A la recherche des biais de l'IA :

Contexte d'apprentissage	
Intervenant	prof doc et prof de xx
Public	Tout niveau
Horaire	❖ 204 le 16/02 14h30
Lieu	CDI ou salle de classe
Modalités et objectifs d'apprentissage	
Prérequis des élèves	Séance qui fait suite à la génération d'images pour le concours "C'est pas juste"
Objectifs de l'activité	- Comprendre le **fonctionnement de l'IA** - Savoir **identifier les biais de l'IA** - **Développer un esprit critique** - Apprendre à distinguer les contenus produits par l'IA, à évaluer leur fiabilité - Avoir conscience des limites de l'IA (fiabilité et écologie)
Notions (Wikinotions)	**Problématiques soulevées par Intelligence Artificielle** **Données d'entraînement** : La qualité, la quantité et la neutralité des données fournies à l'IA pour son développement et ses créations sont primordiales. Elles ont pour conséquences notamment les biais et la désinformation et l'éthique des usages. **Biais, stéréotypes,discriminations et exclusions** : Parfois volontaires, ces phénomènes sont, le plus souvent, inconscients et difficilement repérables. D'autres comme la surreprésentation des hommes dans les métiers intellectuels est un exemple de biais identifiable. **Désinformation et éthique des usages** : Fake news, production volontaire de fausses informations, et manipulations anti-démocratiques peuvent découler d'utilisations non responsables de l'IA. **Enjeux écologiques** : L'entraînement des IA implique des millions d'heures de calculs informatiques énergivores. Poser 25 questions à ChatGPT coûte 1/2 l. d'eau et les 8 millions de centres de données dans le monde génèrent beaucoup de pollution (refroidissement des centres, métaux rares, ...) **Pour aller plus loin…** https://www.saintrapt.com/les-biais-de-lia-un-jeu-interactif-pour-comprendre-les-biais-des-algorithmes/ **1. Biais présents lors de la collecte, la sélection et la préparation des données** ● **Biais de sélection :** sélection de données, consciente ou inconsciente, qui confirment les croyances ou les hypothèses préexistantes. ● **Biais d'omission :** exclusion systématique, consciente ou inconsciente, de certaines informations. ● **Biais de disponibilité :** tendance à utiliser des données facilement accessibles ou immédiatement disponibles, qui peuvent ne pas être complètes. ● **Biais de représentativité :** les données ne reflètent pas correctement la diversité des personnes ou des cas d'utilisation. Peut découler d'une sélection trop drastique ou binaire. ● **Biais d'étiquetage :** Apparaît lorsque les étiquettes attribuées aux données d'apprentissage

	reflètent des opinions subjectives ou stéréotypées, influençant ainsi le processus d'apprentissage. **2. Biais présents lors de l'entraînement des modèles** ● **Biais d'ancrage :** utilisation d'un ensemble initial d'hyperparamètres, par exemple, ceux utilisés dans des travaux ou des modèles antérieurs, ce qui peut limiter l'exploration de configurations qui pourraient être plus optimales ou moins biaisées. ● **Effet Dunning-Kruger :** une compréhension insuffisante des nuances et complexités des modèles mène à un entraînement inadéquat. ● **Biais du survivant :** utilisation d'hyperparamètres qui ont fonctionné dans des projets ou des expériences antérieurs, en ignorant ceux qui n'ont pas abouti mais qui pourraient être pertinents dans le contexte actuel. ● **Biais de confirmation :** sélection des modèles en fonction de l'adéquation de leurs résultats avec nos propres croyances. **3. Biais présents lors de l'utilisation des modèles** ● **Biais d'automatisation :** confiance excessive dans les performances du modèle après son déploiement, négligeant la nécessité d'une surveillance et d'une évaluation continues. ● **Illusion de connaissance :** croyance en le fait que le modèle comprend les sujets ou répond aux questions de manière plus précise qu'il ne le fait réellement. Particulièrement préjudiciable dans le cas des hallucinations des IAG. ● **Biais de confirmation :** interprétation des résultats du modèle d'une manière qui confirme nos croyances préexistantes. Autre piste : https://pedagogie.ac-toulouse.fr/documentation/enseigner/intelligence-artificielle-et-emi et https://www.cnil.fr/fr/comment-permettre-lhomme-de-garder-la-main-rapport-sur-les-enjeux-ethiques-des-algorithmes-et-de
Objectifs info documentaires	● Évaluer les **limites et les biais** et les **enjeux éthiques** de ces technologies ● connaître et **comprendre le modèle technologique** et les outils (culture numérique)
Compétences info-documentaires	● Savoir identifier les biais et les limites de l'IAG
Ressources	
Matériel	● Tableau blanc / Feutres veleda, si besoin ● ordi et vidéo-proj
Ressource pédagogique	- wooclap + tél portable - Le chat mistral - chatGPT - Perplexity
Déroulement - Contenus - Activités	
Titre de l'activité	**A la recherche des biais de l'IA**
Descriptif synthétique de l'activité	Comprendre les clichés véhiculés par l'IA, les limites et les biais de l'IA.

