La régulation des réseaux sociaux : l'intelligence artificielle, une solution pour la modération des contenus sur les réseaux sociaux ?

2021 - 2022

Rédigé par : MENARD Stéfan
Directeur de mémoire : LAKEL Amar

Master 2 DNHD, Université Bordeaux Montaigne

Remerciements

Je remercie mon directeur de mémoire M.LAKEL, enseignant chercheur et maître de conférence dans le domaine des sciences de l'information et de la communication, pour son aide dans la recherche du sujet de mémoire et la structuration du plan.

J'adresse mes sincères remerciements à tous les professeurs et intervenants qui ont participé à ma scolarité et amené jusqu'à la rédaction de ce mémoire.

Je tiens à remercier principalement ma famille, pour leur présence et le soutien moral donné durant les difficultés rencontrées.

De plus, je voulais remercier tous mes amis d'avoir été également là pour me soutenir moralement.

Enfin, mon dernier remerciement est à destination de M.BRINI, mon meilleur ami, qui m'a conseillé pour structurer mon mémoire et qui a pris le temps de le relire.

Résumé

Le sujet de la régulation des réseaux sociaux a été repris par les États Européens dernièrement, critiquant les méthodes appliquées par les plateformes pour le contrôle des communautés. Afin de comprendre une partie de ces critiques et les potentielles améliorations, nous allons mener une analyse des procédés de MDC actuelles. L'objectif à travers ce mémoire, est d'examiner les possibilités d'application d'un processus de modération entièrement automatisé, pour but de limiter les erreurs et éviter les problématiques liées à la santé mentale des modérateurs humains qui figurent derrière nos écrans.

De nombreuses sources textuelles et vidéos viennent alimenter les connaissances sur le sujet pour venir répondre à l'objectif présenté. L'analyse de toutes ces sources permet une analyse théorique sur la sociologie des réseaux sociaux et l'usage des règles/méthodes des plateformes concernant la régulation. Puis, se suit une étude davantage technique sur les procédés de modérations existants et l'emploie de l'IA dans ce domaine.

À la fin des recherches, les clés de compréhension exposées donnent la possibilité de comprendre la place des systèmes établis sur l'intelligence artificielle, dans le domaine de la modération de contenu.

Abstract

The topic of social network regulation has been taken up by European states lately, criticising the methods applied by platforms to control communities. In order to understand some of these criticisms and potential improvements, we will conduct an analysis of current CBM processes. The objective of this thesis is to examine the possibilities of applying a fully automated moderation process, with the aim of limiting errors and avoiding problems related to the mental health of the human moderators, who appear behind our screens.

Numerous textual and video sources are used to provide knowledge to meet the objective presented. The analysis of all these sources allows a theoretical analysis on the sociology of social networks and the use of the rules/methods of the platforms concerning regulation. This is followed by a more technical study of existing moderation processes and the use of AI in this field.

At the end of the research, the keys of understanding exposed give the possibility to understand the place of systems based on artificial intelligence, in the field of content moderation.

Sommaire

Liste des abréviations

RSV = Réseau social virtuel / Réseaux sociaux virtuel

MDC = Modération de contenu

IA = Intelligence artificielle

ML = Machine Learning

DL = Deep Learning

TALN = Traitement automatique du langage naturel

Liste des figures

Introduction

Depuis le premier RSV en 1997, ces plateformes accroissent chaque année leurs nombres d'utilisateurs. Cette croissance a été un sujet prédominant pour ces plateformes, contrairement à la régulation des communautés, qui n'a en comparaison que peu évolué. Une multitude d'affaires viennent montrer les failles que peut comporter la modération actuelle. On retrouve par exemple, les témoignages des modérateurs humains sur leurs conditions de travail ou l'affaire des « Facebook Files » en 2021 qui dénonce les méthodes appliquées pour la modération.

À l'heure de l'émergence de systèmes d'IA dans nos quotidiens, j'ai décidé à travers ce mémoire de me concentrer sur la régulation des RSV. Plus précisément, sur les méthodes autour de la MDC et les possibilités d'une automatisation totale à l'aide de l'IA. Ma motivation de rédiger sur ce sujet provient d'un premier travail effectué lors de la 1ère année de Master, il consistait à créer une note de veille sur un sujet souhaité. J'ai pu y découvrir et développer sur le sujet de la modération de la plus grande communauté que sont les RSV, je me suis axé grandement sur la modération humaine en regardant l'impact sur leur santé mentale. À la découverte des dégâts sur les modérateurs humains, je me suis posé la question : « Pourquoi ne pas soutenir ou remplacer les modérateurs humains par des IA ? ». C'est pour cela que je me suis penché sur cette question lors de ce mémoire. De plus, je voulais développer mes connaissances sur la sociologie des réseaux sociaux et sur le fonctionnement des systèmes d'IA.

La rédaction de ce mémoire s'est réalisée exclusivement par l'analyse de multiples sources textuelles et vidéos.

Au moyen de ces sources, nous allons pouvoir mettre en lumière les clés de réponse à la problématique suivante : « L'intelligence artificielle, une solution pour la modération des contenus sur les réseaux sociaux ? ».

Durant la première partie, on évoquera les aspects sociologiques fondamentaux autour des réseaux sociaux réels et virtuels. La seconde partie poursuivra par la présentation du parcours des RSV jusqu'aujourd'hui, ainsi que l'analyse des méthodes contrôlant l'engagement des utilisateurs et les règles pour les encadrer. Enfin, la troisième et dernière partie se composera

d'une description détaillée des procédés et techniques de modérations actuelles, puis se suivra un éclaircissement sur la composition et le fonctionnement de l'IA.

1. Sociabilité et réseaux sociaux

De nos jours, lorsque l'on parle de réseau social, on pense principalement aux médias sociaux[1] qui nous connectent aux autres de manières virtuelles. Mais, un réseau social ne se résume pas à un média social, c'est une façon de visualiser les échanges humains. Pour reprendre les paroles du chercheur *Pierre MERKLE*[1], « Un réseau social est un ensemble d'unité sociale et des relations que ces unités entretiennent les unes avec les autres soit directement ou indirectement avec des chaînes ou chemin de longueur variable ». Deux entités sont présentes dans un réseau social, l'unité sociale pouvant correspondre à un individu, une entreprise, etc., puis la relation qui est également diverse pouvant prendre la forme de transmission de connaissance ou d'une simple salutation entre deux voisins. De ce point de vue, on peut affirmer que les réseaux sociaux sont vieux comme le monde.

Avec l'émergence de services permettant de nous connecter de plus en plus, on se retrouve hyperconnecté au monde qui nous entoure, nous faisant oublier la distinction entre les types de réseaux sociaux que l'on côtoie au quotidien. Deux types se distinguent, le réseau social réel et virtuel, cette différence est essentielle, pour mieux cartographier ses relations. Dans ce mémoire, nous nous concentrerons sur les deux types de RSV dit de contact et de contenu[2].

1.1 Les caractéristiques des échanges sociaux

Depuis que les relations humaines existent, nous créons et entretenons des réseaux sociaux, et depuis que cela subsiste, des études démontrent des caractéristiques liées aux relations humaines. Ces caractéristiques, au fil d'études et de recherches, figurent existante qu'importe le type de réseau social. L'aspect principal des relations humaines correspond à la sociabilité, représentant la globalité des échanges sociaux qu'un individu peut développer au quotidien.

[1] Les médias sociaux correspondent aux applications web donnant la possibilité d'échanger entre les individus à travers les réseaux sociaux virtuels, créer du contenu (musical, vidéo, etc.). Il existe une grande variété de média sociaux.

[2] Les réseaux sociaux virtuels de contact sont ceux qui privilégient la création de lien entre les utilisateurs et les réseaux sociaux virtuels sont ceux qui se concentrent sur la création de contenu.

C’est à travers cela que des sous-ensembles viennent compléter et augmenter la complexité de la sociabilité humaine.

L’intention de création et d’entretien de relations porte le nom de « Réseautage », cette pratique s’est vue exploitée d’une manière sans précédent avec les RSV. Toutes interactions se trouvent simplifiées, c’est également le cas pour la recherche de liens à créer, en quelques clics, on a la possibilité d’entrer en contact avec un individu provenant de notre réseau ou celui de nos liens. Ce terme peut s’exploiter pour les contacts sociaux, mais s’est vu grandement employé dans le milieu professionnel, pour de la recherche d’emploi simple, ou bien dans des domaines précis, tel que le marketing web.

Toute origine d’un réseau social provient suite à une interaction avec un individu, ce qui après coup deviendra une liaison dans notre réseau social. Ce concept de lien est essentiel dans le fonctionnement des structures relationnelles et se complète avec le réseautage précédemment cité. Évoqué pour la première fois en 1762 dans l’œuvre « Du contrat social » par Jean-Jacques ROUSSEAU, ce concept s’est vu repris dans les premiers écrits sur la sociologie par Emile DURKHEIM, Ferdinand TÖNNIES ou Max WEBER. Enfin, le sociologue et économiste Mark GRANOVETTER l’a saisi à son tour pour en écrire ces fondements[2] auquel on se réfère encore aujourd’hui. Selon ce dernier, on distingue deux types de liens, les forts et faibles. À propos des liens forts, cela représente les relations entretenues régulièrement, avec qui il se créait une confiance réciproque, ça fait référence aux amis proches et l’entourage familial. Pour ce qui est des liens faibles, ils ont une fonction bien différente, ne demandant pas le même niveau d’engagement et liant principalement des individus socialement et culturellement éloignées, ça renvoie à des connaissances. Comment on peut distinguer la force des liens ? Avec 4 variables, la fréquence de contact, l’intensité émotionnelle, l’intimité et la réciprocité des services rendus. Revenons brièvement sur la théorie de Mark GRANOVETTER, concernant la force des liens faibles. Un lien faible de par son nom peut être sous-estimé sur la puissance et les avantages qu’il peut apporter à un individu. Un individu avec une distance culturelle et sociale (lien faible) permettra d’accroître nos opportunités, d’obtenir des informations nouvelles ou d’agrandir notre cercle social. Contrairement à des liens forts comportant des individus avec une proximité culturelle, nous isolant dans le même cercle et les mêmes informations obtenues.

Prenons un exemple type ou un individu recherche activement un emploi, il y a de grandes chances que nos liens forts nous transmettent des offres déjà connues, contrairement à un lien faible ouvert vers d'autres réseaux méconnus, proposant d'autres possibilités d'emploi. Cela confirme l'idée de sous-ensemble venant complexifier les relations humaines et la composante principale qu'est la sociabilité, qui correspond bien à un ensemble de liens forts et faibles qu'un individu rencontre dans son quotidien.

Les individus ne sont pas interchangeables, nous avons tous une structure sociale et des caractéristiques différentes, ce qui permet une distinction. Cette distinction donne la possibilité de schématiser les relations sociales entre des individus dans un réseau, en apportant une meilleure compréhension des liens et des caractéristiques communes entre ces derniers. Le sociologue Pierre BOURDIEU a repris l'idée de Capital de Karl Marx[3] pour ajouter une dimension sociale approfondie, permettant de mieux percevoir les relations dans la société. Sa théorie[4] s'établit sur la présence de 4 capitaux chez un individu, avec un volume et une répartition différente, obtenus par la suite d'un héritage, de son environnement social/familial ou d'actions menées.

