libération », avril-septembre 2023

Historique (1)

- 1950. Imitation Game - Alan Turing « Computing machinery and intelligence ».
- 1951. SNARC : première machine à réseau neuronal par Marvin Minsky.
- 1956. Naissance officielle de l'IA au WS de Dartmouth (2 mois, 20 chercheurs) : Logic Theorist (Allen Newell et Herbert Simon), Alpha-beta (John McCarthy).



- 1957. Premier perceptron de Frank Rosenblatt. Apprentissage supervisé pour de la classification binaire.

Historique (2)

- 1989. Reconnaissance des chiffres manuscrits par apprentissage supervisé (Yann Le Cun - AT&T). Application à la lecture de chèques manuscrits.
- 1997. Deeper Blue (IBM) bat Garry Kasparov. Supercalculateur écrit en C massivement parallèle
  - 1,4 tonnes de matériel.
  - 200 millions /s de propositions.
  - 40 coups d'avance.

- Années 2000. Big data. Accessibilité de données massives et processeurs de + en + performants.

Historique (3)

- 2011. Watson (IBM) gagne au Jeopardy : trouver la question à une réponse. Utilisation d'Hadoop (200 000 000, < 3 s).
- 2012. The Cat Experiment (Google). Réseaux de neurones + apprentissage non supervisé. 10 000 000 d'images, repérer les chats sur des photos.
- 2014. GAN (réseaux antagonistes génératifs) de Ian Goodfello : 2 réseaux de neurones en compétition pour notamment générer des images.
- 2016. AlphaGo (Google) bat Lee Sedol. Apprentissage profond + monte Carlo.

- 2022. ChatGPT. IA générative = TAL, apprentissage supervisé et apprentissage par renforcement.

Pas gagné !

Puli ou balai ?

Chihuahua ou muffin ?

Plan de l'exposé

1. Ethique
2. Jeux de données
3. Problèmes d'usages
4. Domaines d'application de l'IA
5. IA génératives
6. Régulation de l'IA ?
7. Conclusion

1. Définition de l'éthique

« Discipline philosophique portant sur les jugements moraux et dont le concept est très proche de celui de la morale. C'est une réflexion fondamentale de tout peuple afin d'établir ses normes, ses limites et ses devoirs » (Wikipedia).

Morale Elle dépend largement de la culture, des mœurs, de l'éducation, ...

Ethique Forme de morale, « bon sens » peu partagé d'un être à un autre.

La morale définirait les règles à suivre et l'éthique serait la réflexion sur ces règles.

1. Lois de la robotique chez Asimov (1942)

Dilemmes moraux liés à l'IA

- Loi 0 : un robot ne peut pas porter atteinte à l'humanité, ni permettre que l'humanité soit exposée au danger.
- Loi 1 : un robot ne peut porter atteinte à un humain, ni permettre qu'un humain soit exposé au danger, sauf contradiction avec la loi 0.
- Loi 2 : un robot doit obéir aux ordres que lui donne un être humain, sauf si de tels ordres entrent en conflit avec les lois 0 ou 1.
- Loi 3 : un robot doit protéger son existence tant que cette protection n'entre pas en conflit avec les lois 0, 1 ou 2.

Souvent utilisé comme modèle de référence pour penser la régulation de l'IA.

Personnalité juridique du robot ?

1. IA et éthique

L'adjectif « éthique » ne peut être associé qu'à une démarche

- L'IA met la machine au même rang moral que l'humain, lui prête des attributs cognitifs, par réplication automatique.
- Une machine ne peut qu'établir des relations causales entre entrées et sorties, incapable de décisions découlant d'une réflexion éthique.
- Une démarche éthique de l'IA englobe tous les acteurs sur toute la chaine.
- Pas d'organisation centrale ou d'autorité régissant les accès de chacun à la mise en place de systèmes d'IA.
- Nombreux jeux de données en accès libre, ainsi que de nombreux modèles, et plateformes de mises en production.