Déroulement de l'activité **Séance 1 (1h)**	1. Installation, Appel (5 min.) 2. Images générées par IAG véhiculent des clichés. Couleur Rose/bleu. Homme mécanicien avec cambouis, femme mécanicienne sexy sans cambouis… Selon vous d'où proviennent ces biais ? ----> vers comment fonctionne l'IA ? : débat (5 min.) **Plusieurs possibilités selon les résultats du débat** (15 min.) : 1. Simulation d'un **entraînement sur les bases de données** avec Wooclap https://app.wooclap.com/QIGYDN?from=instruction-slide Date de début de la 2nde GM. dire que c'est anonyme pour que chacun mette ce qu'il veut. Entraînement de l'IA et fiabilité des bases de données des IAG 2. Durant la 2nde Guerre Mondiale, on peut dire qu'il y avait 2 camps, les alliés et les puissances de l'axe. Citez le 1er pays auquel vous pensez et qui faisait partie des pays alliés. Questions aux élèves Test sur 2 IAG - Biais des données initiales + stéréotype, discrimination et exclusion - débat https://chatgpt.com/ https://www.perplexity.ai/ 3. Création d'un abécédaire des animaux, montrer seulement https://chatgpt.com/share/698dfb05-23c0-8008-8f01-42d8bee2839c bases de données anglosaxonnes, peu de fiabilité, faire preuve de discernement, mélange des lettres etc... Débat - connaissances, esprit critique 4. Représenter une lycéenne de 17 ans, montrer seulement https://chat.mistral.ai/chat/9fefddc0-12ad-4784-a0d8-f4a3199ff2c6 Sexisme et représentation des femmes jolies et répondant à une norme de beauté. Cliché lieux. https://chat.mistral.ai/chat + Chatgpt 3. Jeu, les biais de l'IA (bases d'entraînement + modèle d'entraînement + réponses statistiques) https://mi.hepl.ch/projects/ia/ia.html (25 min.) 4. Test comparIA avec les question 4 : https://comparia.beta.gouv.fr/ (5 min.) Conclusion

Le 16/02 :

- photo **lycéenne** ---> beauté, peau claire, forme de bouche, yeux, norme et moyenne des critères d'un beau visage -----> **codage** machine IA
- **abécédaire** ----> bases de données **pas d'analyse**, juste des données brutes même si justes elles produisent des aberrations
- **wooclap** guerre mondiale ---> certaines **données fausses** (données dispos)
- discussions avec l'IA ----> **entraînement** qui peut changer l'information donnée par l'IA mais biais encore par le filtre de celui qui entraîne (**CAPCHA** - étiquetage volontaire mais il existe aussi l'étiquetage manuel dans les pays pauvres)
- A retenir, quelques biais :
    - collecte
        - biais de sélection et d'omission
        - biais de représentativité et de **disponibilité**
        - biais d'étiquetage
    - entrainement
        - biais de confirmation et d'ancrage

- biais du survivant, (exemple avion blindage wald)
- effet dunning-kruger (manque de nuance)

Liens :
https://app.wooclap.com/QIGYDN?from=instruction-slide
https://www.perplexity.ai/
https://www.perplexity.ai/
https://chat.mistral.ai/chat
https://chatgpt.com/share/698dfb05-23c0-8008-8f01-42d8bee2839c (abécédaire)
https://chat.mistral.ai/chat/9fefddc0-12ad-4784-a0d8-f4a3199ff2c6 (lycéenne)
https://mi.hepl.ch/projects/ia/ia.html
https://comparia.beta.gouv.fr/

Article Sciences et avenir :

La beauté universelle existe-t-elle ?

Lise Loumé le 12.04.2016 à 18h43, mis à jour le 27.08.2021 à 12h50
Existe-t-il des critères universels permettant de juger si un visage est attirant ? Des chercheurs français ont répondu à cette question grâce à un modèle mathématique.
Pour évaluer la beauté du visage d'autrui, utilisons-nous des critères communs, permettant la définition d'une beauté universelle ? En se basant sur un modèle mathématique reproduisant l'activité de nos neurones, Julien Renoult, biologique au CNRS et spécialiste de l'évolution, avec deux autres collègues, a répondu à cette question dans la revue *The Royal Society Open Science*.