En premier le plus connu, le « Capital Economique », il se compose de l'ensemble des richesses d'un individu, avec son revenu et son patrimoine. En second, le « Capital culturel », qui s'éloigne complètement de l'aspect économique, en représentant l'intégralité des ressources culturelles, en considérant, le savoir et savoir-faire, les compétences, diplômes, etc. Il peut s'exprimer sous trois formes[5], la forme objectivée prenant la forme de biens culturels qu'un individu possède, tel que les livres, peintures, etc. La forme institutionnalisée représentant l'attestation de compétences par un diplôme et la forme incorporée qui va lier la culture à l'individu lui-même, à la suite de différentes activités telles que la lecture d'un livre, la visite d'un musée, etc. Le « Capital social » se rapporte au troisième pilier, qui va se constituer de la globalité des ressources du réseau de relation d'un individu, sur la composition de ce dernier, les stratégies adoptées pour sa création, ainsi que ce qu'il apporte. Enfin, pour qu'un individu obtienne dans son environnement une certaine aura de reconnaissance sociale, il faut compter sur le « Capital Symbolique » se manifestant par des distinctions telles que des médailles ou titres.

À travers les interactions et la structure des réseaux sociaux des individus, nous pouvons visualiser une inégalité sociale qui perdure, affichant la difficulté de ces derniers à sortir de leurs conditions sociales préétablies. De par sa théorie, Pierre BOURDIEU donne des clés de compréhension, pour mesurer et expliquer ses inégalités. Celles-ci sont-elles tout aussi présentes dans les RSV ?

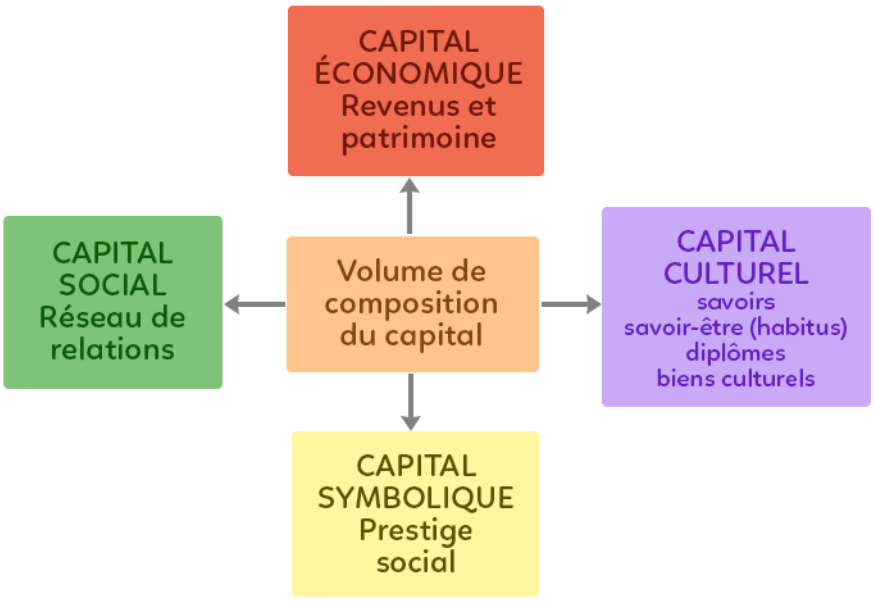

Figure 1 : Les capitaux d'un individu selon Bourdieu

1.2 La sociabilité à l'ère du numérique

Maintenant que nous avons abordé la sociabilité et une partie de ses sous-ensembles, il est intéressant d'analyser comment celle-ci a évolué jusqu'à présent, l'ère des RSV. Dans les débats publics, on entend souvent ressortir que la technologie et les RSV éloignent les individus, en créant une distance entre eux, altérant alors la sociabilité de ces derniers. Cependant, dès les années 1950, le sociologue américain David RIESMAN[6], affiche un constat négatif concernant non pas les RSV mais l'impact de l'évolution des sociétés modernes sur la sociabilité à travers l'analyse de la société américaine. Un bon nombre d'auteurs et chercheurs tels que Robert PUTNAM[7]ou Nathalie BLANPAIN et Jean-Louis PAN KE SHON[8] confirment le constat de David RIESMAN en venant s'accorder à dire que les sociétés modernes apportent un plus grand isolement et éloignement entre les individus. Ces études convergent vers l'idée que la sociabilité se dégrade fortement, bien avant l'émergence des RSV. Qu'est-ce qui en est

de l'émergence des technologies et des RSV dans notre société moderne ? La sociabilité s'est-elle vue encore plus affaiblie ? Les technologies amènent une grande variété de nouvelles interactions, en donnant la possibilité de communiquer de quasiment n'importe où. Cela prend la forme d'un support « en ligne » délivrant une nouvelle pratique de la sociabilité. On aperçoit un entrelacement entre les différentes pratiques de sociabilité, en « face-à-face » et « en ligne » avec les dispositifs technologique et numérique tels que le téléphone, les mails, les RSV, etc. Ces différentes pratiques de sociabilité viennent se compléter ou se substituer aux proportions du besoin des individus. Selon certains, la substitution l'emporte, comme Pierre MERCKLE, qui pour lui la sociabilité ne baisserait pas en France, mais subirait un effet de remplacement des interactions en face-à-face par des interactions en ligne. Pour d'autres telles que Anabel QUAN-HAASE et Alyson YOUNG[9], portent la pensée que les pratiques de sociabilité s'intègrent dans un ensemble de médias utilisés incluant des formes de communications en ligne et hors ligne (face-face).

1.3 Les formes de communications et l'utilisation des RSV

Les formes de communications présentent des différences notables, que ce soit sur la facilité d'accès ou les échanges menées. Elles demandent chacune à l'utilisateur d'adapter sa compréhension de l'autre et de modifier sa manière d'interpréter les discussions.

Chaque individu se compose différemment, dû à sa classe social et son vécu, créant des préférences de canaux de communication. Concernant les RSV, ils proposent un large panel d'avantages pouvant justifier le choix de certains qui se tournent vers eux. On peut principalement citer la réduction des inégalités sociales entre les individus, qui permet d'amplifier le capital social de chacun en élargissant les échanges possibles. De plus, cela atténue également le jugement de classe sociale, justifié par une visibilité moindre du capital économique et culturelle. Puis, l'accès au savoir simplifié accroit le capital culturel qui participe à la baisse des inégalités sociales.

Une étude[10] de Bénédicte AFFO et Olivier ROQUES a permis de dessiner le profil type des individus pouvant pleinement jouir de la forme de communication du RSV. Celui-ci, se défini par un individu moins favorisé en capital social et sociabilité, qui a la possibilité de s'orienter vers les RSV pour combler ce manque dans la pratique en face-à-face. Ils se sont basés sur deux variables composant le concept de « soutien social »[11] provenant des réseaux, le

« capital social », correspondant à l'aisance pour élargir son réseau, puis les « qualités sociales individuelles »[12] qui conditionnent nos échanges avec les autres.

Maintenant que l'on connaît la catégorie d'individu à qui cette forme de communication sert le plus, il est intéressant d'analyser les principales raisons qui poussent l'utilisation des RSV. On en décèle 3 dominantes, celle de poster son quotidien, pour laisser une trace comme une forme de journal intime. La suivante consiste à garder du lien ou s'en créer au moyen des échanges privés avec d'autres individus. Puis, obtenir de l'attention des autres utilisateurs, rappelant le concept de « soutien social », davantage essentiel en ce lieu de communication. Ce sont les chercheurs Lee SANG-HOON et Kim YO-HAN qui ont démontré ces pratiques, lors d'une enquête[13] auprès d'étudiants coréens.

2. Emergence des réseaux sociaux virtuels

La troisième révolution industrielle[3] a mené bien des changements dans le monde professionnel et notre société, elle est causée par les technologies, mais également par la naissance du World Wide Web (www). Jusque-là, internet se rapportait à un immense réseau d'ordinateurs connectés entre eux, étant capable d'échanger des fichiers et envoyer des messages par courrier électronique. C'est alors qu'en 1989, un chercheur Britannique du nom Tim BERNERS-LEE voulait fluidifier les échanges scientifiques sur Internet. Pour répondre à cette ambition, il inventa le World Wide Web, qui par la suite s'est vu intégré et étendu comme nouvelle application du réseau Internet, modifiant la manière de partager les informations, que ce soit sur la forme ou sur l'accès se voyant simplifié. Cette innovation se base sur le système hypertext[4], qui se déploie sur un fonctionnement tripartite. L'URL (Uniform Ressource Locator) se rapportant à l'adresse d'un document, d'une page web, le protocole HTTP (Hypertext Transfer Protocol) définissant les règles de communication entre la machine de l'utilisateur et la page web et le HTML (Hypertext Markug Language) correspondant au

[3] La troisième révolution industrielle désigne un changement industriel et économique dans le secteur d'activité des nouvelles technologies de l'information et de la communication. Introduit et popularisé par l'économiste Américain Jérémy RIFKIN.

[4] Le système d'Hypertext fut créé par le sociologue Américain Ted NELSON, cela représentante un document qui donne la possibilité à l'utilisateur de naviguer entre plusieurs informations à l'aide de redirection (hyperliens).

langage structurant l'information, qui demande à la machine d'interpréter ce code pour que la page web soit compréhensible par l'utilisateur.

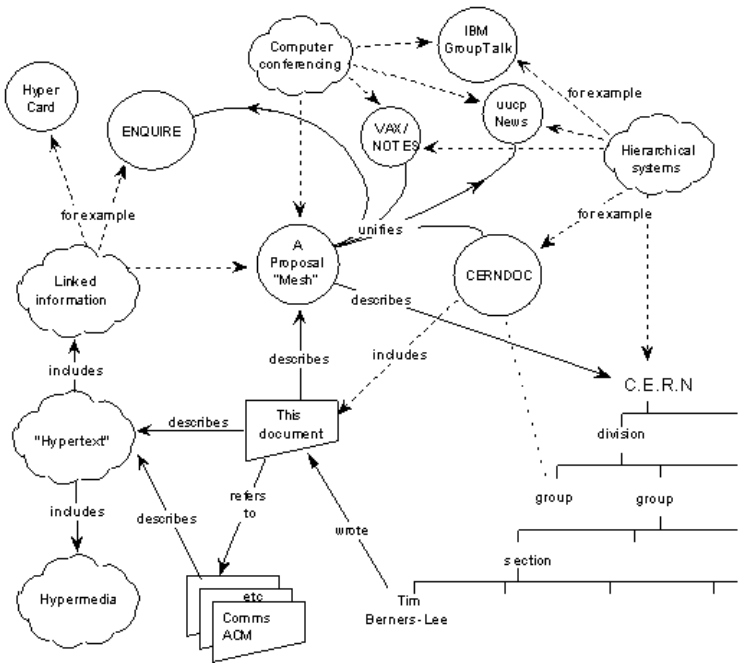

Figure 2: Première représentation d'une requête sur le World Wide Web par Tim BERNERS-LEE

Par la suite, internet et le web ont pu se développer et s'insérer dans les entreprises et les foyers, grâce à la baisse du prix des ordinateurs, la création des premiers navigateurs web facilitant la recherche d'informations et les serveurs web hébergeant les ressources contenants les pages web.

Depuis son implantation dans notre société, le Web est en constante mutation, se décomposant sur plusieurs générations, exposant des changements majeurs relatifs à sa puissance, son usage, son architecture, etc. À ce jour, ces différentes versions du web[5] viennent coexister et se compléter sur le réseau informatique pour créer le web d'aujourd'hui et de demain.

À propos des médias sociaux, et en particulier des RSV, le premier considérait comme tel est né en 1997, sous le nom de Six Degrees qui offrait des pratiques élémentaires de la sociabilité virtuelles, avec la création, l'invitation et la consultation de profils. Cependant, à ce moment

[5] Majoritairement le web se décrit en 5 générations à ce jour, mais d'autres le définis autrement, en nommant les mouvements et utilisations qui font partie intégrante dans l'histoire de l'évolution du Web (exemple : le Web mobile, le Web Sémantique, etc.).

de l'histoire du web, ce type d'usage ne se trouvait pas adapté, préférant celui de la lecture et l'acquisition d'information sur la toile, lié à la première génération (Web 1.0) qui se limitait surtout à l'emploi d'annuaires ou de moteurs de recherche. Les RSV ont réellement pris place dans notre société lors de l'expansion de la vision des utilisateurs sur l'utilisation du web, ils ne se cantonnaient plus seulement à lire du contenu, mais également à en créer. Ce changement représentait le web 2.0, définition qui s'est vue utilisée pour la première fois en 2003[6]. À partir de cette date, ce fut la création d'une multitude de RSV, prenant différentes formes, en proposant des fonctionnalités exclusives et ciblant des communautés. Ils se distinguent en deux types, ceux qui privilégient la mise en relation entre les utilisateurs (Facebook, LinkedIn, etc.) et ceux qui favorisent le partage de contenu sous les différents aspects possibles (YouTube (Vidéo), Twitch (Vidéo), Spotify (Audio), Pinterest (Image), etc.).