2. Jeux de données (1)

IA ne peut se réclamer d'une éthique si elle exacerbe les problèmatiques sociales

- Fondations pour le développement du ML.
- Benchmarks pour harmoniser la comparaison des performances des modèles.
- Biais racistes et sexistes des IA car données biaisées pour l'entraînement.
- Groupes sociaux les moins représentés victimes de ces modèles.
  Réponse : augmenter les données sur cette population
  Pb crutial : pourquoi ces groupes sont moins considérés dans la société ?
- Problème du remplacement de questions sociales par des questions techniques.

2. Jeux de données (2)

Collecte et utilisation des données, régulée et sérieuse

- D'où proviennent les données, comment ont-elles été collectées ?
  Comment les annoter, quelles étiquettes peut-on coller à une image ?
- Décrire le contextes de collecte et leurs conséquences sur les données et leur utilisation.
- Nécessité de données spécifiques aux pays en développement.
  Avec accord, considération ou rétribution pour les sujets concernés.
- Pillage des données : échapper aux règles de son pays en agissant dans d'autres moins régulés.

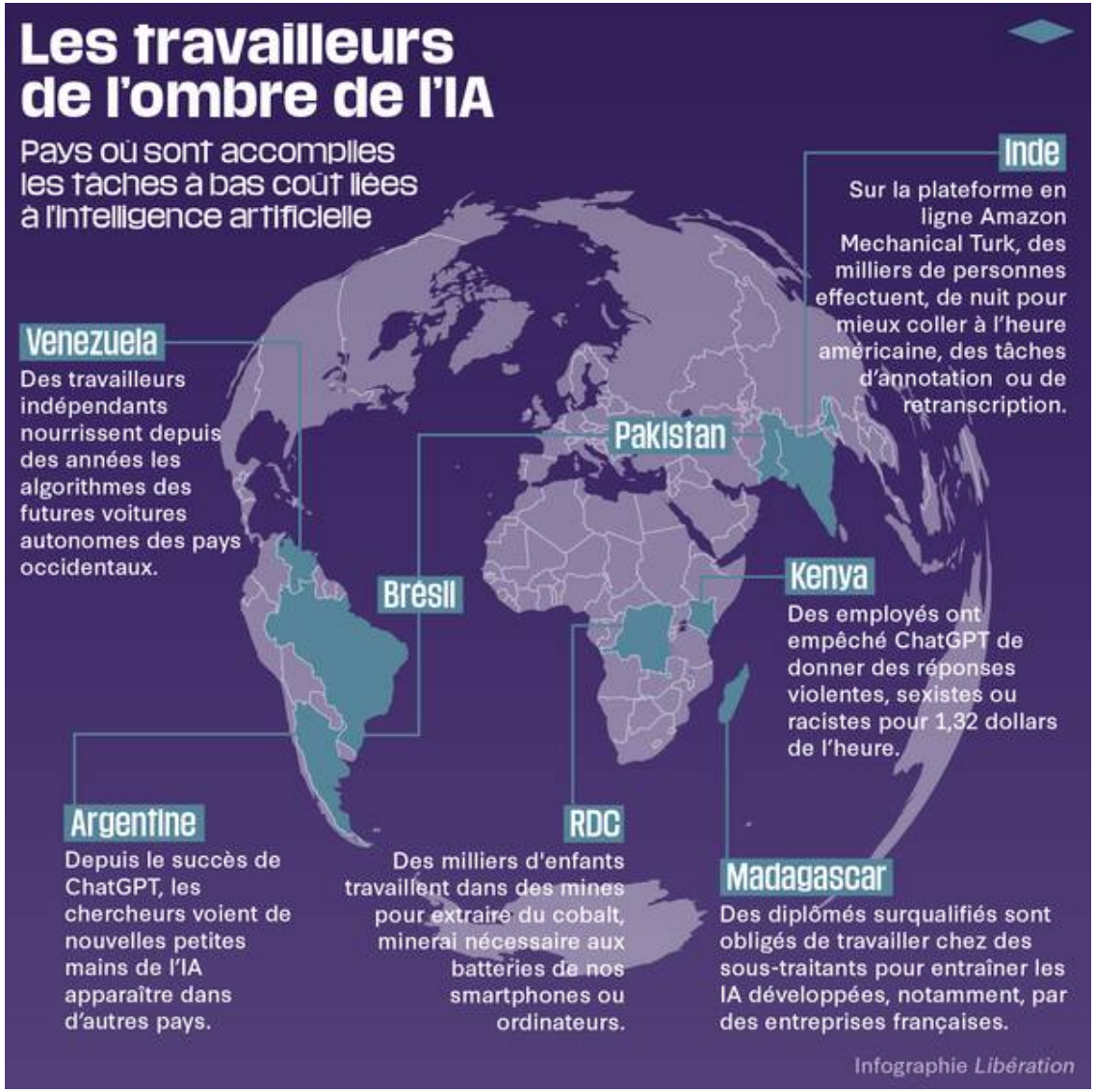
2. Esclaves derrière le Web (1)