Des visages arrondis et aux traits réguliers plus attractifs

Les chercheurs se sont appuyés sur la théorie du "codage efficace". *"Cette théorie, qui date de 1996, a été appliquée initialement dans le domaine de l'intelligence artificielle. Elle définit un algorithme qui doit coder un signal en activant le moins d'encodeurs (des neurones artificiels) possible en même temps"*, explique à *Sciences et Avenir* Julien Renoult*. "Puis l'on a découvert que cette stratégie de codage est très présente dans le cerveau d'animaux, vraisemblablement parce qu'elle est économe en énergie du fait qu'il faille activer peu de neurones à la fois. Notre cortex visuel primaire l'utilise aussi pour traiter l'information provenant des yeux."* Ainsi, ce que l'on trouve beau correspond à une information facile et rapide à coder pour notre cerveau. Des études précédentes ont d'ailleurs révélé que le traitement rapide de l'information visuelle (que l'on perçoive des visages, des paysages ou des formes géométriques abstraites) entraîne une sécrétion de dopamine, active le circuit de la récompense et génère donc du plaisir.

Pour réaliser leur expérience, les chercheurs ont demandé à une centaine d'hommes français (âge moyen : 36 ans) de noter des photos de femmes selon leurs préférences. *"Nous avons modélisé le cortex visuel primaire afin d'estimer à partir d'une photo le nombre de neurones qu'il fallait activer pour traiter l'information visuelle qui arrivait",* commente Julien Renoult. Plus les visages étaient bien notés (donc attractifs), moins l'image activait de neurones dans cette région du cerveau. *"Ce sont des visages arrondis à la peau homogène, aux contours réguliers, avec des courbures fines, qui sont jugés les plus attractifs"*, conclut le chercheur de son étude. On peut penser que ces visages sont préférés car ils révèlent une personne jeune et en bonne santé. "*Mais dans notre étude, nous avons contrôlé statistiquement l'effet de l'âge"*, précise Julien Renoult. De plus, les femmes prises en photo, toutes françaises et de type caucasienne, ont entre 18 et 26 ans...

L'amour de la symétrie

L'étude ne porte que sur des hommes... Si elle avait inclus des femmes, le résultat aurait-il été tout autre ? *"Non, il y a de fortes chances pour que nous obtenions la même conclusion*, répond Julien Renoult. *Le*

cortex visuel des femmes utilise également la même stratégie de "codage efficace" et le traitement rapide de l'information génère du plaisir également chez les femmes." Le résultat de l'étude ne suprend pas le chercheur. *"Des travaux de psychologie évolutive ont montré que les individus préfèrent en général les visages très symétriques, car ce critère indique un bon développement in utero, ce qui est corrélé à une bonne santé et une bonne résistance immunitaire",* précise-t-il. Mais n'est-ce pas aussi lié au fait qu'un stimulus symétrique est codé plus rapidement qu'un stimulus asymétrique, et donc que l'attractivité est là encore liée au codage efficace ?

D'autres facteurs jouent dans l'attractivité

Julien Renoult reconnaît que ses travaux présentent quelques limites. *"Il faudrait reproduire l'étude en demandant aux personnes de noter l'attractivité des visages, tout en mesurant leur activité cérébrale par exemple avec une IRM fonctionnel. Mais une telle étude est très coûteuse !". "De plus, pour tester plus généralement la théorie esthétique du codage efficace, il faudrait regarder si les visages arrondis aux traits fins reçoivent des notes d'attractivité plus rapidement que les autres, et s'ils sont maintenus en mémoire plus longtemps"*, ajoute-t-il. Sans oublier que le "codage efficace" ne fait pas tout dans l'attractivité d'un visage. *"Il n'explique qu'au mieux 17% de la variation dans les notes d'attractivité",* précise le chercheur. Cela veut dire qu'il y a beaucoup d'autres facteurs (biologiques, culturels, etc.), qui influencent nos préférences. Une étude récente a d'ailleurs montré que les goûts en matière d'attirance physique diffèrent même d'un vrai jumeau à un autre et dépendent davantage des expériences personnelles vécues que de facteurs génétiques.

Une étude en ligne pour mieux comprendre l'origine du beau

Le CNRS a lancé récemment une étude en ligne pour mieux comprendre les facteurs influençant l'évolution culturelle et biologique de la beauté et de son évaluation. *"La réussite de l'étude dépendra du nombre de participants et de leur origine géographique. Le site de l'étude est traduit en 18 langues"*, précise Julien Renoult. N'hésitez donc pas à y participer !