Depuis le premier réseau social en 1997, le seul support utilisable pour utiliser les RSV demeurait l'ordinateur jusqu'à que les Smartphones soient accessibles au marché mondial et qu'Instagram apparaissent, faisant de lui en 2010, le premier RSV sur ce support. L'apparition de ce nouveau support modifia complètement la façon d'échanger, de consommer du contenu et également de manipuler le web. Les chiffres de l'étude annuelle de We Are Social et Hootsuite viennent confirmer ce changement survenu par l'émergence des smartphones, en révélant qu'à ce jour, 67,1% de la population mondiale utilise un mobile[14].

3. Le flux sur les réseaux sociaux virtuels

Aujourd'hui, on détecte 4,62 milliards d'utilisateurs présents sur les RSV, ce qui représente 58,4 % de la population mondiale. Comment les géants du web ont pu attirer autant de personnes sur leurs plateformes ? Cela peut s'expliquer par leur Business Model demandant un mouvement continuel vers l'acquisition de profit. Le fondement de celui-ci s'établit sur la revente des informations des utilisateurs, mais également sur l'affichage des publicités des annonceurs. Les réels clients se trouvent être les annonceurs et entreprises achetant les données utilisateurs et non pas ces derniers occupants le statut de produit. Concernant les annonceurs, on remarque l'application d'une économie type, le capitalisme de surveillance,

[6] Le terme Web 2.0 a vu son apparition pour la première fois en 2003 par Dale DOUGHERTY (Cofondateur de la société d'édition O'Reilly Media).

créant du profit avec la traque des faits et gestes des utilisateurs sur les RSV, entrainant la prédiction de leurs actions, permettant l'assurance d'une réussite de l'engagement des publicités des annonceurs. C'est devenu un commerce des données humaines à échelle industrielle.

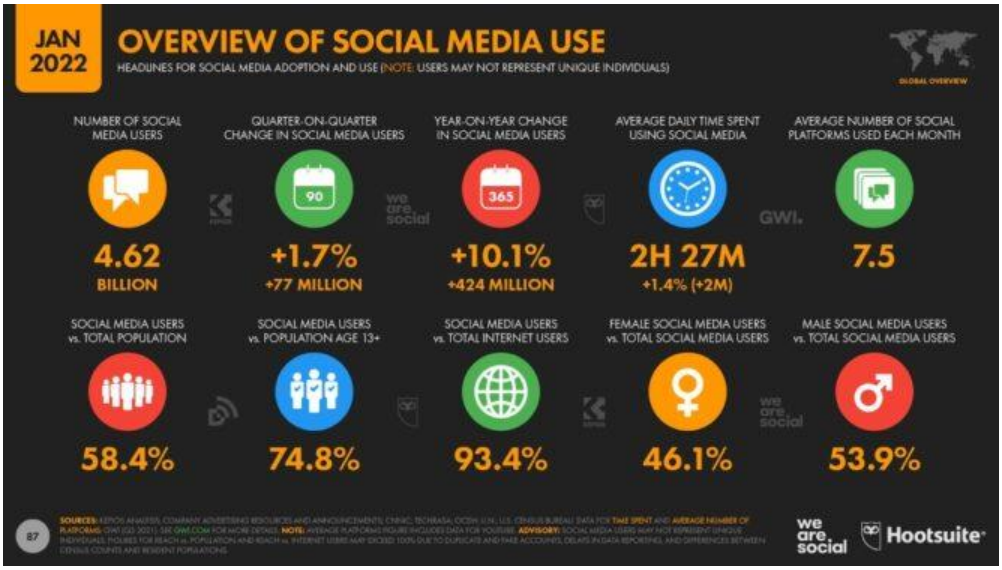

Figure 3 : Données sur l'utilisation d'Internet et des RSV (source : Etude de We are social et Hootsuite)

3.1 Enjeux et fonctionnement

Pour mener à bien ce Business Model et garder la confiance des annonceurs, ces plateformes se concentrent sur la maximisation du temps d'utilisation et l'engagement des utilisateurs, la croissance du nombre d'utilisateurs présents, passant par l'influence des autres utilisateurs envers leur entourage proche ou non, puis la récupération de revenu provenant des publicités (coût par clic par exemple).

La pierre angulaire est l'engagement cité précédemment, pour ses entreprises c'est un point central de leur modèle économique qui se révèle être l'économie de l'attention[7]. Si les utilisateurs passent peu de temps sur la plateforme, cela engendrera un effet boule de neige, en attirant peu de nouveaux utilisateurs et en abaissant l'efficacité des publicités promise aux annonceurs. En réponse à ce besoin, l'analyse comportementale des utilisateurs est devenue

[7] L'économie de l'attention est une branche des sciences économiques qui considère l'attention d'un individu comme un produit à grande valeur, pour l'inciter à effectuer un acte monétaire. Elle vu le jour au début du 20 ème siècle par le sociologue Français Gabriel TARDE.

essentielle (temps passé sur une vidéo ou photo, les likes attribués, les réactions données sous les posts, etc.), donnant les clés pour instaurer des dispositifs communs ou personnalisés, pour capter notre attention et nous influencer augmentant l'engagement. Avec la récupération de ces informations comportementales, ses plateformes sont devenues des spécialistes de la manipulation à travers les différents dispositifs mis en place. Cette manipulation se rapporte à une science, la captologie[8] ou plus récemment surnommé le design persuasif, qui s'apparente à des techniques de persuasion s'appuyant sur les vulnérabilités comportementales, dans le but de modifier la façon de penser et le comportement des utilisateurs. Ceci passe par l'étude de l'influence et des motivations de ces derniers. Ces techniques de persuasion vont utiliser des concepts de sociologies en se traduisant sous des dispositifs tels que des algorithmes de recommandation ou des fonctionnalités.

Figure 4: Exemples de vulnérabilités utilisées en captologie

Les algorithmes sont de plus en plus en précis, profitant des données comportementales captées, enrichissant les modèles pour prédire le comportement à venir et le contenu à recommander aux utilisateurs. On retrouve cette pratique sur quasiment tous les RSV, avec l'exposition de contenu quasiment exclusif en rapport avec les intérêts et les idées de

[8] La captologie a été inventé par le sociologue Américain B.J FOGG en 1996. Cet acronyme provient de l'anglais Captology qui décrit « Computers As Persuasive Technologies ». Cela correspond à l'étude de l'influence de l'informatique sur le comportement des utilisateurs qui s'inspire de la théorie de la manipulation créé par le professeur en psychologie sociale Charles Adolphus KIESLER en 1970.

l'utilisateur, on appelle ça les bulles de filtres[9]. Pour décrire ces bulles de filtres, on peut citer les concepts de chambre d'écho[10] et d'auto-propagande qui conviennent à l'affichage exacerbé d'informations en lien avec les idées de l'utilisateur. Cette notion est régulièrement contestée, se justifiant par les problèmes qu'elle met en lumière, en amplifiant la polarisation et la radicalisation de certains en les renfermant sur leurs opinions.

Pour ce qui est des fonctionnalités, on en retrouve une multitude sur les différents RSV coïncidant avec des mécanismes de persuasion, venant hacker les cerveaux pour modifier nos comportements au quotidien. Voici une liste non exhaustive pour donner des exemples concrets appliqués sur les RSV :

- Le mécanisme de la « Récompense aléatoire »[11], provient des travaux (Boite de Skinner[15]) du psychologue américain B. F SKINNER – Exemple : le swipe, les notifications, le rafraîchissement du fil d'actualité (correspond au mouvement des machines à sous), etc.
- L'approbation sociale et la réciprocité sociale, techniques de détournement jouant sur le besoin d'appartenance, d'acceptation de son réseau, le besoin de maintenir des interactions sociales et de se comparer socialement – Exemple : les tags, les likes, les commentaires, les réactions (a permis d'analyser le comportement émotionnel des utilisateurs sur la base des 6 émotions universelles[16]), le « vu » dans les messages, l'animation des « … » de suspensions lorsque notre interlocuteur écrit, filtre photo, etc.
- La peur de manquer un événement, technique de détournement aussi nommé FOMSI (Fear of Missing Something Important) - Exemple : Lié directement au comportement des utilisateurs modifié par la totalité des designs persuasifs. Ce qui amène à des

[9] Les bulle de filtres correspond au processus suivi par les algorithmes recommandant du contenu en lien avec les préférences des utilisateurs. Apparu en 2011 par le militant internet Américain Eli PARISER.

[10] La chambre d'écho est un mécanisme médiatique amplifiant la répétition de l'apparition de contenu sur les idées de l'individues. Ce terme est introduit pour la première fois par le lobbyiste John SCRUGGS en 1998.

[11] Le système de récompense est une fonction fondamentale des mammifères, il vient fournir les récompenses pour motiver la réalisation d'actions pour notre survie. Il existe plusieurs situations ou substances qui viennent altérer ce système de récompense, comme celui de la récompense aléatoire en lien avec les jeux d'argent. La récompense aléatoire vient augmenter le désir de l'individu à effectuer la tâche de manière plus régulière, sans se lasser. Cette récompense s'est révélée lors de l'expérience de la boite de Skinner.

situations comme le scrolling infini de son fil d'actualité par peur de manquer une information.

- Les mécanismes de gamification[12] et d'aversion à la perte[13] sont sollicités avec la mise en place d'une forme de jeux quotidiens liés à des interactions sociales et produisant une peur de la perte plus importante que celle du gain– Exemple : la fonctionnalité des flammes sur le RSV Snapchat (phénomène de Streak, série consécutive de photo/snap envoyé à une personne, provoquant le lancement de la fonctionnalité des flammes. Cette dernière entre dans notre quotidien en demandant un envoi de photo toutes les 24H sinon elle disparait).
- Le flux infini, technique de détournement pour conserver les utilisateurs un maximum de temps sur les RSV, s'appuie sur l'effet de Zeigarnik[14] - Exemple : l'autoplay à la fin d'une vidéo.

3.2 Les conséquences des pratiques des RSV

Toutes ces pratiques de manipulations des utilisateurs enrichissent les plateformes en affectant le comportement des utilisateurs. Cet agissement à des conséquences néfastes sur ces derniers, on observe une forme d'addiction qui s'installe dans leurs quotidiens, c'est l'addiction comportementale. Elle consiste selon le psychiatre américain Aviel GOODMAN *« processus dans lequel est réalisé un comportement qui peut avoir pour fonction de procurer du plaisir et soulager un malaise intérieur, et qui se caractérise par l'échec répété de son contrôle et sa persistance en dépit des conséquences négatives »*. Elle conduit à l'apparition d'une perte de contrôle sur le niveau de consommation, d'une instabilité émotionnelle, de problèmes médicaux selon la pratique, etc. Présentement, les comportements associés à ce genre d'addiction sont le jeu d'argent, l'addiction à internet et aux jeux vidéo, le sport, le sexe, le travail et les achats compulsifs.

---

[12] La gamification ou en français ludification, correspond à la mise en place d'un mécanisme de jeu pour des tâches hors de ce domaine.