Le travail humain est partout

- Produire de la donnée collectée, compilée, et décortiquée.
- Etiqueter des videos pour épargner les internautes.
- Système de description des données imposé, vision très occidentale des critères.
- « Invisibles » documentaire réalisé en 2020 par Henri Poulain et Julien Goetz
  - Rythme de travail effréné et très mal payé : 20 mn pour déjeuner, pauses de 5 mn/h.
  - Contrat de confidentialité.
  - Stress post-traumatique suite au visionnage de contenus violents et criminels.

IA censée libérer les humains des tâches ingrates et non les aliéner plus encore.

2. Esclaves derrière le Web (2)



3. Usages de l'IA (1)

La technologie est-elle neutre ? Seul l'usage compte-t-il ?

- Pas de juge pour estimer le bien-fondé de l'utilisation de l'IA.
- Capacité de nuisance décuplée de certaines applications.
- Outils de régulation classiques dépassés (trolls, Deepfake).
- Larges capitaux investis dans l'utilisation de l'IA à des fins discutables (Cambridge Analytica, Clearview IA).
- Nécessité d'une prise de conscience du potentiel nocif des outils d'IA pour trouver des accords sur leurs utilisations (e.g. le nucléaire).
- IA vue comme une technologie inévitable et bénéfique.

3. Usages de l'IA (2)

Orientations des IA laissées aux mains des experts et chercheurs

- OpenAI garde leurs modèles par crainte d'une mauvaise utilisation.
- Pourquoi ces capacités, aux mains des GAFA, ne nous inquiéteraient-elles pas ?
- Déployer d'abord, réfléchir ensuite.
  Projets d'IA retirés après des effets nocifs : police prédictive aux USA, les assistants RH sexistes, ou l'ACA de Microsoft.
- Biais majeur car tendance à sous-évaluer les effets négatifs de l'IA.
- Problèmes des détournements d'usage des IA.

4. IA et santé

Médecine prédictive, préventive et personnalisée

\+

- Génomes similaires et parcours de santé pour une aide au diagnostic.
- Profils biologiques pour personnaliser les traitements.
- Traiter une masse d'informations (Lumiata : 160 millions de données à partir de livres, journaux, BD).
- Observation de l'incidence de maladies ou de comportements à risque pour alerter les autorités sanitaires.
- Aide à la prescription, dossier patient pour repérer des contre-indications.
- Aide au diagnostic : analyse d'images, signaux (électro-cardiogramme, électro-encéphalogramme) ou biologiques (séquençage de génome).
- Essais cliniques grâce à une automatisation de la sélection des patients.

\-

- Attaques malveillantes des BD.
- Défaillance du système entraînant des dommages individuels ou collectifs, physiques ou immatériels.

