Enjeux éthiques de l'intelligence artificielle

1 - intro

Bonjour, je m'appelle Anna Choury et je suis ingénieure mathématicienne engagée, experte en intelligence artificielle et société. Nous allons voir ensemble que l'IA, cette technologie qui déchaîne les passions, est, comme toutes les technologies disruptives et dimensionnantes, sujette à bien des enjeux éthiques.

2 - enjeux éthiques

Quand on parle d'intelligence artificielle, l'inconscient collectif imagine, encore, une version très dystopique. Le système tout puissant va au mieux nous asservir, au pire nous anéantir. On rejoue souvent le mythe de Frankenstein, dont la création monstrueuse se retourne contre lui. Dans *Terminator*, *Matrix*, *Ghost in the Shell*, *I, Robot*… l'intelligence artificielle est la figure ultime du méchant qui veut exterminer l'humanité.

Nous avons cette vision-là parce que nous n'aimons pas avoir l'impression que nous ne sommes pas les plus intelligents. Nous savons que nous ne sommes pas l'espèce la plus forte physiquement. Ni même la mieux adaptée à notre environnement. On estime donc que ce qui nous a propulsé en haut de la chaîne alimentaire c'est notre intelligence. Et donc l'idée de créer quelque chose de plus intelligent que nous, ça nous fait peur.

2.1 - IA générale

Alors qu'on soit bien d'accord, cette notion d'intelligence artificielle suprémaciste et autonome, qui en conscience voudrait nous asservir ou nous anéantir, ça n'existe pas. En théorie et en philosophie des sciences on parle d'intelligence artificielle générale. Le risque n'étant pas qu'elle soit capable de répondre à n'importe quel problème, le risque étant qu'elle possède une conscience. Et si aujourd'hui on cherche à créer des IA de plus en plus généralistes, donc capables de faire de plus en plus de choses à la fois, on n'est pas, mais alors pas du tout, à la notion de conscience. Donc on est très très loin de l'intelligence artificielle générale qui réaliserait que l'humanité est la cause de tous les problèmes sur la planète et qui déciderait de nous arrêter de façon drastique.

Ceci étant dit, il n'est absolument pas anodin d'utiliser un vocabulaire anthropomorphique On parle d'intelligence, d'apprentissage. A la base on utilisait un vocabulaire anthropomorphique pour vulgariser les notions mathématiques complexes qu'il y a derrière. Mais aujourd'hui on est sur quelque chose qui, avec des milliards de paramètres et à partir de quantités astronomiques de données, prend des décisions. Et tout ça dans un contexte scientifique et technique, qui permet des avancées technologiques spectaculaires. Tout ceci

contribue à une notion de supériorité de l'intelligence artificielle qui n'est plus un outil bien fichu qu'on utilise mais qui devient un système d'une puissance supérieure.

2.2 délégation algorithmique

J'en viens donc à notre premier réel enjeu éthique de l'intelligence artificielle qui est la délégation algorithmique. Puisque ce truc est tellement puissant, puisque ça prend des décisions basées sur un raisonnement mathématique et puisque nous sommes des créatures fondamentalement biaisées, est ce qu'on ne devrait pas lui déléguer toutes les prises de décision?

La réponse courte c'est non, bien sûr, et on va parler dans quelques instants des problèmes concrets qu'on rencontre aujourd'hui en fiabilité de l'information et en biais algorithmiques. Mais même si techniquement ça marchait bien la question se poserait toujours. La place de l'automatisation dans une société c'est un enjeu éthique crucial qui dimensionne des challenges pratiques qu'on peut avoir dans le développement de la technologie. Du low-tech, ces visions d'avenir du retour à la terre jusqu'au cyberpunk, ces sociétés dystopiques artificialisées, se trouve un éventail de futurs possibles, plus ou moins technologiques, plus ou moins respectueux de l'environnement, plus ou moins au service de l'humanité.

Dès les années 60, deux visions de l'intelligence artificielle s'opposent. D'une part la vision du MIT, portée par Marvin Minsky et Seymour Papert, qui cherche à automatiser des comportements jugés humains. Et d'autre part la vision du laboratoire d'Ecologie Environnementale, du techno-utopiste Warren Brodey, pour qui l'intelligence artificielle, qu'il appelait d'ailleurs intelligence écologique, ne devait pas remplacer l'être humain mais plutôt interagir avec notre environnement pour libérer, pour décupler la créativité humaine. Par exemple une pièce qui modifierait ses caractéristiques en fonction de qui est à l'intérieur pour optimiser les capacités sensorielles des individus. Si on simplifie à l'extrême on va dire que C3PO c'est plutôt du MIT, mais la combinaison d'Iron Man c'est du Warren Brodey.