[13] L'aversion à la perte est une notion provenant du champ de l'économie comportementale, exposant un biais comportemental de l'Homme qui montre davantage d'importance dans la perte que dans le gain.

[14] L'effet Zeigarnik provient d'une expérience menée en 1929 sur des humains par la psychologue Russe Bljuma ZEIGARNIK. Cela consistait à demander aux sujets d'exécuter une série de tâches, en interrompant les tâches pour certains et en laissant d'autres aller jusqu'au bout. Elle en conclut que les sujets retiennent mieux les tâches non terminées (interrompues) que les tâches terminées et que le fait de débuter une tâche crée l'envie de la continuer.

Les acteurs responsables de cette addiction se situent dans le cerveau, avec le circuit de récompense et le neurotransmetteur portant le nom de dopamine. C'est ce dernier qui joue sur le comportement, avec le déclenchement de la motivation de mener des tâches. Une addiction se bâtit sur trois étapes, avec dans un premier temps, **la recherche de plaisir** qui débute par le stimulus du circuit de récompense et la libération de la dopamine dû à la pratique. Lorsque la consommation de cette pratique est répétée, l'individu va recevoir de la dopamine par anticipation pour prédire l'arrivée de la récompense, on appelle ce processus le renforcement. Dans un second temps, vient le **déséquilibre de l'état émotionnel** qui provient d'une baisse de dopamine libérée lors de la pratique, provoquant un état émotionnel négatif. La consommation excessive de la pratique se voit alors être la solution pour palier à la baisse de dopamine et pour satisfaire le système de récompense ainsi qu'équilibrer l'état émotionnel. On ne cherche plus à prendre du plaisir, mais à se soulager. Puis survient la **perte de contrôle** dans un troisième et dernier temps, qui va altérer le circuit de récompense et des émotions, corrompant nos prises de décisions et capacités à résister à la consommation de la pratique. C'est problématique puisque comme nous avons pu le voir précédemment, notre mécanisme de défense se fonde sur nos émotions, nous induisant maintenant en erreur lors de certaines pratiques. Les experts de la captologie ont bien compris que ce biais cognitif était primordial pour rendre addict les utilisateurs aux RSV.

« Seulement deux industries appellent leurs clients des « utilisateurs », celle de la drogue et celle du logiciel », Edward TUFTE, professeur américain de statistiques, informatique, design de l'information et économie politique à l'université de Yale

L'émergence des RSV sur nos téléphones en 2010 a étendue d'une manière importante leurs temps d'utilisation, aggravant par la même occasion l'addiction comportementale. Comme énoncé lors des 3 étapes à la genèse d'une addiction, on aperçoit un état émotionnel déséquilibré qui peut mener à de l'anxiété et même de la dépression. C'est ce qu'ont remarqué les hôpitaux en 2010, avec une forte croissance concernant les mutilations[17] et suicides[18] chez les jeunes. Les RSV peuvent représenter une des causes de cette croissance.

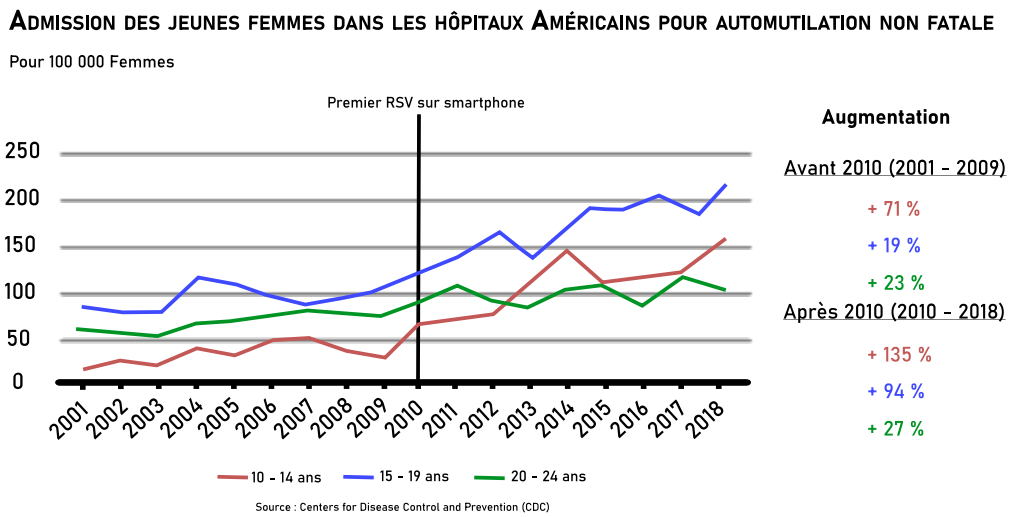

Figure 5: Chiffre des auto-mutilations non meurtrière chez les jeunes femmes Américaines (voir annexe 1)

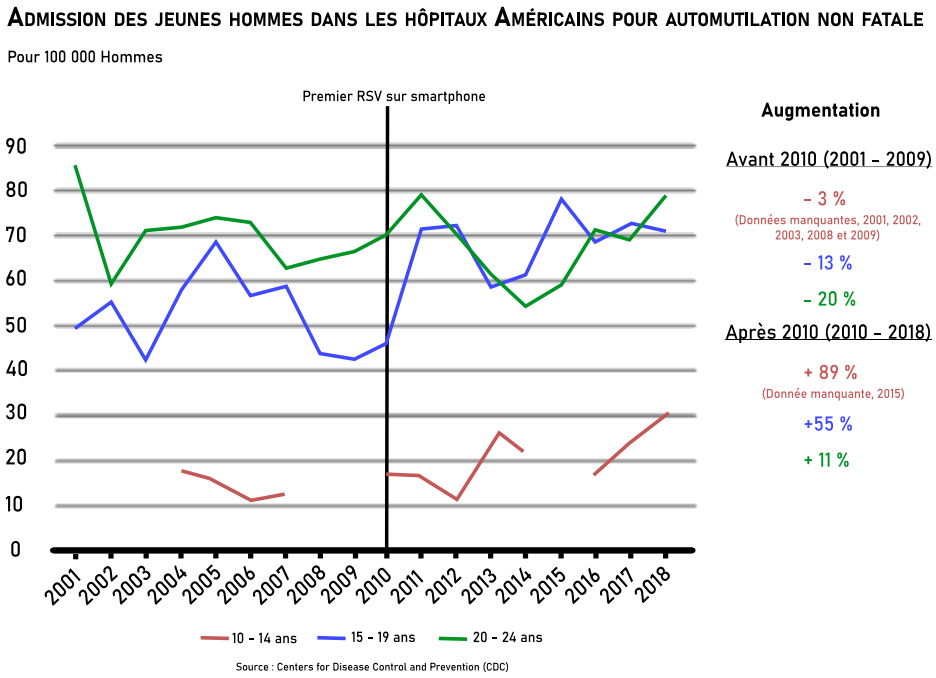

Figure 6: Chiffre des auto-mutilations non meurtrière chez les jeunes hommes Américains (voir annexe 1)

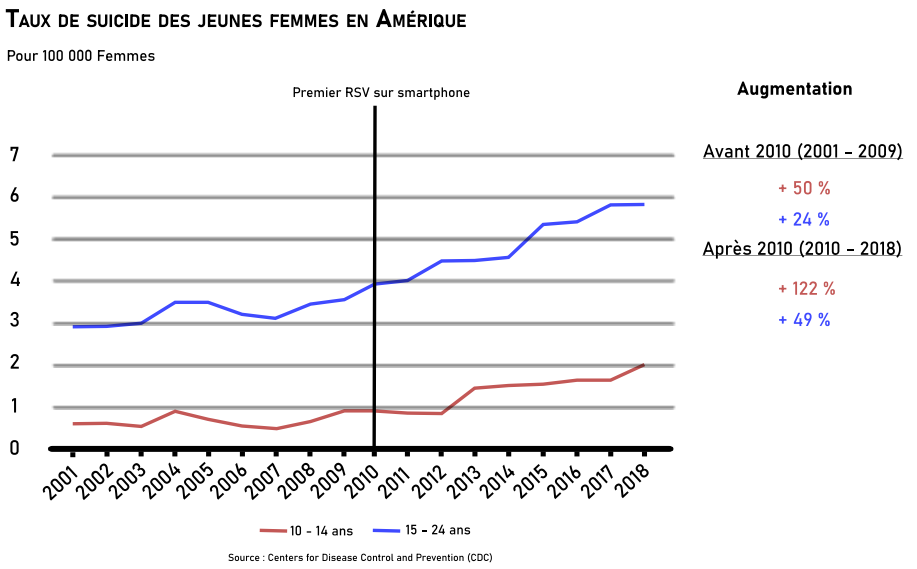

Figure 7: Chiffre des suicides chez les jeunes femmes Américaines (voir annexe 1)

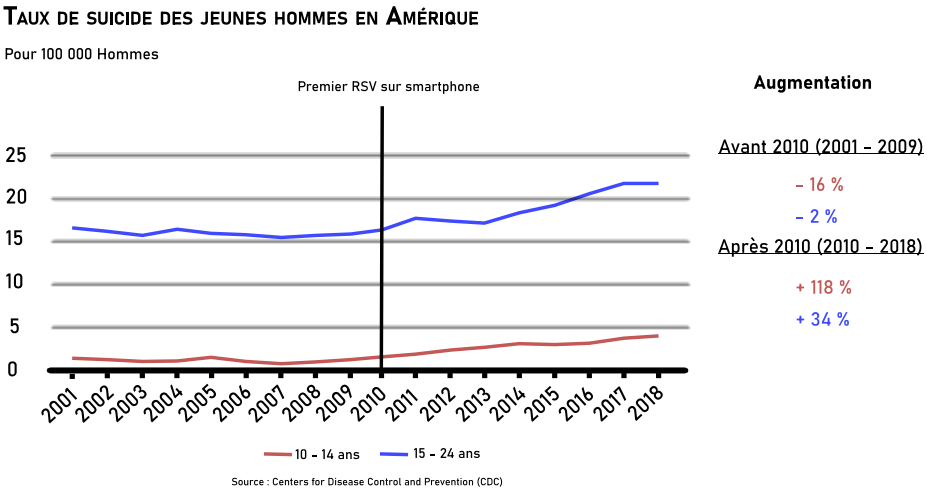

Figure 8: Chiffre des suicides chez les jeunes hommes Américains (voir annexe 1)

3.3 L'éthique dans le design

Ces dernières années les RSV ont abusé du design persuasif pour continuer leur croissance économique, sans se soucier réellement du besoin des utilisateurs et l'impact sur ces derniers. Au vu des résultats sur la société, des anciens employés de ces plateformes (Tristan HARRIS, Mike MONTEIRO, etc.) se repentissent et dénoncent les méthodes de manipulations qu'ils ont pu appliquer lors de leurs passages. Ils demandent davantage d'éthique au sein des équipes de designer, c'est le lancement du mouvement pour obtenir un design éthique. La prise de responsabilité des designers sur l'importance de leur rôle est la principale composante pour apercevoir un changement au sein des plateformes dans le futur. Cette composante n'est pas neuve, le designer autrichien et américain Victor PAPANEK en fait déjà mention en 1970,[19]

en déclarant que « Le design est devenu l'outil le plus puissant avec lequel l'Homme forme ses outils et son environnement. ». Le design influe sur plusieurs sujets essentiels de notre société. Dans les grands principes du design éthique, on identifie les problèmes d'accessibilité pour inclure tous les types d'utilisateurs, de confidentialité des données, de transparence et d'implication des utilisateurs sur les méthodes d'engagement employées pour leur en laisser la décision, de la liberté de la concentration des utilisateurs en mettant à disposition des services utiles ponctuellement et la durabilité des services développés et façonnés. Comme initiative qui appuie ce mouvement et donne des clés pour éduquer les designers sur ces enjeux, nous avons le User Experience Professionnals Association (UXPA) qui propose depuis 2005 un code de conduite concentré sur le besoin des utilisateurs. En France s'est vu la création en 2016 d'une association, les Designers Éthiques, qui instruise et sensibilise les designers sur ces questions à travers un blog et des conférences annuelles. Néanmoins, l'influence comportementale liée au design persuasif peut s'aborder de manière responsable, positive et éthique selon l'angle utilisé lors de sa création. On en a la preuve dans le sport avec des applications qui récompense les utilisateurs selon la distance parcourue ou avec le nudge dans divers domaines qui incite à se diriger vers la meilleure décision possible.