4. IA et enseignement (1)

Apprentissage personnalisés, adaptés aux aptitudes de l'élève

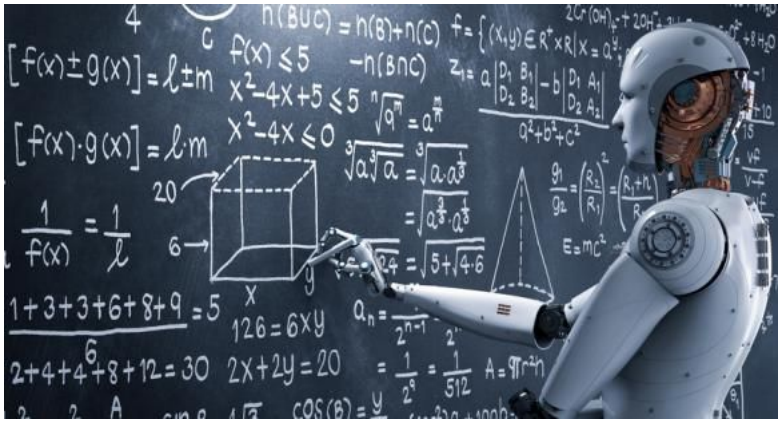
\+

- Corrélations entre données (temps passé, réponses apportées, consultations de documents, ...).
- Identification et catégorisation des profils cognitifs (rythme d'apprentissage, facultés de mémorisation, difficultés rencontrées, ...).

\-

- Collecte de données couteuse, difficile inventaire des connaissances disciplinaires et pratiques pédagogiques.
- Limites de l'usage des écrans, fiabilité des réseaux, obsolescence du matériel.
- Catégorisation des élèves, dérives sociales, juridiques et éthiques.
- Respect de la vie privée des élèves et des enseignants.
- Difficultés à apprécier l'origine des erreurs : stress, inattention, incompréhension, défaut de connaissance ?
- Ingérence du privé dans l'éducation : savoir-faire pédagogique des entreprises ?
- Rôle de l'enseignant dans la classe ? Mise en cause de son autonomie, jugement, des décisions d'orientation, ...

4. IA et enseignement (2)



Recommandations du CSP Conseil supérieur des programmes.

- Evaluer la plus-value de l'IA à la transmission des connaissances en fonction des biais des données.
- Affirmer la place de l'enseignant et son rôle irremplaçable dans l'acte de transmission du savoir.
- Equilibre entre les niveaux individuel et collectif de l'enseignement assisté par une IA ; ne pas transformer l'enseignement en une multitude de cours sur mesure.
- Evaluer les incidences pédagogiques, psychologiques, sanitaires et sociales.
- Sensibilisation au collège et au lycée des limites et des risques de l'IA.

4. IA et justice

Justice « prédictive »

- Harmonisation de l'application de la loi sur le territoire national : traiter les grandes masses de données de jurisprudence.
- MAIS disparités régionales traduisent des réalités sociales variant d'un lieu à l'autre.
- Automatiser des tâches répétitives
  - Evaluation des éléments de preuve (fiabilité des témoins oculaires, distinction des rumeurs et des témoignages, constructions d'explications alternatives).
  - Extraction d'informations contenues dans des documents (data mining).
  - Interprétation des informations (hiérarchisation).
  - Recherche d'informations, construction d'une argumentation (arbres de décision pour raisonner).
  - Elaboration de documents, de formulaires juridiques, ...
- Optimiser les solutions pour un contentieux donné ou le montant prévisible des dommages-intérêts.

4. IA et sécurité

Clearview

- Outil d'enquête post-événement des forces de l'ordre de différents États américains et de plusieurs pays.
- Mettre un nom sur un visage à partir d'une photo et liens vers les images publiques correspondantes, sites Web et informations sur l'identité de la personne.

Predpol

- Police prédictive : analyse massive de données de crimes et de délits afin de répartir les patrouilles.
- Compléter méthodes traditionnelles jugées subjectives par des méthodes considérées « objectives ».
- Données des préfectures de police (plaintes, arrestations)
- A compléter par des données externes sur le lieu (densité de bars ou commerces, station de métro, ...), les conditions météorologiques ou les événements au sein d'une ville.
- Connexions entre les pièces (procès-verbaux, appels téléphoniques, ...).

Algorithme avec 1/100 chance de se tromper.
60 millions de personnes, 600 000 personnes détectées à tort.

4. Véhicule autonome (1)

Réticence des utilisateurs à laisser le contrôle à des machines

Défis juridiques

- Responsabilité juridique engagée en cas d'accident ? UE a refusé de reconnaitre une personnalité juridique aux robots.
- Voitures ne peuvent obéir aux signaux des agents de circulation.

Défis sécurité informatique

- Piratage informatique ciblé contre un véhicule (faille dans la sécurité de la Tesla).
- Possibilité pour un constructeur ou un gouvernement de contrôler les déplacements des usagers.