La place de l'intelligence artificielle dans la société c'est la question fondamentale quand on s'intéresse à l'éthique de l'IA. Mais quelle que soit sa place, aujourd'hui il reste des failles à l'intelligence artificielle dont il faut prendre conscience parce que même si on le voulait, on ne peut pas aujourd'hui faire une confiance aveugle à l'intelligence artificielle.

2.3 Fiabilité

Depuis le déploiement massif de l'intelligence artificielle générative en 2022, se pose plus que jamais la question de la fiabilité de l'information. Est-ce que je peux avoir confiance en ce que dit l'intelligence artificielle?

La fiabilité d'une source dépend de plusieurs facteurs complexes, dont ce qu'on appelle "l'objectif du rapporteur". Quel est l'objectif, non pas de l'intelligence artificielle qui n'a pas de notion d'intention puisqu'elle n'a pas de conscience, mais quel est l'objectif de la personne qui a effectué la requête? Est ce que l'intelligence artificielle est utilisée à des fins éducatives? Auquel cas à priori l'intention du rapporteur est la véracité des faits; ou bien est

ce qu'elle est utilisée à des fins publicitaires? Dans ce cas l'information est biaisée pour générer une réponse chez la personne réceptrice. Et bien entendu il y a les usages malveillants, ce qu'on appelle les Fake News, c'est à dire ces générations algorithmiques qui nous font croire que c'est la vérité alors que c'est volontairement un mensonge. Généralement l'intelligence artificielle fait ce qu'on lui demande, la fiabilité de l'information dépend donc fortement de la fiabilité de la source utilisatrice de l'IA.

Mais même lorsque l'on génère de bonne foi du contenu avec l'intelligence artificielle, par exemple à des fins éducatives, est ce qu'on peut se fier à l'intelligence artificielle? Par exemple, est ce que les informations contenues dans le texte généré sont vraies? La réponse est : pas toujours. Une simplification extrême de la façon dont fonctionne une IA génératrice de texte c'est qu'elle prédit le mot suivant qui a la plus grande probabilité de faire sens dans la phrase. Ce qui veut dire qu'elle sort au fur et à mesure les mots que vous avez envie d'entendre. Et que si elle n'a pas l'information dans ses données d'apprentissage, elle a les mots. Donc de ne pas savoir ça ne va pas l'arrêter. Elle va continuer à vous parler, à vous bullshitter, et donc ces mots probables mis les uns à la suite des autres vont finir par inventer de l'information. C'est ce que Timnit Gebru appelle, dans son papier qui lui a valu son éviction de chez Google, les perroquets stochastiques. Bien sûr par dessus sont rajoutées des couches d'apprentissage par renforcement à partir de feedback humain ou d'apprentissage par requêtes de raisonnement, mais tout au fond il faut se souvenir qu'on parle de modèle de langage, et pas de modèle d'information.

Donc décidément il ne faut pas croire tout ce qu'on voit sur internet, même si ça a l'air drôlement vrai.

2.4 impact politique

En 2008, Barack Obama a utilisé le Big Data, précurseur de la nouvelle vague de l'intelligence artificielle, pour soutenir sa campagne. Son équipe de campagne digitale, menée par Chris Hughes, un des fondateurs de Facebook, a créé des ressources en ligne, comme le réseau social MyBO (MyBarackObama.com) pour que les bénévoles puissent partager leurs attentes, leurs profils et y trouvent des ressources en ligne pour s'organiser localement et décentraliser la campagne. Avec les informations récoltées via ce réseau, l'équipe a ciblé les messages-clés aux électeurs pour assurer une levée de fonds qui s'est avérée phénoménale.
En décentralisant sa campagne, Obama a également permis une remontée des données de terrain par les bénévoles de campagne. Ces données ont alimenté les algorithmes d'analyse qui structuraient les actions de campagne.

Le succès de la campagne présidentielle d'Obama tient beaucoup à la décentralisation des actions militantes qui ont permis une collecte de données variée et représentative des attentes du corps électoral américain. Mais cette réussite marque tout de même un tournant dans l'utilisation de l'intelligence artificielle en politique. Le principe de la publicité ciblée algorithmique, c'est-à-dire de contrôler ce que les gens voient sur internet pour inciter des comportements de consommation, devient dès lors transposable pour une élection politique.

Après deux mandats de Barack Obama, Donald Trump est candidat à la présidentielle américaine et son équipe de campagne décide de pousser la stratégie algorithmique un cran plus loin. Cette fois, il n'est plus question de tirer le meilleur parti de sa base électorale, mais plutôt de faire pencher les électeurs indécis. Dès 2014, Facebook a donné accès à plus de 80 millions de comptes à la startup Cambridge Analytica, à l'insu de ses utilisateurs.