4. Encadrement de la parole des réseaux sociaux virtuels

La création du web repose sur des idées libertaires, prônant la liberté d'expression, ou l'on apercevait un usage décentralisé, avec de petites communautés intéressées par un même sujet, se trouvant réparties sur le web, à travers diverses plateformes tel que des forums. Parmi ces petites communautés, des utilisateurs portaient le rôle de modérateur bénévolement, en se référant à une charte de modération[15] disponible de tous pour une transparence.

Lors de l'avènement des RSV, un chamboulement du système a eu lieu, avec la centralisation de nos usages autour de services communs (Instagram, YouTube, etc.) proposant une multitude de sujets traités hébergés sous des interfaces modernes, regroupant une communauté massive composés de tous types d'individus. Auparavant, la communication

[15] La charte de modération est un document contenant les règles d'usage pour publier sur un site. Utile pour l'utilisateur et le modérateur, définissant les limites à ne pas dépasser.

pour atteindre un large public était lente et couteuse et se menait par des journalistes dans des canaux médiatiques tels que les journaux papier ou télévisés. De nos jours, les plateformes ont permis de démocratiser la parole à tous les utilisateurs d'internet en supprimant les obstacles de temps et de prix. Ce qui a engendré l'apparition d'une masse de contenus indésirables et une modification du rapport à l'information et la connaissance en mettant chaque parole à la même importance, rendant compliqué la distinction de la véracité des discours publiés. Ces conséquences ont créé un besoin de MDC, qui se rapporte aux mécanismes pour gouverner les échanges d'utilisateurs dans une communauté, facilitant la cohabitation et prévenant les abus.

Dans les deux cas d'usage centralisé ou décentralisé, les données des utilisateurs qui transitent sont majoritairement traitées par une entité supérieure et centrale, n'affichant aucune transparence sur l'emploi et les modifications apportées. C'est le système de centralisation des données (adopte un réseau de type client-serveur[16]), mais celui-ci se voit aujourd'hui confronté à un mouvement de décentralisation des données, avec des systèmes décentralisés[17] (exploite un réseau de type peer-to-peer[18]) tel que la blockchain.

4.1 Les acteurs et les règles de modération

Auparavant, l'État conduisait plusieurs tentatives sans résultat, pour modérer la parole sur le web, cet échec peut s'expliquer par la présence d'une communauté massive étendue sur le web provenant de la décentralisation des usages. Cependant, cela s'est vu simplifié avec la centralisation de ses derniers, en rapprochant autour de mêmes plateformes les communautés autrefois réparties. Avec ce changement, les États prit la décision de déléguer le droit de modération à ses plateformes.

Comment ses plateformes effectuent la MDC ? À savoir, qu'actuellement il n'existe aucune législation propre aux RSV. Aux yeux de l'État, les RSV sont considérés comme des espaces publics, ils doivent appliquer les droits fondamentaux que sont la liberté d'expression, le droit

[16] Le modèle de réseau client-serveur qui s'établit sur un serveur qui stocke les données et renvoie les données demandées par les clients (les machines des utilisateurs).

[17] L'idée des systèmes de centralisation de la donnée, c'est de rendre garants plusieurs utilisateurs de l'intégrité des données du réseau et non plus une seule et même entité.

[18] Le modèle de réseau Peer-to-Peer ne possède pas de serveur fixe, chaque machine d'un réseau peut être client et serveur, partageant ainsi les données entre les machines des utilisateurs.

à l'image et le droit d'auteur. Pour ce qui est de la liberté d'expression, celle-ci se voit protégée par les documents du Pacte International Relatif aux Droits Civils et Politiques[20] (PIDCP, article 19) et la Déclaration Universelle des Droits de l'Homme[21] (DUDH, article 19) qui dit « Tout individu a droit à la liberté d'opinion et d'expression, ce qui implique le droit de ne pas être inquiété pour ses opinions et celui de chercher, de recevoir et de répandre, sans considérations de frontières, les informations et les idées par quelque moyen d'expression que ce soit. ». Néanmoins, le droit à la liberté d'expression n'est pas absolu, la loi Française en précise les limites, « sont considérés comme illicites tous les propos qui portent atteinte à l'honneur, à la vie privée ou à la réputation, les injures ciblées, la diffamation ainsi que les propos qui incitent à la haine raciale, à la xénophobie ou qui font l'apologie des crimes contre l'humanité ». La loi interdit également les services illicites et la diffusion de fausses informations (fake news) au vu des conséquences que ça engendre sur notre société, dû à la place importante qu'ils occupent sur l'actualité. Chaque RSV applique différemment ses règles de modération selon les lois de chaque État, à travers une charte de modération ou les conditions générales d'utilisation (CGU).

4.2 Les problèmes que ça pose

Ces plateformes ont maintenant un rôle structurant sur notre société et le débat public, devenant aussi puissant que les autres médias, demandant une régulation juste pour préserver les droits individuels et non les contraindre. On peut mesurer la puissance d'influence de ce média avec les élections présidentielles, ou nombre de contenus et de débat apparaissent sur ces plateformes, influant le choix du vote pour certains. En effet, en 2010, lors de l'élection américaine de mi-mandat, apparait une fonctionnalité, le bouton « J'ai voté », permettant d'indiquer que l'on a voté, avec un statut le signalant. Selon Facebook, cela a conduit à 340 000 bulletins de vote supplémentaires, pouvant être confirmé par la hausse de 5 millions de votes entre 2006 et 2010[22].

Cette responsabilité de la régulation s'est vue attribuée aux plateformes, provoquant plusieurs problèmes, tel que l'accroissement d'une censure moderne, plus étatique, mais privatisé, comme une majeure partie de notre société[23]. La censure est devenue alors incontrôlée de par les abus de suppression de contenus légitime ou à l'inverse des contenus

illicites non supprimés. De plus, elle est opaque, n'exposant aucunement les processus de censure qu'importe le type de modération employé.

De surcroit, leur statut de quasi intouchable face à la justice se voit reproché, considéré seulement comme des hébergeurs[19], cela les protège lors d'un procès en cas d'erreur de modération, arguant que seul le propriétaire du contenu émis est responsable. Enfin, la régulation des RSV se retrouve hissée en haut de l'actualité avec l'affaire des « Facebook Files » en Septembre 2021, ou le Wall Street Journal publient des documents internes du RSV Facebook. On apprend que ces documents ont été remis par Frances HAUGEN, une ancienne employée de l'entreprise. Ces documents internes témoignent de l'utilisation par la firme d'algorithme de recommandation favorisant la désinformation et les contenus de haines, ainsi que la présence d'une modération à double vitesse avec son programme nommé « XCheck » distinguant les utilisateurs basiques et VIP.

4.3 Les solutions des Etats

Toutes ces problématiques soulevées par bon nombre de médias, viennent appuyer le désir des pays de reprendre un certain contrôle de la régulation de ces plateformes. Ils envisagent d'appliquer un encadrement pour redonner une lisibilité de la modération, pour protéger réellement les données utilisateurs et pour rééquilibrer le rôle des plateformes en leur faisant respecter les principes de l'État de droit[20] des différents États. Plusieurs initiatives apparaissent dans les États pour venir encadrer ses géants du web, on peut citer dans notre cas, la France, voulant contrôler la haine en ligne avec la création d'un observatoire pour l'analyse de ce phénomène et la mise en place de contrainte aux plateformes, concernant le traitement des signalements. Ce désir s'est vu retranscrit à travers la loi Avia[24], officiellement adoptée par l'Assemblée Nationale le 13 mai 2020, toutefois une partie des contraintes fut

[19] Il existe deux statuts pour les plateformes web, les hébergeurs sont les plateformes qui mettent à disposition sans modification tout type de contenu du public composant ses utilisateurs de celle-ci (4Chan, Reddit, etc.). Le second statut les éditeurs, qui permettent également d'afficher du contenu mais avec un travail sur l'information avant la communication à ses utilisateurs (Google, Bing, les blogs, etc.).

[20] L'Etat de droit est un système institutionnel dans lequel la puissance publique et tous les sujets de droits sont soumis de la même manière aux droits. Le but est de préserver les libertés et les droits fondamentaux. Rapport d'égalité entre l'Etat et les citoyens.

supprimée par le Conseil Constitutionnel, qui jugea une atteinte excessive à la liberté d'expression, allant à l'encontre de la Constitution. Le pouvoir européen a également cette intention, avec la loi Digital Act Service (DSA)[25] dévoilé en décembre 2020 qui a pour objectif d'appliquer un encadrement commun des géants du web pour tous les États membres. Elle s'attaque à deux parties, le « service » visant la partie modération de contenu, en venant responsabiliser davantage les hébergeurs des plateformes, exigeant l'application de demandes telles que la simplification du signalement pour les utilisateurs, l'augmentation des moyens pour la modération ou la transparence de ces dernières concernant les algorithmes utilisés. La seconde se concentre sur le « marché », en empêchant une domination totale de ces géants, pour laisser une émergence possible à des acteurs alternatifs, pour garder un milieu concurrentiel.

5. Les types de modérations présentes

L'enjeu de la modération à l'ère des RSV est primordiale, regroupant des communautés massives et représentant les lieux principaux du débat public, influant sur les opinions et choix d'avenir des peuples concernant leurs pays. C'est une grande complexité de construire une cohabitation saine entre des utilisateurs hyperconnectés de tous âges et ethnies différentes, en protégeant les plus jeunes des contenus inadaptés ou en essayant de garder une neutralité dans les débats publics lors de la MDC pour respecter les régimes politiques des États. À ce jour, on ne décompte pas moins de 100 millions de photos et vidéos postées par jour sur Instagram[26], sur Facebook, c'est 350 millions de photos postées chaque jour[27].

Au début de l'émergence des RSV dans notre société, les entreprises dans ce secteur se concentraient surtout sur les technologies et fonctionnalités pouvant améliorer leurs plateformes. Elles ont pu prospérer efficacement, en portant le statut de Fournisseur d'Accès Internet (FAI) aux Etats-Unis, les considérant comme des réels hébergeurs de contenus. Ce statut leur a permis d'écarter toutes les problématiques liées aux contenus présents, les dispensant de mener une politique de MDC stricte et de subir des problèmes juridiques. Néanmoins, ces entreprises ont toujours effectué de la modération, elle était minime et invisible au regard des utilisateurs. En réponse à la croissance du flux des utilisateurs, de l'augmentation des contenus répréhensibles et de la pression du public et des

gouvernements, les plateformes ont dû créer et mettre en œuvre des politiques de contenu et des processus de MDC stricte, tout en conservant une opacité autour de son fonctionnement. L'instauration de ce dispositif s'explique également par le Business Model qu'elles abordent, voulant satisfaire les utilisateurs (engagement) et les annonceurs/marques (publicités des marques présentes sur des plateformes friendly) en prônant la sécurité et l'expérience utilisateur positive. Récemment, on a retrouvé des exemples qui démontrent la pleine prise en charge du rôle d'organisateur du débat public en ligne, en prenant des décisions importantes tel que l'exclusion de l'ex-président Américain de plusieurs RSV, ou avec Facebook qui a monté une cour suprême qui délègue le jugement de certains cas de MDC à des acteurs indépendants et externes de l'entreprise. Concernant leurs décisions et la régulation des plateformes, on a pu apercevoir de nombreuses critiques dans les parties antérieures, incitant les États à agir.