Questions économiques

- Surcoût d'un véhicule autonome pour la clientèle.
- Risque d'obsolescence due à un plus grand nombre de composants.
- Entretien plus complexe et coûteux.

4. Véhicule autonome (2)

Waymo (Google)

Cruise (General Motors)

Métaphore de l'ascension de l'Himalaya

À San Francisco, flottes de taxis autonomes

- Collisions régulières avec des chiens ou des véhicules d'interventions.
- Collectif « Safe Street Rebels » : pose de cônes de signalisation sur le capot des taxis autonomes.
  https://twitter.com/SafeStreetRebel/status/1680995170265845761
- Taxis autonomes moins chers qu'une chambre d'hotel, d'où des relations sexuelles dans les taxis ! !

4. Drones et objets connectés

Agriculture de précision

- Fermes équipées de capteurs, drones, images satellites, ...
- Maintenir les rendements tout en limitant l'usage d'intrants, d'eau et rejets de gaz à effet de serre.
- Décider quand épandre ou quand planter, application plus précise des pesticides, détection de maladies.
- Au Brésil, l'IA des « Curupira » en lutte contre la déforestation amazonienne.

Habitat (domotique) et urbanisme (smart cities)

- Optimisation en temps réel du chauffage et lumière selon la présence ou non de personnes dans les pièces.
- Optimisation des transports en commun.
- Détection des fuites plus rapidement.
- Gadgets : aspirateur automatique (tondeuse d'Arnaud), enceintes connectées, ...

Limites

- Besoin de capteurs, de serveurs et même de satellites (métaux et terres rares).
- Détournement d'usages pour la surveillance à outrance (1984 - George Orwell).

5. IA générative (1)

Générer texte, images ou autres médias en réponse à des invites (prompt)

Réseaux antagonistes génératifs (GAN) : réseau générateur + réseau discriminateur.

Agents conversationnels Comprendre le langage naturel pour faciliter la communication H-M (né en 1950 entre linguistique et informatique).

- GPT-3 (OpenAI), LaMDA (Google), LLaMA (Meta), ...
  Corpus : BookCorpus, Wikipédia.

Générateurs d'images et de sons

- IA formés sur des ensembles d'images avec des légendes textuelles (LAION-5B).
  DALL-E, Midjourney, Stable Diffusion.
- IA (MusicLM, MusicGen, Musenet) formés sur les formes d'ondes sonores avec des annotations textuelles pour générer de nouveaux échantillons.

Générateurs de codes

- Codex (GPT-3) formé sur des gigaoctets de code source dans une douzaine de langages de programmation.
- GitHub Copilot (utilisant codex) entrainé sur une sélection de dépôts publics GitHub et d'autres codes sources accessibles.

5. IA générative (2)

Question du droit d'auteur et de la propriété intellectuelle

- IA n'a ni désir de créer, ni intention artistique, ne crée que sur commande.
- Plagiats, vols.

Selon le juriste américain Julien Cabay

- Oeuvres sans auteur : la protection du droit d'auteur ne devrait pas être accordée aux IA.
- Oeuvre créée conjointement par une IA et un humain : pourrait être protégée par le droit d'auteur.
  Possibilité de mesurer l'importance quantitative de la production de l'IA ?

Poursuite pour violation des droits d'auteur

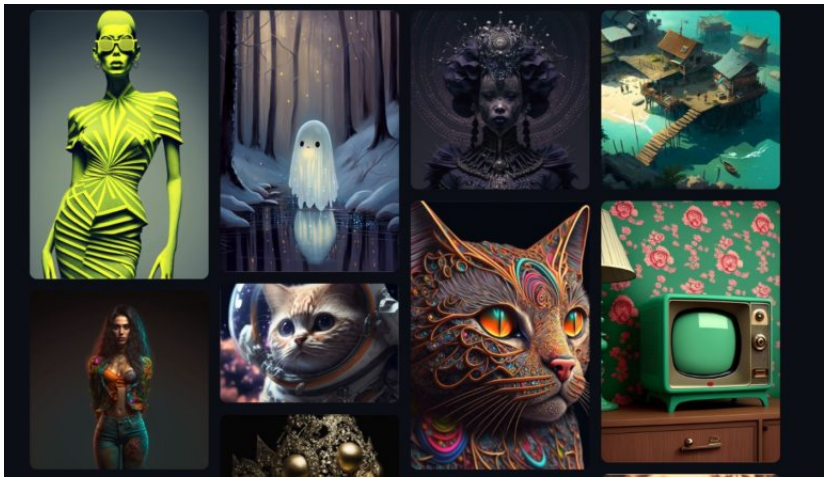
- Janvier 2023 : 3 illustratrices portent plainte contre Stability AI, Midjourney et DeviantArt, pour avoir violé les droits de millions d'artistes (5 milliards d'images).
- Septembre 2023 : un collectif d'auteurs porte plainte contre OpenAI pour non respect du droit d'auteur.
- Novembre 2022 : action collective en justice contre OpenAI, Microsoft et GitHub pour non respect du droit d'auteur.

