GLOSSAIRE EMI
Data, algorithmes, Intelligence artificielle

Avec la disponibilité d'une masse de données toujours plus grande et la généralisation d'algorithmes dotés d'intelligence artificielle, la production et la consommation des médias se modifie rapidement. Désormais, il s'agit de profiter des nouvelles opportunités qui se dessinent pour l'information et la liberté d'expression, tout en donnant à tous les citoyens les moyens de lutter contre une désinformation poussée par les algorithmes et l'IA.

NB: Ce glossaire critique définit les termes avant de leur donner l'interprétation en EMI (italiques).

• Algo-littératie
L'algo-littératie est l'une des composantes de l'EMI. Elle permet aux citoyens de mieux comprendre une information ou une actualité modelée par les algorithmes et l'intelligence artificielle. L'algo-littératie couvre les compétences suivantes : comprendre les algorithmes, dont ceux qui intègrent l'intelligence artificielle; être capable de les analyser et de les critiquer; savoir les utiliser à bon escient ; modifier ses comportements et usages en connaissance de cause; réagir à leur sujet sur les médias de masse et les médias sociaux.
Comme la plus grande partie des algorithmes intègrent désormais un ou plusieurs systèmes d'intelligence, on parle aussi d'IA-littératie. Vous pouvez consulter dans ce poster imprimable les 10 points clés de l'algo-littératie selon Savoir Devenir

• Algorithme
Un algorithme est une suite finie et non ambiguë d'opérations ou d'instructions. Les algorithmes permettent de résoudre une classe de problèmes ou d'exécuter une tâche. Par exemple, les algorithmes de recommandation permettent de déterminer ce que les gens aiment pour leur faire des propositions personnalisées.
En EMI, on s'intéresse aux algorithmes qui ont une influence à la fois sur la production et l'accès à l'information. Et on alerte sur les biais algorithmiques, dus à des bases de données de mauvaise qualité, qui peuvent entraîner des représentations du monde sexistes, racistes, etc.

• Algorithmes de prédiction
Ces modèles mathématiques sont dotés d'intelligences artificielles qui permettent d'apprendre du passé pour faire des prédictions sur le futur. Par exemple, en étudiant les performances passées de certaines équipes et de leurs joueurs contre d'autres équipes, certains algorithmes sont utilisés pour essayer de prévoir le résultat des matchs de football.
Si ces types d'algorithmes sont pour l'instant surtout utilisés dans les domaines de la santé, des sciences, de la détection des fraudes, de l'économie et des finances, des paris sportifs et de la vente, en EMI on surveille également leur utilisation en termes d'information.

• Algorithmes de ranking (ou de tri)

Utilisés notamment par les moteurs de recherche, ces algorithmes sélectionnent et listent les informations qu'ils jugent les plus pertinentes en réponse aux requêtes qu'on leur pose.

En EMI, on estime que ces algorithmes ne sont pas neutres et forgent en partie notre représentation du monde. On invite donc les citoyens à multiplier et diversifier les sources de leur information plutôt que de toujours tout demander… à Google par exemple.

• Algorithmes de recommandation

Les algorithmes de recommandation sont des modèles mathématiques qui permettent de personnaliser les résultats de recherche sur les moteurs ou les réseaux sociaux en proposant des contenus censés correspondre aux goûts des utilisateurs. Ils prennent en compte différents éléments tels que votre navigation passée, l'endroit où vous vous trouvez, les interactions que vous avez avec votre communauté, ou encore les publicités sur lesquelles vous cliquez. C'est le principe du "si vous avez aimé ceci, alors vous aimerez aussi…"

Les algorithmes de recommandation informationnels sont particulièrement scrutés dans la mesure où les comprendre et développer des stratégies pour les utiliser et non les subir est crucial, en particulier sur les médias sociaux.