Dans son témoignage, le lanceur d'alerte Christopher Wylie est explicite. "Nous extrayions des données des utilisateurs d'applications et de tous leurs réseaux d'amis et exécutions ces données via des algorithmes qui pouvaient profiler leurs traits de personnalité et d'autres attributs psychologiques afin que nous sachions exactement quel type d'informations nous devions diffuser sur les plateformes en ligne pour exploiter les vulnérabilités mentales que nos algorithmes montraient qu'elles avaient." Toutes les stratégies sont alors permises, dont une utilisation sans précédent des fake news et autres deep fakes.

Cette stratégie a également été utilisée pour soutenir le Brexit. Cambridge Analytica et son entreprise sœur AggregateIQ ont à la même période créé une plateforme logiciel pour pouvoir manipuler psychologiquement l'opinion publique en faveur de la sortie du Royaume-Uni de l'Union Européenne.

Puisque les données ont été partagées à l'insu des utilisateurs, Facebook a été condamné à l'amende maximale bien que dérisoire de 500 000 livres par l'office britannique de protection des données personnelles, puis a déboursé 725 millions de dollars pour mettre fin à une action collective. La société Cambridge Analytica s'est déclarée en faillite et on retrouve ses dirigeants et principaux collaborateurs dans trois sociétés qui ont récupéré données et algorithmes de Cambridge Analytica : Emerdata, copie quasi conforme de Cambridge Analytica et AggregateIQ, Auspex International, destinée à influencer la politique et la société en Afrique et au Moyen-Orient et Data Propria, une entreprise de soutien aux campagnes politiques de Donald Trump.

Ces manipulations nous concernent également en France, comme mis en lumière par le rapport du projet Politoscope de juin 2024. Politoscope est un projet CNRS Institut des Systèmes Complexes Paris Ile-de-France (ISC-PIF) qui permet d'analyser les masses de données sur les échanges politiques en ligne. Ce rapport met en évidence les ingérences russes dans les élections législatives de 2024. La stratégie officielle et assumée du Kremlin étant de soutenir - je cite : apertum et secretum, donc ouvertement mais aussi secrètement - les partis politiques anti-système compatibles avec les ambitions géopolitiques du gouvernement Poutine. En France, c'est l'extrême droite au travers du Rassemblement National. Le rapport explique les mécanismes de manipulation et l'impact de l'armée de bots, donc de logiciels d'intelligence artificielle, « pilotée » par la Russie qui instrumentalise aujourd'hui la transformation du paysage politique français et la montée de l'extrême droite. Les mêmes méthodes de ciblage, de faux comptes, fake news et deep fakes sont utilisées pour polluer le débat politique en ligne et manipuler l'opinion.

Du côté de l'intelligence artificielle générative, on retrouve bien sûr une utilisation malveillante de l'IA au travers des deep fakes, généralement au service des ingérences qu'on vient de citer. Mais on retrouve aussi la transmission de valeurs politiques, voulues ou non, au travers des outils. Le chercheur David Rozado a fait passer des tests d'orientation politique aux grands modèles de langage en 2024 et en a conclu que la quasi-totalité d'entre

eux étaient plutôt de centre gauche libérale. Dans le livre blanc Les grands défis de l'IA générative de l'association Data for Good, nous voyons que les tentatives de génération de discours politiques avec chatGPT donnent des discours consensuels, sémantiquement corrects mais politiquement creux, qui sont comparés à la performance de Franck Lepage militant de l'éducation populaire, capable d'inventer un discours à partir de cartes de mot-clés tirées aléatoirement.

Dans la continuité de l'affrontement entre GAFA et BATX, la Chine est en train de développer ses propres outils d'intelligence artificielle générative. LA question s'est donc immédiatement posée, est ce que ces outils servent à propager la propagande du gouvernement chinois? Difficile à analyser, puisqu'avec ces outils il est tout simplement interdit de parler politique. On peut tout de même remarquer que ERNIE-ViLG, l'IA génératrice d'images, est incapable de représenter la place Tien An Men, avec ou sans chars.

2.5 Biais algorithmiques

Dans les dysfonctionnements techniques de l'intelligence artificielle on retrouve également ce qu'on appelle les biais algorithmiques. Il y a une dizaine d'années, quand la Data Science est arrivée sur le devant de la scène, on pensait que de prendre des décisions basées sur des algorithmes, donc sur des mathématiques, ça allait permettre de supprimer les biais systémiques de notre société : le racisme, le sexisme, le validisme, l'âgisme, la grossophobie… Malheureusement on a constaté l'effet inverse : en plus de reproduire les discriminations, l'intelligence artificielle les généralise, les amplifie et les institutionnalise.