Des alternatives existent aux différents RSV, on les appelle les technologies alternatives ou alt-tech. Celles-ci sont nombreuses, prônant la liberté d'expression quasiment absolue, n'effectuant que peu de MDC. Cette direction les a popularisés majoritairement auprès des communautés de l'extrême droite, offrant la liberté de parole sur des sujets tels que le racisme, la xénophobie, etc. On recense plusieurs actes dramatiques orchestrés par des utilisateurs de ces alt-tech, qui postaient du contenu idéologique extrême, justifiant leurs actes barbares. Par exemple, l'attentat de la synagogue de Pittsburgh[28] le 27 Octobre 2018 par un utilisateur du RSV alt-tech Gab ou celui d'El Paso[29] le 03 Aout 2019 par un utilisateur de 8kun. Il existe divers RSV alt-tech reprenant les codes des RSV grand public, en voici une liste non exhaustive :

- *Gab*, RSV Américain lancé en 2016 qui reprend en partie le fonctionnement de Twitter
- *4chan*, RSV Anglophone lancé en 2003 qui correspond à un réseau d'échange d'images, de discussions, de dessins et de liens
- *8chan* ou *8kun* présentement, RSV Anglophone lancé en 2013 qui reprend en partie le fonctionnement de 4chan
- *Voat*, RSV Américain lancé en 2014 et fermé en 2020 qui reprenait en partie le fonctionnement de Reddit

- *Palrer*, RSV Américain lancé en 2018 qui reprend en partie le fonctionnement de Twitter
- *DLive*, RSV Américain lancé en 2018 qui reprend en partie le fonctionnement de Twitch
- *Odysee*, RSV Américain lancé en 2020 qui reprend en partie le fonctionnement de YouTube

La résultante de ces agissements, montre l'importance de l'existence d'une modération des échanges menées par des communautés massives regroupées en une même plateforme.

5.1 La structure d'une modération

Pour répondre à ce besoin de modération, les plateformes utilisent une variété d'outils pour appliquer les politiques de contenu, en s'occupant de la suppression des contenus et des comptes utilisateurs à l'origine. On répertorie 3 grandes parties, décidant de la forme de modération qu'appliquera une plateforme, avec le choix des acteurs de la modération, de la politique de contenu et du cycle de vie des contenus.

5.1.1 Approches de la MDC

Généralement, les plateformes se tournent vers 3 approches de la MDC, la première représente la **modération manuelle**, qui va compter sur une modération humaine pour examiner les contenus et prendre les décisions. La seconde correspond à la **modération automatique**, qui va consister à l'utilisation d'IA[21] pour détecter et filtrer de manière automatisée pour prendre des décisions telles que la suppression ou le signalement. La troisième étant la **modération hybride** qui fait appel aux deux modérations précédentes. La modération automatique vient compléter la modération manuelle, en signalant et hiérarchisant les contenus pour faciliter le travail réalisé par les modérateurs humains.

5.1.2 La politique de contenu

La politique de contenu des plateformes se voit conduite de manière **centralisée**, modéré par la plateforme elle-même, elle établit une politique de contenu étendue mondialement, avec

[21] L'intelligence artificielle repose sur des algorithmes et concepts mathématique avec pour objectif de répliquer l'intelligence humaine.

des cas particuliers pour se conformer aux lois des États (Facebook, Twitter, YouTube). Ou d'une façon décentralisée, modéré à la fois par la plateforme, mais principalement par les utilisateurs eux-mêmes. Elle instaure également une politique de contenu commune, en laissant la possibilité aux utilisateurs modérateurs de compléter cette politique avec l'ajout de règles de modération adaptées à leur communauté (Reddit, Twitch).

5.1.3 Cycle de vie des contenus

À propos du cycle de vie des contenus, on peut pratiquer de la **pré-modération**, qui entraine le contrôle du contenu avant qu'il figure dans le fil d'actualité de la plateforme, de la **post-modération**, avec la vérification du contenu après sa mise en ligne, la **modération réactive,** employant les utilisateurs pour signaler les contenus inappropriés qui remontent aux modérateurs et la **modération distribuée**, impliquant également les utilisateurs pour noter les contenus ce qui valide ou non la légitimité de celui-ci sur la plateforme.

5.1.4 Le droit de contestation des contenus modérés

Le droit de regard des utilisateurs sur les contenus modérés diffère selon le choix des plateformes. Lorsqu'un contenu est considéré comme nuisible à la plateforme, le modérateur prend une décision, de le déréférencer[22], le supprimer ou le bloquer temporairement. Dès lors que la modération est effectuée, le propriétaire du contenu peut ne pas être prévenu, c'est ce que l'on nomme la **modération secrète**. Dans un autre cas, avec la **modération transparente**, il peut être informé du contenu modéré et des raisons. Dans la position de modération transparente, la plateforme pourra décider que la décision est **non contestable**, ou au contraire **contestable**, mettant alors à disposition des voies de recours.

5.2 La modération manuelle

Ces plateformes ont pris la décision de privilégier l'approche manuelle pour la prise de décision, qui assure une qualité de la modération avec l'analyse humaine, mais de surcroit présente des avantages juridiques et financiers. L'analyse des contenus se pratique de deux manières, **internalisée**, usée pour les contenus davantage complexes, avec des volumes peu

[22] Le déréférencement, invisibilisation ou shadowban correspond à une action de modération, masquant sur un RSV ou un moteur de recherche un contenu considéré comme sensible/nuisible aux autres utilisateurs.

importants, nécessitant la présence de modérateurs humains familiers à la culture web et celle de la région modérée. Puis le choix largement privilégié, **l'externalisation**, utilisée pour les contenus génériques avec des volumes importants, n'exigeant pas de connaissances particulières.

Pour ce qui de **l'externalisation**, une pluralité d'entreprises spécialisées dans ce domaine propose de sous-traiter le service de MDC aux géants du web. Celles-ci, traite en majeure partie la MDC dans des pays émergents, profitant de la main d'œuvre à bas prix, comme en Inde ou aux Philippines devenu capitale mondiale des calls centers. Tout le processus de MDC externalisée écarte pour les géants du web les problèmes juridiques liées aux erreurs de modération, mais également offre un intérêt économique leur évitant le coût de création d'un pôle de modération dans un pays développé (main d'œuvre plus cher, création d'infrastructure dédié, psychologue, etc.). On peut citer les entreprises tel que **TaskUs**, **Micro Sourcing**, **CPL Ressources** ou **Competence Call Center** filiale de Telus International.

5.2.1 Les reproches à la modération manuelle

Une pression et des reproches grandissantes afflux des utilisateurs et des États pour que les géants du web prennent plus en considération l'importance du rôle de la MDC sur leurs plateformes. Dernièrement, ils ont commencé à réagir en prenant de vraies positions comme vu précédemment, mais de cela se rajoute un autre souci, l'opacité autour des processus de MDC. Selon Sarah T. ROBERTS enseignante américaine spécialisée dans les médias sociaux et la MDC, le choix de l'invisibilisation de ces processus a pour objectif de renforcer l'idée des utilisateurs que la technologie est quelque chose de magique, que l'implémentation des CGU s'applique automatiquement et cela révèle également, une intention de supprimer les traces humaines présentes derrière la MDC. La volonté derrière cette décision est de cacher l'amas d'incidents et d'affaires suspectes autour des algorithmes, mais surtout des modérateurs humains que l'on nomme les « Eboueurs du web ».

Ces « Eboueurs du web » réalisent le sale boulot en se confrontant quotidiennement aux pires atrocités de l'Humanité, affectant naturellement leur état mental. Leur parole se libère progressivement depuis 2010 grâce aux recherches de Sarah T. ROBERTS et d'autres journalistes, recensant des témoignages[30] pour éclaircir le flou autour de ce processus. En les lisant, on y comprend l'ampleur des dégâts psychologiques dû à la modération, ainsi qu'aux

conditions de travail imposées par les géants du web eux-mêmes ou les entreprises de sous-traitance. Cette charge mentale s'alourdit avec les conditions instaurées par les entreprises (contrats de confidentialité stipulant l'interdiction de divulguer ce qu'ils voient au quotidien, des temps de pauses chronométrés, interdiction de communiquer avec ses collègues, des soutiens psychologiques faible, des salaires de misère par rapport à ce qu'ils subissent, etc.). L'accumulation des évènements négatifs et d'une condition de travail anxiogène entrainent une instabilité professionnelle (change pour la plupart rapidement de travail) et des dégâts psychologiques et physiques lourds tels que la dépression, l'addiction à la drogue ou l'alcool, tentatives de suicide, insensibilisation à l'horreur, paranoïa, etc. La résultante est que l'on se retrouve dans une situation ou les personnes avec la responsabilité de modérer les contenus sont ceux qui sont les plus visés par la cyberhaine.

Enfin, la modération manuelle est critiquée pour les biais que comporte l'analyse humaine. Le manque de diversité au sein des équipes et l'étude insuffisante des angles morts (par exemple le mot « Lesbienne » considéré principalement comme un mot clé du domaine de la pornographie, influence la suppression de contenu comportant ce mot) associés aux stéréotypes que l'on retrouve dans notre société augmente le biais décisionnel. Si aucun travail n'est mené sur ses angles morts, on constate la reproduction des stéréotypes dans les choix et processus de modération. On visualise la même problématique dans les équipes de développeurs qui établissent les algorithmes.

5.3 Les outils automatiques

Les outils automatiques pour la modération sont nombreux, venant agir sur les formes de contenus possibles (texte, image, vidéo, audio, métadonnées) à des niveaux de complexité différents. Ces niveaux de complexités passent de l'algorithme simple qui recherche les mots-clés interdits pour en supprimer les commentaires, jusqu'à des techniques complexes de compréhension des sentiments dans un texte qui se base sur l'IA. Les algorithmes simples sont implémentés pour des tâches fixes et répétitives, ne faisant appel à aucune forme d'intelligence, employé dans une approche de modération automatisée occupant le rôle de vérificateur et décideur de la sentence. La création et l'implantation de ces outils automatiques ont un coût, ils varient selon les technologies et niveaux de complexité souhaités (voir annexe 2). En ce qui concerne l'application de l'IA dans la modération, elle se voit rarement tournée vers une approche de modération automatique, mais plutôt vers une

approche hybride combinant la machine et l'Homme pour maximiser l'efficacité des traitements et la qualité des décisions.

Durant le processus d'une approche de modération hybride, les outils de modérations automatiques sous IA vont venir récupérer le contenu, l'analyser et le classer par niveau de nuisance. Le contenu pourra être classé comme nuisible, provoquant sa suppression définitive, comme légitime, le laissant actif sur la plateforme, ou comme potentiellement nuisible, transférant la responsabilité de son analyse aux modérateurs humains. Pour rappel, les utilisateurs ont également la possibilité de signaler les contenus potentiellement nuisibles aux modérateurs humains. Lors des erreurs de jugements de l'IA et des décisions prises par les modérateurs humains, les résultats seront renvoyés à celle-ci pour renforcer son modèle, ajustant son discernement des contenus.