• Biais algorithmique

Un biais algorithmique est un défaut dans le fonctionnement de l'algorithme, qui tend à traiter différemment des situations ou personnes. Il se produit lorsque les données utilisées pour entraîner le système d'apprentissage automatique reflètent des valeurs implicites.

En EMI, les biais algorithmiques les plus étudiés sont ceux qui conduisent les médias et les médias sociaux à renforcer les stéréotypes, comme les stéréotypes de genre, et favorisent la discrimination.

• Bulle de filtres

Le terme désigne le mécanisme de filtrage de l'information parvenant à un usager d'Internet sur la base de ses centres d'intérêt. Elle résulte des dispositifs de personnalisation des contenus en ligne basés notamment sur les algorithmes de recommandation, et sur la gestion des fils d'actualités. Elle se traduit par des propositions de contenus très similaires entre eux, dans lesquels l'usager retrouve ses goûts, ses opinions, ses biais.

L'impact des bulles de filtres sur l'information est actuellement en débat : certains pensent qu'elles ont pour conséquence de réduire la diversité des informations des usagers et de favoriser la désinformation. D'autres que l'on a exagéré la portée de ces bulles. Dans tous les cas, il est certain que les bulles de filtres ne favorisent pas l'esprit critique !

• Big data

Le terme de Big Data réfère à la collecte, au traitement et au stockage de données massives, précieuses pour alimenter les algorithmes. Ces techniques permettent d'analyser les propriétés statistiques de très grandes bases de données. Elles sont au cœur de notre économie numérique. Le big data est caractérisé par les 5 V : Volume, Vitesse, Variété, Véracité, Valeur.

En EMI, l'alerte est mise sur le fait que beaucoup de services dans le domaine des médias proposent des services "gratuits", en échange de ces données précieuses, sans que l'usager soit conscient de toutes les conséquences pour sa vie privée.

• Chatbot (agent conversationnel)
Un chatbot est un programme informatique développé spécifiquement pour simuler et traiter une conversation humaine (écrite ou parlée), permettant aux humains d'interagir avec des leurs machines comme s'ils communiquaient avec une personne réelle.
Doté d'une intelligence artificielle très performante, d'une base de données pour s'entraîner immense et d'une bonne interface, ChatGPT a popularisé l'usage de ce type d'outil.

• Cookies
Les cookies sont des petits fichiers texte déposés sur l'ordinateur ou le téléphone portable à l'insu de l'internaute, lors de la consultation de certains sites web, qui conservent des informations en vue d'une connexion ultérieure.
Les algorithmes de recommandation et de prédiction sont friands de ces cookies, qui aident à "profiler" les utilisateurs, c'est-à-dire les qualifier pour leurs proposer des offres ou informations qui ont plus de chance d'avoir du succès.

• Data
Une donnée ou data est une information numérique (image, son, texte, vidéo, signaux) produite par un individu, collectivité, institution. Elle peut être qualitative, quantitative ou technique. Par exemple, une date, un lieu, le nom d'une personne ou une photo sont des données.
En EMI, on distingue en termes juridiques les données personnelles et sensibles, qui sont protégées, de toutes les autres, qui ne le sont pas (comme les données collectées lorsque l'on navigue sur Internet).

• Data Journalisme (ou journalisme de données)
Le data journalisme date de la fin des années 50, où il visait à l'origine à exploiter des données statistiques pour aller dans le sens d'un "journalisme de précision " Dans son acceptation large, le terme couvre désormais toutes les pratiques journalistiques utilisant le traitement de données (fact-checking, traitement de l'information, data visualisation, robot-journalisme, études d'audience…).
Il convient de s'interroger sur les opportunités et les risques liés à ce nouveau mode de journalisme, et à ses répercussions sur le métier de journaliste.