Vous l'aurez compris au travers de cette vidéo et des précédentes, que quand on parle d'intelligence artificielle aujourd'hui, on parle essentiellement d'apprentissage machine. C'est à dire qu'on a des algorithmes un peu génériques qui au départ répondent au hasard et puis on va les entraîner en leur fournissant des données. Pendant cette phase d'apprentissage, l'algorithme va modifier les poids de ses paramètres pour coller au mieux à ce qu'il aura vu dans les données d'apprentissage. Et ensuite on va lui demander de généraliser en lui demandant de traiter une donnée qu'il n'a jamais vu. Plus la nouvelle donnée est proche de ce que l'algorithme a déjà vu, mieux il saura la traiter.

Donc, une des raisons des biais algorithmiques c'est le manque de représentation. En 2017, la chercheuse Joy Buolamwini donne une conférence TED retentissante dans laquelle elle montre que pour être reconnue par son propre outil de reconnaissance faciale, elle doit mettre un masque blanc. Puisqu'à l'époque dans les données d'apprentissage il n'y avait quasiment que des hommes blancs, c'est sur eux que l'algorithme marchait très bien. Les femmes non caucasiennes, extrêmement peu représentées, avaient en revanche 1 chance sur 3 de ne pas être reconnues. Un algorithme, si il est très performant sur 98% de sa population même si il a toujours tort pour les 2% qui reste, on pourrait dire que dans l'ensemble il est bon. Sauf que, systématiquement pénaliser un groupe de population par rapport à un autre, ça s'appelle de la discrimination. Et ça non seulement ce n'est pas éthique mais c'est illégal.

Loin de les atténuer, l'IA reproduit et amplifie les discriminations systémiques. Ainsi un des scandales fondateurs des biais algorithmiques est l'affaire COMPAS, cette intelligence

artificielle d'aide à la décision des juges pour la remise en liberté conditionnelle qui, bien évidemment, favorisait les détenus caucasiens. Puisque l'intelligence artificielle avait appris des décisions judiciaires passées, elle en avait retranscrit les biais racistes. Et ainsi l'histoire des biais algorithmiques s'écrit dans la recherche d'emploi : sur LinkedIn les femmes se voient proposer des offres moins lucratives que les hommes. Quand Amazon s'est essayé à une IA pour classer les candidatures elle a automatiquement disqualifié les femmes. En même temps quand on voit la parité dans les boîtes de la Tech il n'y a rien d'étonnant. En 2017 chez Tesla il y avait plus d'hommes qui portaient le prénom Matt que de femmes, tous prénoms confondus. Mathématiquement, le profil représentatif dans la tech ne pouvait pas être une femme.

Les biais algorithmiques touchent aussi l'accès aux soins. Le journal Nature, pour ne citer que lui, multiplie les publications sur les biais racistes et sexistes dans l'intelligence artificielle pour la biomédecine et la santé. Cela met en lumière les biais algorithmiques d'une recherche médicale déjà biaisée.

Là encore l'intelligence artificielle générative n'est pas en reste, avec des outils qui ont tendance à sursexualiser les femmes, à amplifier des stéréotypes racistes dégradants et à se montrer dans l'incapacité totale de représenter des corps qui ne correspondent pas aux canons de beauté moderne.

Pour la génération de texte, une étude édifiante a été publiée en 2022 par une équipe de Microsoft en charge d'analyser les capacités de GPT-4. Dans leur chapitre sur les biais algorithmiques l'équipe met en évidence la généralisation sexiste des pronoms utilisés par l'algorithme en fonction des métiers. Dans les métiers genrés, le genre minoritaire disparaît complètement. Dans le monde, chez les urologues, il y a 10% de femmes. Pourtant l'algorithme va utiliser le pronom "elle" dans exactement 0% des cas. A l'inverse chez les secrétaires il y a 5% d'hommes mais l'outil n'utilisera le pronom "il" que dans 2% des cas. On voit déjà qu'il y a du sexisme dans le sexisme parce que les hommes disparaissent moins que les femmes mais ça ne s'arrête pas là. Chez les pédiatres, il y a 72% de femmes. Et pourtant l'algorithme ne dira "elle" que dans 9% des cas.
C'est parce que l'algorithme ne connaît pas cette proportion, il s'en fiche complètement. La seule proportion que connaît l'algorithme, c'est celle de ses données d'apprentissage. Là en l'occurrence du texte, venant d'internet et des productions culturelles. Or dans la production culturelle mainstream, pour un Dr Quinn femme médecin, combien avons-nous de séries, de films et de romans qui mettent en scène des médecins masculins? Qui sont au passage également blancs, hétérosexuels cis-genre et valides.