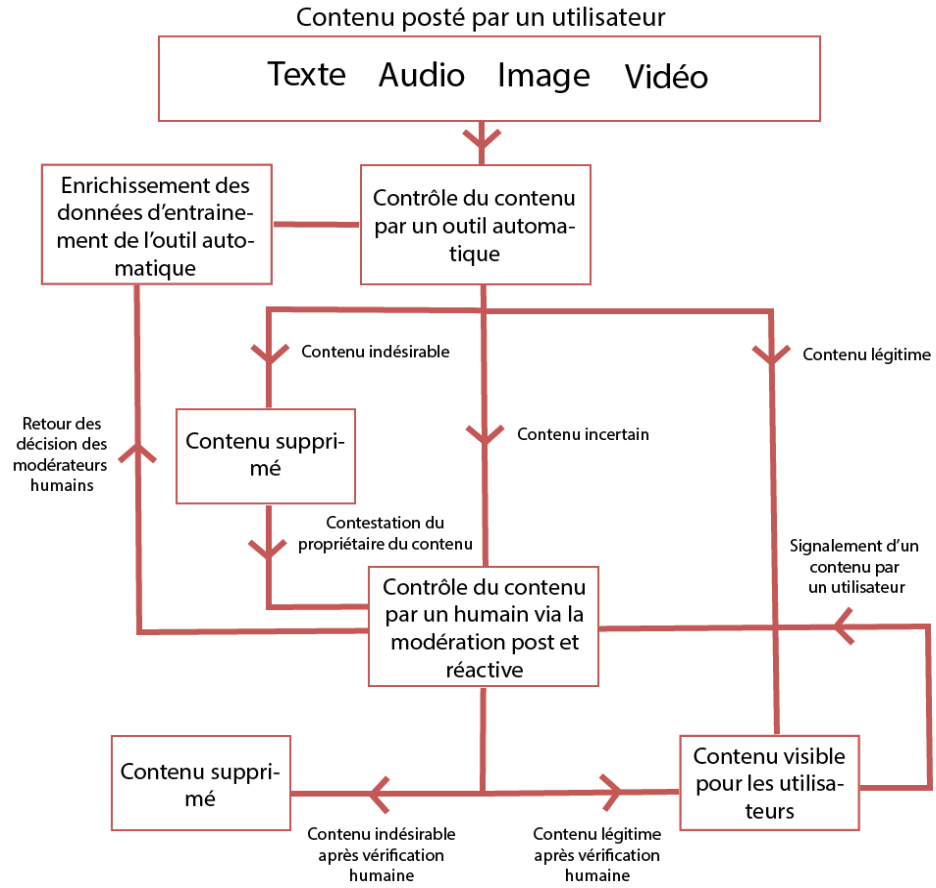

Figure 9 : Processus de la modération hybride

5.3.1 Derrière le fonctionnement des outils automatiques

Intéressons-nous au spectre de compétence que possède l'IA actuellement et les processus derrière ces outils automatiques. La reproduction d'une certaine intelligence humaine est l'objectif recherché à travers l'IA. Néanmoins, on constate encore une intelligence rudimentaire, qui se spécialise dans des tâches précises et des domaines spécifiques, manifestant d'aucune conscience, ni prise de décision identique à celle d'un humain. C'est pour cela qu'on leurs attribuent encore le statut d'IA faibles. Toutefois, bien qu'elles soient considérées comme faible, nous ne pouvons plus nous en priver dans les différentes tâches menées lors de notre quotidien, elles sont partout pour proposer un appui à l'Homme. Les tâches traitées par l'IA se répartissent en plusieurs sous-branches d'application. On identifie dans un premier temps, la **vision par ordinateur**, qui propose d'une part la vision artificielle ou machine, utile dans le secteur industriel pour le contrôle de la forme et la dimension des pièces dans les chaînes d'assemblage, et d'autre part, la fonctionnalité de reconnaissance d'image, que l'on aperçoit dans la modération, pour détecter les images violentes. Seconde branche, **le langage parlé**, qui affiche des possibilités de conversion d'un texte vers du langage (text to speech) oral ou inversement (speech to text). Troisième branche qui continue l'exploration du domaine de la langue, avec le **traitement naturel du langage**, qui donne la capacité aux machines de traduire des textes dans des langues différentes ou d'analyser les sentiments exprimés dans les textes (sentiment analysis). Quatrième branche, les **systèmes experts** qui vont aider à la décision dans des domaines spécifiques, comme celui de la santé par exemple, avec des systèmes pour appuyer les médecins à diagnostiquer certaines maladies. Cinquième branche, la **robotique** qui comprend toutes les entités mécaniques automatisées, allant des robots industriels jusqu'aux voitures autonomes. Sixième branche, la **planification**, optimisant l'organisation de divers projets tout en tenant compte des contraintes, précieux pour la logistique.

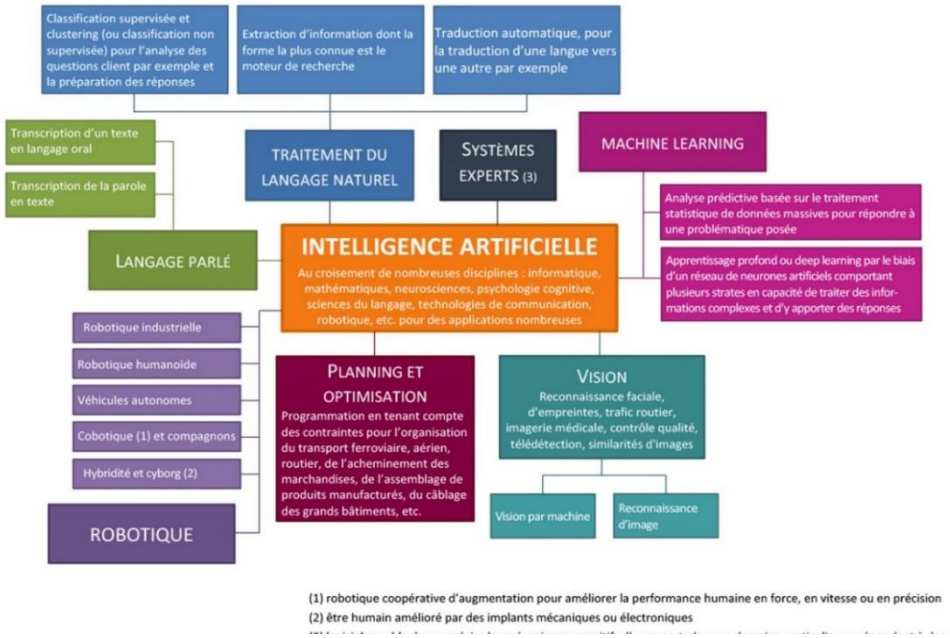

Figure 10 : Différentes branches de l'IA (source : Université Grenoble Alpes)

Pour obtenir une quelconque intelligence se rapprochant de celle de l'Homme, les machines doivent également passer par une forme d'apprentissage, pour en comprendre les missions demandées. Une dernière sous-branche de l'IA et pas des moindres vient assurer cet apprentissage, c'est le **ML (apprentissage automatique/statistique).** Cette même sous-branche, inclut des procédures d'apprentissages multiples (supervisé, non supervisé, semi-supervisé, par renforcement, profondeur) et des techniques variées (modèles mathématiques et réseaux neuronaux) que nous allons décrire par la suite.

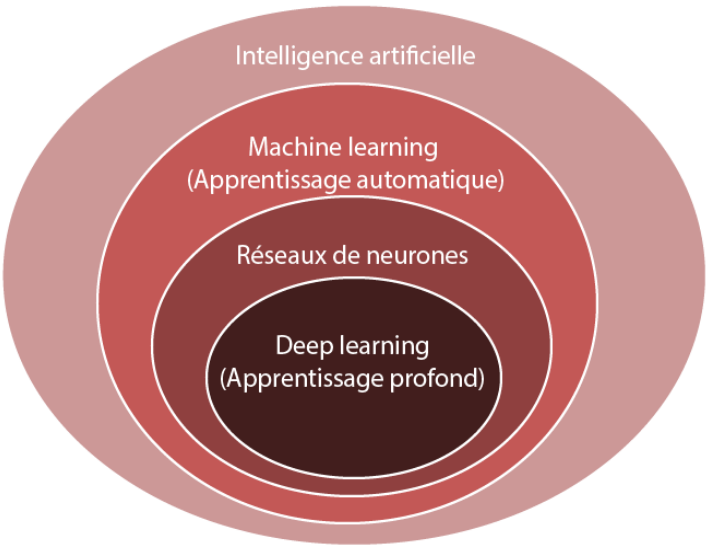

Figure 11 : Imbrication des notions de l'IA

5.3.1.1 Le Machine Learning

Le ML est l'approche d'apprentissage la plus ancienne et la plus maitrisée, elle a vu le jour dans les années 50 par le mathématicien Britannique Alan TURING qui en a établi les bases. Son premier usage était en 1959 avec l'invention du premier jeu de dames « intelligent » par Arthur Samuel[23]. L'inventeur de ce jeu de dames « intelligent » a défini le ML comme le don de la capacité d'apprendre à une machine sans la programmer de façon explicite. Son fonctionnement se fonde sur l'application de modèles mathématiques en suivant des processus d'apprentissages différents.

a. Apprentissage supervisé

L'apprentissage supervisé est le protocole basique, utile pour la prédiction de données (Y), qui demande obligatoirement l'appui de l'Homme pour étiqueter les données[24] (f(x)) d'entrainement et préciser les variables souhaitées que le modèle doit évaluer pour les corrélations (Y = f(x), prédire Y avec les données étiquetées f(x) à travers l'association/corrélation entre f(x) et Y). C'est un système monocouche qui a besoin d'un grand ensemble de données et qui marche avec une donnée étiquetée en entrée, suivi par le traitement de celle-ci par le modèle, puis l'apparition en sortie de la variable prédite. Le développement de modèles prédictifs s'emploie dans tous types de domaines tels que l'immobilier, la bourse, santé, etc. Sur le fonctionnement, on peut comparer de manière simplifiée ce processus d'apprentissage avec celui de l'école, ou un professeur de langue écrit au tableau que le mot « fish » = « poisson », en tant qu'élève, on a les données étiquetées et la variable à acquérir. Le mot « poisson » se voit être la traduction française, « fish » correspond à la variable cible que l'élève doit assimiler et la corrélation est simple, vu que c'est seulement la traduction d'un mot.

L'avantage de cette approche d'apprentissage se voit être la précision et la fiabilité des résultats. Néanmoins, la plupart des projets de nos jours requièrent un volume de données

[23] Arthur Samuel est un professeur et chercheur sur l'IA, c'est le pionnier de l'IA et de l'apprentissage automatique avec la création de son jeu de dames « intelligent » sur ordinateur en 1952.

[24] L'étiquetage de données ou data labelling correspond à l'action d'annoter la variable ciblée, la variable que l'on veut que la machine prédise. C'est une notion utilisée dans l'apprentissage supervisé, mais également d'autres approches.

conséquents, ce qui implique un coût de main d'œuvre et un temps élevé pour l'étiquetage des données.

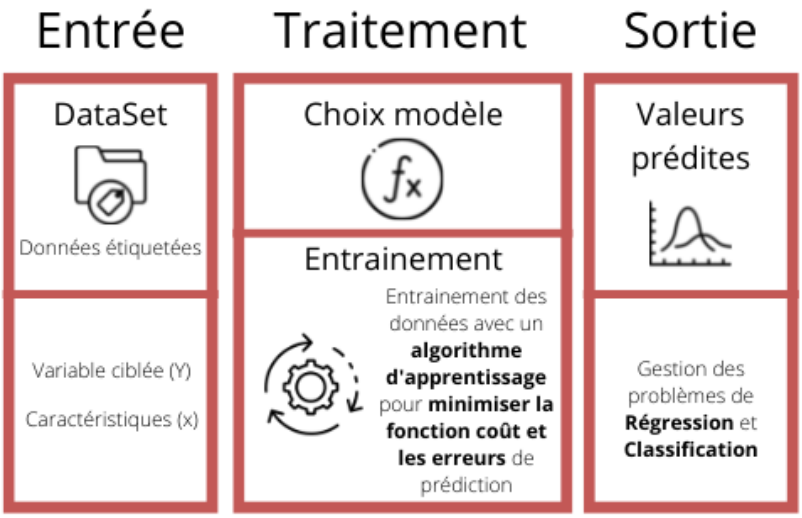

Figure 12 : Système de l'apprentissage supervisé

Deux types de problèmes sont gérés avec cet apprentissage, la **régression** qui consiste à prédire une valeur continue/quantitative (peut prendre une infinité de valeurs) en sortie, que l'on aperçoit dans la prédiction du prix de biens immobiliers. Puis la **classification**, qui à l'inverse procède à la prédiction d'une valeur discrète/qualitative (peut prendre un nombre limité de valeurs) en sortie, que l'on rencontre dans notre quotidien avec le classement des messages sur les applications de discussion instantanée dans les différentes catégories (Exemple avec Messanger la discussion instantanée de Meta : « Vous connaissez peut-être » ou « Spam » ou « Légitime »).