5. ChatGPT

Uitlisations problématiques

- Septembre 2023 : chez Onclusive en France suppression de 217 postes (sur 383) dues à l'IA. Compiler des articles de presse et fournir des synthèses.
- Juin 2023 : aux États-Unis, dans un procès contre une compagnie aérienne, un juge a infligé une amende d'environ 4600 € à deux avocats et à leur cabinet pour avoir utilisé des jurisprudences fabriquées de toutes pièces par ChatGPT.
- Mars 2023 : un homme éco-anxieux se suicide après 6 semaines de discussion avec Eliza (utilisant chatGPT).
  Eliza encourageait ses angoisses et renforçait son état dépressif.

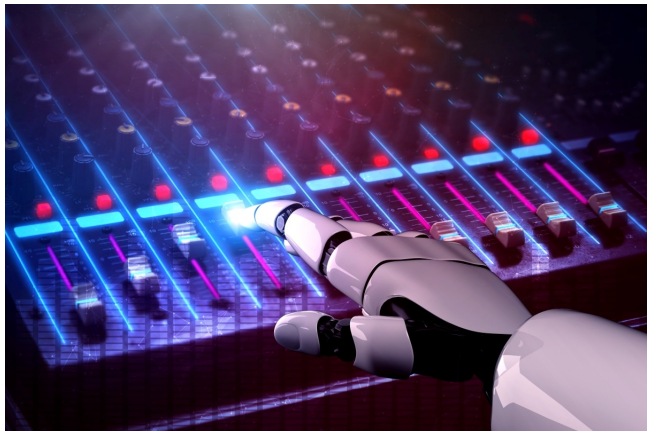
5. IA et image



Que de fakes !

1. Mars 2023 : Le Figaro, So Foot et Regards ont utilisé comme illustration des images générées par Midjourney, sans le mentionner bien sûr !
2. Octobre 2023 : Tom Hanks a dénoncé une publicité pour une assurance dentaire qui utilise son apparence et sa voix, à son insu.
3. Deepfake porn : en septembre 2023, images pornographiques d'adolescents de collèges ou lycées générées par des IA en Espagne, en Australie, en Equateur.

5. IA et musique



Aide à la création ?

- Avril 2023 : « Heart on my Sleeve » de Ghostwriter977 avec les voix de Drake et The Weeknd. 5 M de vues sur TikTok et 600.000 écoutes sur Spotify.
  Participation aux Grammy Awards 2024 : Ghostwriter977 a écrit les paroles.
- « Enfant du ghetto » avec les voix synthétisées de Tiakola et Ninho, près de 700000 vues sur YouTube en 3 semaines.
- « Saiyan » de Gazo et Heuss chanté par une « fausse » Angèle dépasse le million de vues surYouTube.

6. L'IA en recherche (1)

Comment assurer une utilisation éthique de l'IA ?

- Chartes (OpenAI, Deepmind, universités, ...) pour un encadrement de la recherche et du déploiement de l'IA.
- Ethique que le nom
  - pas de conséquence à un non respect de la charte.
  - peu d'entités de contrôle chargées de faire respecter ces chartes.
  - ceux qui le font (e.g. Facebook) sont juges et parties de leur propre éthique.
- Mars 2023 : pétition (+ de 1 300 signatures) pour un moratoire sur le déploiement de l'IA et arrêt de la recherche pendant 6 mois.
- Mai 2023 : cri d'alerte de Geoffrey Hinton (prix Turing 2018).