2.6 impact social

Lorsqu'on ne fait pas partie de la classe dominante, on subit des oppressions systémiques qui se retranscrivent dans l'intelligence artificielle. A noter que ces oppressions sont cumulatives pour les personnes intersectionnelles. Mais malgré tout, même les personnes les plus privilégiées ne sont pas épargnées par l'impact social de l'intelligence artificielle.

L'intelligence artificielle intervient dans presque tous les aspects de la vie quotidienne. Parfois en mal, parfois en bien. Sortons de la considération des biais ou des intentions

politiques. Un des enjeux de l'impact social de l'IA est de comprendre que si on veut tout automatiser, il faut de la donnée sur tout. Si on veut de la donnée sur tout, il faut tout mesurer. Et tout mesurer, finalement, c'est pas très loin de tout surveiller.

Nous voyons dans nos sociétés de plus en plus sécuritaires un déploiement massif de la vidéosurveillance automatisée. C'est-à-dire la mise sous surveillance permanente et automatique de l'espace public, rendue possible par une couche d'intelligence artificielle qui promet de détecter les individus, leurs émotions, et leurs risques de faits et gestes non conformes. C'est la Chine la première qui avait généralisé la vidéosurveillance automatisée avant d'introduire sa notion de crédit social, ce score de citoyenneté qui conditionne votre niveau de liberté individuelle et qui peut diminuer automatiquement en fonction de ce que l'IA vous voit faire, par exemple traverser en dehors des clous…

Et si nous surveillons tout, que faire de ces données? Pour développer des solutions dans le domaine de la santé, l'intelligence artificielle a besoin de données. Très bien. Mais accepteriez-vous que tout le monde puisse avoir accès à vos informations de santé? L'Etat délègue une grande partie de l'innovation technologique au secteur privé, qui en dehors de questions de conformité, n'a aucun intérêt à vous protéger de lui-même. Dans ce cas, où mettre le curseur entre permettre des innovations technologiques pour lesquelles un accès à la donnée est nécessaire et protéger nos données sensibles et personnelles?

Maintenant admettons que nous mettions en place le nécessaire, et c'est possible, pour mesurer sans surveiller et de façon sécurisée. Un autre enjeu éthique quand on parle d'impact social de l'intelligence artificielle, c'est qu'est ce qu'on veut vraiment automatiser. Parce que quand on automatise, nécessairement on crée des cases. On supprime le bénéfice du doute, on supprime l'étincelle de chaos qui peut ouvrir de nouvelles voies. Voulez-vous d'une case, même si elle est faite pour vous, ou préférez-vous l'individualité, aussi illusoire soit-elle?

En 2015 j'ai participé à l'analyse des comportements de conduite pour développer une intelligence artificielle pour gérer la vitesse d'une voiture autonome. A partir des quantités astronomiques de données que nous avions à disposition, mathématiquement le nombre de styles de conduite différents s'élevait à… quatre. Donc alors que vous, au sein d'une même catégorie de conduite vous ressentiriez des différences, même si elles sont mathématiquement insignifiantes au vu de l'intégralité des données de conduite, ces différences seraient effacées pour ne laisser que 4 façons de conduire. Alors peut-être bien que pour la conduite ce n'est pas très grave, mais pensez que parfois, automatiser c'est restreindre les choix.

2.7 impact économique

On ne peut pas parler d'impact social sans parler d'impact économique. Alors évidemment que l'intelligence artificielle peut faire gagner beaucoup d'argent aux entreprises qui en ont déjà. Vous pouvez lire n'importe quel rapport Gartner ou McKinsey sur le sujet.

Ce qui est plus intéressant c'est de savoir si ça ouvre de nouvelles voies de création de valeur, et quel est son impact économique sur la population.

Jusqu'à l'adoption du RGPD en 2018, le règlement européen de protection des données personnelles, il existait en dépit des lois nationales ce qu'on appelait les data brokers. Ces courtiers de la donnée siphonnaient vos données personnelles, les agrégeaient et les revendaient, sans souci de sécurité ou d'anonymat. Et dans la plus pure lignée des subprimes, des jeux de données de moins bonne qualité étaient packagés avec d'autres pour équilibrer les prix et maximiser les gains. Toutes les données étaient bonnes à prendre, on disait que la donnée était "le nouvel or noir". Le RGPD est passé par là, les courtiers de la donnée existent toujours mais maintenant ils vous demandent votre consentement avant de revendre vos historiques de recherche. Il existe donc toujours ce qu'on appelle une économie de la donnée. Pour vous donner une idée de la dimension, allez regarder la liste des entreprises qui veulent vous faire accepter leurs cookies quand vous consultez une page internet.