• Datavisualisation
La datavisualisation - ou dataviz - est un ensemble de techniques qui permet de résumer de façon graphique des données. Elle permet de visualiser les informations importantes et les tendances au sein d'un ensemble de données. Dans les médias, la datavisualisation est appelée à la rescousse pour présenter de grandes quantités d'informations dans un format agréable, qui parle à tout le monde.
En EMI, on apprend à analyser ces visuels en évitant de les croire… juste parce qu'ils sont chiffrés et jolis.

• Deep learning (apprentissage profond)
L'apprentissage profond est un mode d'apprentissage automatique reposant sur l'utilisation de réseaux de neurones, avec un nombre extrêmement élevé de paramètres, qui nécessitent des bases de données très importantes.
Le deep learning est l'origine des deep fakes visuelles, ces images ou vidéos trompeuses souvent très difficiles à repérer, comme celles où une personnalité politique fait des déclarations absolument inventées.

● **Economie de l'attention**

Dans un univers où nous sommes saturés d'informations, suivant le principe de l'offre et de la demande, notre attention (et donc le temps que nous passons devant un document en ligne) devient une ressource rare, et donc précieuse, en particulier pour vendre de la publicité. Le contrôle de cette ressource est au cœur de la concurrence.

L'économie de l'attention favorise la diffusion des Infox et de toutes les informations sensationnalistes qui captent l'attention. Les algorithmes de ranking et de recommandation sont ainsi conçus pour nous apporter des contenus qui nous plaisent… et nous feront rester plus de temps sur les plateformes.

● **Boîte noire**

Un algorithme en boîte noire (par exemple certains à l'œuvre sur Tik Tok) est un algorithme pour lequel un utilisateur ne peut que fournir des données en entrée et regarder ce qu'il en résulte en sortie, sans savoir comment l'algorithme a procédé pour obtenir ce résultat.

En EMI, la question qui se pose est celle de la transparence. Si même les créateurs d'un algorithme ne connaissent plus son mode de fonctionnement parce qu'il a évolué "tout seul", comment peut-on exiger qu'ils nous expliquent comment il marche ?

● **Infobésité (ou surcharge informationnelle)**

Le terme renvoie à l'excès d'information auquel nous soumettent les technologies numériques et que nous avons du mal à traiter, car trop nombreuses. Dans le domaine de l'actualité, on parle aussi de fatigue informationnelle, un phénomène qui explique que de nombreuses personnes cessent de s'informer.

Désormais, l'une des compétences clés pour bien s'informer n'est plus de "trouver une information" mais de savoir trier les informations.

● **Intelligence artificielle (IA)**

A proprement parler, l'intelligence artificielle est un domaine de recherche qui étudie les mécanismes de l'intelligence humaine et cherche à les modéliser. Les différents systèmes d'IA tentent ainsi d'imiter le cerveau humain pour aider à réaliser des tâches ou se substituer à des activités humaines. Ces systèmes n'ont en fait rien d'intelligents et reposent le plus souvent sur des algorithmes dotés de machine learning.

En EMI, l'accent est mis sur la connaissance et la maîtrise les systèmes d'IA que les médias en ligne et les médias sociaux utilisent pour contrôler les flux d'information et influencer ce que les utilisateurs voient et sont invités à faire sur leurs plateformes.

● **Intelligence Artificielle Générative (IAG)**

L'IA générative fait référence à des algorithmes qui sont capables de générer de nouveaux contenus (texte, images, musique, code…) à partir des bases de big data sur lesquels ils ont été formés. Contrairement à la plupart des autres IA qui sont utilisées pour l'analyse ou l'aide à la décision, l'IA générative vise à créer des contenus de manière autonome, en imitant les créations humaines. L'IA générative la plus connue est celle utilisée par la société Open IA dans ChatGPT, qui permet de créer de toute pièce des textes, des images, des vidéos à partir des demandes formulées par les utilisateurs de ces services. Il y en a beaucoup d'autres !

En EMI, les enjeux sont de comprendre comment fonctionnent ces systèmes, de savoir les repérer et surtout de savoir bien les utiliser à fins d'information et de création.