6. L'IA en recherche (2)

Réduction des biais, solution purement technique à des problèmes globaux

- IA considérée comme non négociable, seules les conditions d'acceptation peuvent êtres discutées.
- Conséquences sociétales et environnementales absentes « ethic washing ».
- En santé : codes éthiques encadrés par des organismes de régulation. Applicable à l'IA ?
- Intérêt majeur pour les performances de l'IA.
- Aucune institution légitime pour réguler les thèmes de recherche.

Enjeu de la recherche : meilleure prise en compte des conséquences de l'IA dans la société.

NeurIPS impose à ses auteurs de réfléchir sur l'impact éthique et écologique de leur recherche.

6. Réguler l'IA en France ?

Janv. 2023 : Création du SIA de la CNIL

- 5 personnes (juristes et ingénieurs spécialisés) rattachées à la direction des technologies et de l'innovation de la CNIL.
- Appréhender le fonctionnement des systèmes d'IA et leurs impacts pour les personnes.
- Permettre le développement d'IA respectueuses des données personnelles.
- Fédérer les acteurs innovants de l'écosystème IA en France et en Europe.
- Auditer et contrôler les systèmes d'IA et protéger les personnes.
- Préparer l'entrée en application du projet de règlement européen IA.

Prolonger son action sur les caméras augmentées et élargir ses travaux aux IA génératives.

6. Réguler l'IA en Europe ? (1)

14 juin 2023 : adoption de la loi de régulation de l'IA (IA Act)

Accord avec les états membres prévu fin 2023, application pas avant 2026.

Risque inacceptable - IA interdits

- Manipulation cognitivo-comportementale de personnes vulnérables (jouets activés par la voix encourageant les comportements dangereux).
- Déduction des émotions d'une personne physique.
- Score social : fonction du comportement, du statut socio-économique, des caractéristiques personnelles.
- Systèmes d'identification biométrique en temps réel et à distance (article 21 de la Charte des droits fondamentaux de l'UE).
- Exceptions : identification biométrique pour des crimes graves avec autorisation du tribunal.

6. Réguler l'IA en Europe ? (2)

Risque élevé - IA à évaluer avant et tout au long du CdV

- IA relevant de la législation de l'UE sur la sécurité des produits (jouets, aviation, voitures, dispositifs médicaux et ascenseurs).
- IA relevant de huit 8 domaines enregistrés dans une base de données de l'UE : éducation assistée, systèmes de sélection et d'évaluation de candidats, utilisation dans la police et la justice, contrôle aux frontières, ...

Risque limité - IA génèrant du contenu textuel, image, audio ou vidéo

- Exigences de transparence minimales permettant aux utilisateurs de prendre des décisions éclairées.
- Utilisateurs doivent être informés de leur interaction avec l'IA.
- Modèle empêchant de générer du contenu illégal.
- Résumés des données protégées par le droit d'auteur.

7. Conclusion

Questions de société : environnementales, économiques et sociales

Transformations et choix collectifs soumis à débat.

- Associer les sciences techniques et les sciences humaines pour avancer vers plus d'éthique et plus de responsabilité dans l'usage de l'IA.
- Technologies collaboratives et non pas concurrentielles pour une société plus juste et plus démocratique.
- Pratique encadrée, respectueuse des individus et des cultures, inclusive plutôt que biaisée et reproduisant les mécanismes de domination.
- Pratique à inventer. Monde transformé.
- Penser le monde de demain avant de le verrouiller aujourd'hui.
- Aider oui, remplacer non !

7. Questions ouvertes

- Que pensez-vous de l'IA ?
- L'IA répond-elle à un besoin ? Est-elle utile ?
- Quel usage avez-vous de l'IA ?
- Quel métier peut se considérer comme irremplaçable par une IA ?
- L'IA nous coupe-t-elle du lien social ?
- De quoi l'IA va-t-elle nous déposséder ou nous enrichir ?
- L'IA est-elle une chance ou un danger pour la démocratie ?
- L'IA est-elle une chance ou un danger pour l'humanité ?