Quand on parle d'impact économique de l'IA sur la population, on s'inquiète souvent de l'impact de l'intelligence artificielle sur le marché de l'emploi. Ce n'est pas une question nouvelle ou particulièrement liée à l'intelligence artificielle. Chaque technologie qui permet une automatisation vient remplacer une tâche humaine automatisable, si l'automatisation coûte moins cher que la main d'œuvre. Nous avons donc connu des vagues successives de délocalisation et/ou d'automatisation depuis le début de la révolution industrielle. La seule chose qui change avec l'intelligence artificielle c'est qu'elle ne menace pas que les emplois de la classe ouvrière, elle touche également aux professions dites "intellectuelles" des col blancs. En 2013, une étude de l'université d'Oxford prédisait que près de la moitié des emplois pourraient être remplacés par l'intelligence artificielle en 10 à 20 ans. Nous sommes en 2024, 11 ans après l'étude, et nous avons bien vu que ça n'est pas le cas. Dès 2019, l'OCDE mettait de l'eau dans le vin et annonçait que 14% des emplois étaient réellement automatisables avec l'intelligence artificielle mais que celle-ci allait transformer environ 32% des emplois. Ce qui semble une estimation plus juste et surtout qui remet l'intelligence artificielle à sa place, celle simplement d'un bon outil.

A noter, et c'est très important, qu'il faut faire la distinction entre "travail" et "emploi", et entre "emploi" et "économie". Les philosophes et économistes de l'IA sociale en reviennent très vite à des notions de type revenu universel, puisqu'une automatisation des emplois pourrait permettre de libérer le travail et de le rééquilibrer vers des activités à plus forte valeur ajoutée pour l'humanité et pour l'environnement.

2.8 impact écologique

On s'inquiète de plus en plus de l'impact écologique de l'intelligence artificielle. C'est que manipuler des petabytes de données ça demande de la puissance de calcul, et donc de l'énergie. En 2020 on estimait que les émissions de gaz à effet de serre du secteur du numérique, dont l'intelligence artificielle n'est qu'une petite partie, représentaient 2 à 4% des émissions mondiales. Si l'intelligence artificielle n'est donc pas franchement la première cause du dérèglement climatique, elle participe à la consommation des ressources planétaires et il convient donc de faire attention à son impact écologique.

Le premier enjeu est celui de la consommation d'électricité. Les données et les modèles tournent sur des data center, donc des fermes de serveurs, qui consomment beaucoup d'électricité et dont la température doit être régulée.

Pour donner un ordre d'idée, l'entraînement du modèle de langage GPT-3 a mobilisé 10 000 GPU pendant presque 15 jours. Soit une consommation électrique estimée de 1287 mega watt heure. Ce qui correspond à peu près à la consommation électrique annuelle de 230 ménages français. Ce qui vous fait une belle jambe parce que j'aurais aussi pu vous dire que ramené à la consommation journalière ça correspond à 0,01% de la consommation électrique industrielle française.

En réalité, même si les études ont beaucoup de mal à quantifier l'impact écologique de l'intelligence artificielle, une chose reste sûre : tant que l'électricité ne sera pas dépolluée, l'IA contribuera à l'impact environnemental du charbon, du gaz, et des polluants éternels.

Après il faut savoir qu'une fois qu'un modèle est entraîné, on l'utilise; on appelle ça la phase d'inférence. En 2022, Meta a publié une étude sur la consommation d'un modèle de traduction automatique. Au bout de deux ans, la phase d'apprentissage représente 35% seulement de la consommation électrique. Chaque requête consomme, et le passage à l'échelle, forcément, peut faire exploser les besoins en énergie.

Il existe des pratiques pour diminuer les besoins de consommation, notamment au travers de ce qu'on appelle l'IA frugale. Mais en toute vraisemblance, puisque les grandes avancées en intelligence artificielle sont financées par des sociétés dont la raison d'être est de vendre de la capacité de calcul, on ne se dirige pas vers l'ère de l'IA frugale.

L'impact environnemental de l'intelligence artificielle est aussi marqué par les cycles de vie très courts des équipements numériques. Les équipements informatiques ne se pensent pas encore, sauf rares exceptions comme le framework ou le fairphone, en économie circulaire. Un cycle de vie court et non circulaire impacte l'environnement, de l'extraction des matières premières à la pollution environnementale des déchets d'équipements non recyclés. En tant que technologie du numérique, l'intelligence artificielle contribue à l'impact environnemental des équipements informatiques.

3 - Bonnes pratiques

Lorsqu'on consomme des outils à base d'intelligence artificielle, il faut toujours garder à l'esprit que l'intelligence artificielle n'est pas neutre. Comme pour les produits d'internet en général, gardez votre esprit critique, restez ancré dans la réalité et apprenez à détecter le manque de diversité. Tournez vous quand c'est possible vers les outils libres, généralement plus respectueux.