• LLM (grands modèles de langage)
Les LLM sont des modèles de machine learning particuliers de type réseaux de neurones utilisés par les IA génératives. Leur objectif est d'apprendre la complexité du langage humain pour pouvoir répondre à tout type de demande des usagers, et d'être aussi capable de produire des textes en langage naturel. Les LLM fonctionnent de façon statistique en analysant des masses de données linguistiques, qu'ils "découpent "en séquences appelées tokens, pour repérer les suites qui sont les plus probables dans une langue ou contexte donné.
Les LLM pourraient bien modifier radicalement l'information mais aussi la création. Ou pas. Il est un peu tôt pour se prononcer.

• Média artificiel (ou synthétique)
Les médias artificiels tendent à utiliser des textes et/ou des images conçus entièrement par une IA développée à cette fin.
En mai 2023, le premier bulletin météo 100% conçu par IA et présenté par un avatar a été lancé sur une chaîne de télévision suisse. Les médias artificiels viennent s'ajouter aux médias de masse et aux médias sociaux.

• Métadonnées
Les métadonnées sont des données qui décrivent les données. Elles sont utilisées pour indexer, trier, analyser et faciliter la recherche d'information.
Par exemple, le nom de l'auteur d'un livre et sa date de publication ; le format, la définition et la durée de lecture d'une vidéo; le lieu et la date de la prise de vue ainsi que le type d'appareil photo et l'objectifs utilisés; le titre d'une page web, son type de codage, son auteur et les mots clés associés à sa publication…
En EMI, connaître la nature et le mode de collecte de ces métadonnées aide à comprendre comment sont organisées les informations qui nous sont présentées.

• Open data (données ouvertes)
Les données ouvertes sont des données numériques dont l'accès et l'utilisation sont laissés libres aux usagers. D'origine privée mais le plus souvent publique, elles peuvent être réutilisées par tout le monde, sans restriction technique, juridique ou financière. *L*'open data est beaucoup utilisé par les journalistes pour mener des enquêtes.
Les données ouvertes rappellent la diversité des systèmes de données et le besoin de contribuer aux "communs de l'information".

• Profil utilisateur
Un profil utilisateur est un ensemble de données et de métadonnées qui permet de créer des catégories de personnes ou de groupes. Les systèmes informatiques les traitent de façon différente, selon les besoins. Par exemple, un média proposera davantage de contenus "de droite" aux personnes qui ont été catégorisées comme telles. De même, un site marchand mettra en avant certains produits auprès de certains profils dont les caractéristiques laissent à penser qu'ils seront intéressés.
Il est important de savoir choisir le type de données que l'on souhaite voir associé à son profil, et à décider de façon avertie si l'on souhaite ou non "bénéficier" de flux d'informations personnalisés en fonction de son profil.

• **Prompt**

Un prompt est un court texte (une question, une requête, une instruction) qui indique à une IA ce que l'on attend d'elle. L'IA interprète le prompt puis génère un résultat sous la forme d'un texte, d'une image, d'un son ou d'une vidéo selon le type d'IA.

En EMI, l'une des nouvelles compétences essentielles est la capacité de rédiger des prompts efficaces, la qualité des prompts permettant d'améliorer considérablement la qualité de la réponse des IA.

• **Token / Tokenisation**

En informatique le terme token (ou jeton) peut avoir divers sens selon ses fonctions. En ce qui concerne l'analyse des textes, il représente l'unité lexicale sur laquelle se base le modèle. Ainsi, les LLM au cœur des IA génératives fonctionnent en découpant les textes en tokens. Ces tokens peuvent être, selon les modèles des mots, des ensembles de caractères, des combinaisons de mots et de ponctuation.

Non, les IAG ne nous comprennent pas. Elles analysent nos requêtes et produisent des textes en recombinant, sur une base statistique, les tokens lexicaux stockés et étiqueté dans leurs bases de données, en suivant des règles… secrètes.