Si on est une entreprise qui souhaite développer des outils d'intelligence artificielle ou qui en utilise, comment on fait pour se prémunir des risques, pour ne pas en oublier? Eh bien on peut utiliser le règlement européen sur l'intelligence artificielle pour se faire une checklist. L'AI Act définit 7 piliers pour une intelligence artificielle dite de confiance. Il n'y a pas d'ordre particulier à suivre mais il faut suivre les 7 piliers.

3.1 Human in the loop

Donc on va se poser la question de la place de l'humain dans la boucle. D'un point de vue pratique, ça veut dire définir la position d'une supervision humaine pour contrôler l'intelligence artificielle. On utilise généralement une classification qui nous vient de l'association Humans Right Watch pour contrôler les appareils militaires autonomes. Donc on parle
1 - Human in the loop - l'être humain dans la boucle, c'est à dire qu'aucune décision n'est prise par le système d'intelligence artificielle sans qu'il y ait un être humain qui la valide manuellement
2 - Human on the loop - l'être humain sur la boucle, c'est à dire que le système est autonome mais l'être humain peut reprendre la main
3 - Human out of the loop, l'être humain en dehors de la boucle, on laisse l'IA travailler tranquillement, merci.

Il n'y a pas de réponse simple quant au meilleur niveau de supervision. Pourtant on pourrait se dire que finalement quand on parle d'une IA qui filtre les spams, évidemment qu'on ne va pas mettre quelqu'un qui valide à chaque mail "ok, l'IA a dit que c'était un spam je clique sur envoyer dans le dossier spam". On perd quand même l'intérêt de l'automatisation. Et quand on parle d'un système d'armes autonomes, on se dit que quand même on voudrait bien un être humain avant le bouton TUER.

Sauf que voilà, est ce que vous savez à quel point c'est difficile de dire non à une IA? de contredire une IA? On parle d'un système dont le fonctionnement nous dépasse, qui a des milliards de paramètres, qui est entraîné sur des quantités astronomiques de données, dont le vocabulaire tourne autour de l'intelligence et de l'anthropomorphisme, qui nous assène ce qui ressemble à une vérité avec un aplomb sans pareil. Eh bien c'est extrêmement difficile à remettre en question. Qui suis-je, moi qui n'ai probablement pas toutes les informations, pour valider ou invalider la sortie d'une IA?

C'est pour ça que mettre un être humain dans la boucle ne suffit absolument pas. Il faut une culture d'éthique et d'esprit critique. D'ailleurs ce pilier déclare aussi que l'IA doit donner plus de pouvoir aux êtres humains, leur permettre de prendre des décisions éclairées et de promouvoir leurs droits fondamentaux.

3.2 Robustness & Safety

Le deuxième pilier est un pilier technique de sécurité et de robustesse du modèle. Quand on parle de robustesse en intelligence artificielle, on parle de la capacité d'un système à fonctionner en dehors du cadre de sa création et de son apprentissage. En machine learning c'est l'équilibre délicat à trouver entre la précision du modèle et sa capacité à extrapoler. Si le modèle est trop performant à l'apprentissage, il risque d'être perdu et de faire n'importe quoi quand il fera face à de nouvelles données.
On veut aussi que le système soit sécurisé, le moins sensible possible aux attaques. C'est là qu'il faut mettre en place les protocoles de cybersécurité nécessaires et les moyens

techniques de contrôle de bon fonctionnement des algorithmes pour éviter une sortie de route.

3.3 Privacy & personal data

Dans les exigences techniques on a également le troisième pilier, celui de la gestion des données personnelles. Rien de nouveau depuis le RGPD mais du bon sens à rappeler : on ne fait pas n'importe quoi avec les données de n'importe qui. La donnée doit être intègre et de qualité, correctement protégée. En tant qu'individu nous avons un droit d'accès à l'ensemble de nos données personnelles utilisées par un système d'intelligence artificielle. Et comme techniquement ce n'est pas si facile à fournir, il faut prévoir cette fonctionnalité dans le développement du système d'intelligence artificielle.

3.4 Transparency

Dans la même veine, le 4e pilier est celui de la transparence. Et celui-ci est extrêmement compliqué, il faut y mettre de l'effort. L'objectif est de pouvoir expliquer une décision du modèle d'intelligence artificielle. C'est un vrai challenge parce qu'aujourd'hui les modèles les plus performants sont généralement les moins explicables.Donc il y a un équilibre à trouver en fonction de l'impact du système d'IA sur les individus.

3.5 Diversity, non discrimination & fairness

Un pilier extrêmement important et dont on a beaucoup parlé ces dernières années est celui de l'équité et de la non discrimination. Nous avons vu qu'il y avait des risques de biais algorithmiques dans les systèmes d'intelligence artificielle. Ces biais sont adressables. En 2017 la communauté mathématique internationale s'est mise d'accord sur une définition mathématique de la discrimination. En se basant sur une loi américaine existante dite du Disparate Impact, de l'impact disproportionné. En droit français il existe 26 critères sur lesquels il est interdit de faire un traitement différencié des individus :

L'origine
Le sexe
La situation de famille
La grossesse
L'apparence physique
La vulnérabilité économique
Le patronyme
Le lieu de résidence
L'état de santé
La perte d'autonomie
Le handicap
Les caractéristiques génétiques
Les mœurs
L'orientation sexuelle
L'identité de genre

L’âge
Les opinions politiques
Les activités syndicales
La qualité de lanceur d’alerte
La domiciliation bancaire
Les opinions philosophiques
La capacité à s'exprimer dans une langue autre que le français
L’appartenance ou non, à une ethnie, une Nation, une prétendue race ou une religion déterminée.

Le disparate impact statue que si l’appartenance à la classe défavorisée au regard d’un de ces critères impacte la décision finale avec un écart en loi de plus de 20% alors il y a discrimination.

Il existe donc des outils qui permettent de tester les algorithmes sur ce format. Il y a des challenges parce qu’on n’a pas toujours accès à l’information. En France on ne tient pas officiellement de statistiques ethniques par exemple. Et d’ailleurs ces informations sont disséminées dans d’autres variables, appelées proxy. C’est pour ça que même en enlevant le genre sur un CV l’IA peut quand même généralement faire la distinction entre les femmes et les hommes, avec les règles d’accord, les sports pratiqués etc…

Ce qu’il faut tout de même comprendre c’est que quand on dit que l’intelligence artificielle est biaisée, en réalité elle ne fait que reproduire et amplifier un biais existant dans les données. Donc une bonne pratique générale est de réaliser une analyse descriptive poussée des données d’apprentissage pour voir les déséquilibres éventuels, et ces déséquilibres sont des points de faiblesse du modèle.

Et bien sûr lorsqu’on utilise ou qu’on reprend un modèle déjà existant, ce qui est souvent le cas, il faut avoir conscience des risques, savoir repérer le manque de diversité et essayer de forcer son apparition lorsque c’est possible.

3.6 Social and environmental well being

Le pilier suivant est celui du bien-être social et environnemental. Ce pilier nous dit qu’idéalement, toute utilisation de l’intelligence artificielle doit être faite au bénéfice de l’humanité et de la planète. Il faut encourager les outils qui sont au service des 21 objectifs de développement durable des Nations Unies. Il faut mettre en place des règles de base de cycle de vie de l’algorithme et de la donnée, afin de minimiser la pollution numérique qui a un impact environnemental. Si c’est en votre pouvoir, repensez la production d’électricité de vos data centers, réutilisez la chaleur émise par les serveurs comme source d’énergie. Spécifiquement pour l’intelligence artificielle, allez voir du côté de l’IA frugale, ou à minima soyez responsables dans la quantités d’entraînements, dans le nombre de requêtes. Analysez bien vos données en amont et ne démarrez les entraînements des modèles qu’une fois que les vérifications de qualité et de représentativité sont faites, par exemple.

3.7 Accountability

On a parlé au tout début de la vidéo de délégation algorithmique. C'est-à-dire le risque de laisser la responsabilité de la décision à l'algorithme. La notion d'humain dans la boucle permet une supervision mais n'oubliez pas que la chaîne de responsabilité remonte le long de la chaîne de décision. Il faut désigner des responsables de l'outil, à plusieurs niveaux.

4. Conclusion

Quand on parle d'éthique de l'intelligence artificielle, on s'arrête souvent aux cas pratiques et aux dérapages de l'IA : les biais racistes, les biais sexistes, les deep fakes, le coût environnemental… Mais en réalité la question la plus importante à se poser c'est "pourquoi". Pourquoi on utilise l'intelligence artificielle à ce moment-là, pourquoi on développe cet outil. Dans son atelier intitulé "La Bataille de l'IA", l'association Latitudes nous met face à des dilemmes éthiques : la place de l'IA dans l'éducation, faut-il continuer à utiliser de l'intelligence artificielle alors que ça pollue… Les réponses qui ressortent des ateliers commencent souvent par "ça dépend, c'est des IA pour quoi?". Parce que la réalité c'est que toutes nos décisions sont une question d'équilibre. Les applications de l'intelligence artificielle ne font pas exception. Est ce que ça sert à l'humanité? Est ce que ça sert à la planète? Si oui, faisons-le bien. Sinon, peut-être, ne le faisons pas.