<u>Cas de régression :</u>

Pour expliquer la conduite d'un problème de régression, nous allons rester sur le domaine de l'immobilier.

1. Dans la première étape, on rentre nos données pour la machine dans un tableau nommé « DataSet », qui se compose de deux types de données, la variable ciblée (target variable = Y), qui convient à ce que l'on veut que le modèle apprenne et les caractéristiques (features extraction = x), qui permettent au modèle de prédire la variable ciblée en menant des corrélations.

Variable ciblée (Y)	Caractéristiques (f(x))			
Y	x1	x2	x3	x4

Prix	Adresse postale	Date de construction	Surface (m2)	Orientation
150 k	64400	1980	60	Ouest
350 k	33600	2005	45	Sud
250 k	64500	1995	50	Sud

2. La seconde étape se poursuit par le choix de l'algorithme. Il doit se faire par rapport aux types des données auquel il sera confronté. Lorsque qu'il a été sélectionné, il faut compléter les paramètres de sa fonction mathématique, généralement on débute par donner des paramètres aléatoires. Les étapes suivantes vont fournir des précisions sur les valeurs optimales à attribuer aux paramètres de la fonction, pour obtenir une justesse accrue sur les prédictions. Concernant les algorithmes les plus employés dans un modèle de régression, on peut citer la régression linéaire, l'arbre de décision, la forêt d'arbres décisionnels, etc.

Application d'un modèle linéaire dans l'exemple ci-dessous, avec la fonction affine **f(x) = ax + b**

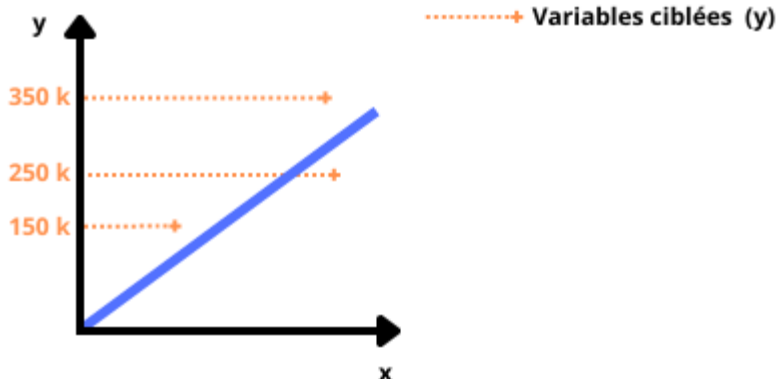

Figure 13 : Choix du modèle pour l'apprentissage supervisé | Régression

3. La troisième étape va servir à représenter l'ensemble des erreurs présentes sur les premières prédictions du modèle, en comparaison avec les variables ciblées (Y) présent dans le dataset. De cela, on analyse la distance des erreurs pour chaque prédiction (trait rouge sur la figure ci-dessous), puis on assemble et modélise toutes ces erreurs la forme d'une fonction nommée « Coût », pour que l'on puisse l'utiliser pour améliorer le modèle par la suite.

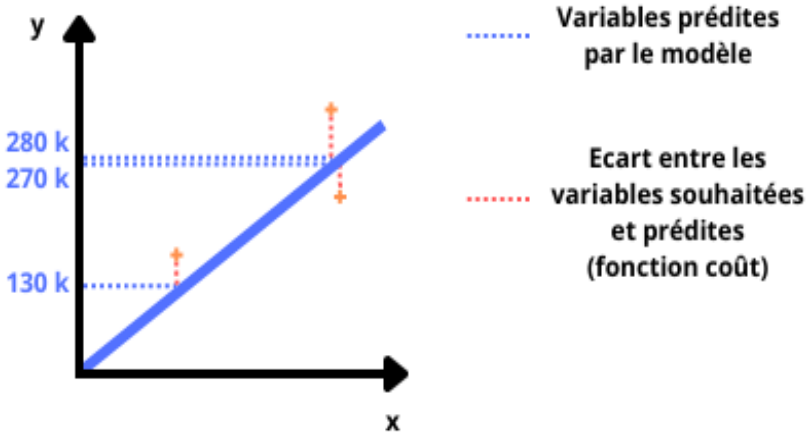

Figure 14 : Analyse de l'écart des erreurs pour la création de la fonction coût | Régression

4. Dernière étape, qui concorde à la partie dite d'apprentissage, ou l'on veut minimiser l'ensemble des erreurs obtenues, modélisées sous la forme de la fonction Coût. En réponse à ce besoin, des algorithmes de minimisation et d'apprentissage existent pour trouver les paramètres adéquats au modèle, procurant une diminution de la fonction Coût et des erreurs de prédiction du modèle. Pour les modèles linéaires, on peut citer l'algorithme de minimisation de la Descente de gradient ou la méthode des moindres carrés. Pour les arbres de décision, on trouve l'algorithme de CART.

<u>Cas de classification :</u>

Pour exposer le déroulement d'un cas de classification, nous allons rester sur la catégorisation des messages reçus sur l'application de discussion instantanée Messanger. Les étapes de la classification sont quasiment identiques à celle de la régression, seulement un changement intervient à l'étape de la prédiction des données par le modèle.

1. Mise en place du dataset avec l'étiquetage des données.

0 -> SPAM ; 1 -> Vous connaissez peut-être ; 2 -> Légitime

Variable ciblée (Y)	Caractéristiques (f(x))			
Y	x1	x2	x3	x4

Type Message	**Première discussion**	**Groupe de discussion**	**Présence du contact dans les amis**	**Nombre de lien dans le message**
0	Oui	Oui	Non	1
2	Non	Non	Oui	0
2	Non	Non	Oui	1

2. Sélection de l'algorithme approprié aux données que l'on traite et intégration de paramètres aléatoire pour le premier jet de prédiction des données. Sur les premières prédictions, l'algorithme place une frontière de décision, qui délimite le classement des prédictions en comparaison aux variables ciblées. D'un côté se trouve les messages considérés comme des spams et de l'autre les légitimes. Les algorithmes les plus populaires au sein du modèle de classification sont la régression logistique, le support vecteur machine, Naive Bayes, l'arbre de décision, la forêt d'arbres décisionnels, etc.

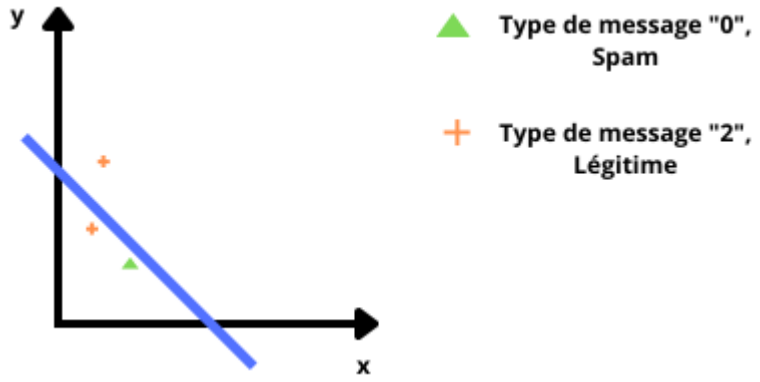

Figure 15 : Choix du modèle pour l'apprentissage supervisé | Classification

3. Déduction de la distance des erreurs et de la fonction Coût.
4. Mise en exécution d'un algorithme de minimisation pour réduire la fonction Coût et les erreurs de prédiction.

b. Apprentissage non supervisé

À l'inverse de l'apprentissage supervisé, l'apprentissage non supervisé s'opère de manière indépendante, avec des données non étiquetées (f(x)) par l'Homme et avec aucun exemple de résultats (Pas de variable ciblée, Y). La machine va parcourir l'ensemble des données non étiquetées pour chercher des relations significatives. Son objectif n'est pas de prédire, mais de trouver du sens et des liens sur des données larges et évasives pour l'Homme. C'est un système qui demande également un grand ensemble de données, il va analyser les données, les attribuer à des catégories, puis les trier pour finalement améliorer leur compréhension. On remarque son emploi dans les domaines du marketing ou du web pour segmenter les communautés donnant la possibilité de proposer des pubs ou produits adaptées, mais également dans le champ de la reconnaissance vocale et de la santé. Pour donner un exemple concret, si un individu décide d'apprendre une langue en rejoignant un pays sans aucune base ni ressources pour l'aider, cela correspond au processus de l'apprentissage non supervisé.

L'avantage de cette approche d'apprentissage est la rapidité d'exécution et l'amputation du coût de main d'œuvre pour étiqueter les données. Toutefois, la précision des résultats est inférieure à l'apprentissage supervisé.

Une variante existe pour entrainer les modèles, alliant l'approche supervisée et non supervisée, elle porte le nom d'apprentissage semi-supervisé. Selon les circonstances, l'objectif et la composition du projet, l'ordre d'utilisation des approches d'apprentissage seront différents durant les cycles de ce dernier. Par exemple, l'apprentissage supervisé peut accroitre la précision de l'apprentissage non supervisé, en procédant au préalable à l'étiquetage d'une partie du jeu de données. En intervertissant l'ordre, l'apprentissage non supervisé vient préparer le terrain en éclaircissant les données (mise en lumière de nouvelles métriques), ce qui offre une meilleure compréhension pour ajuster la phase d'étiquetage des données de l'apprentissage supervisé.

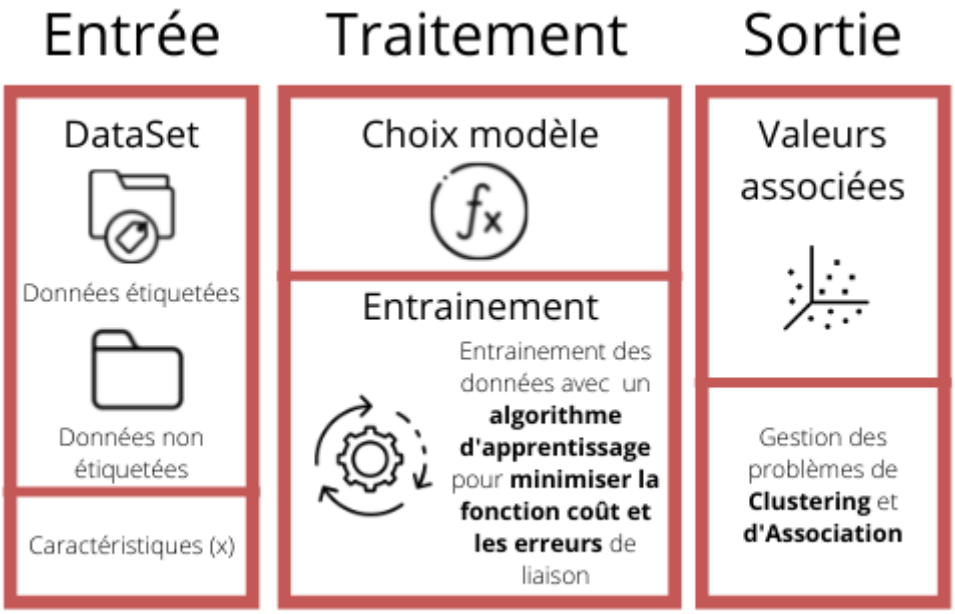

Figure 16 : Système de l'apprentissage non supervisé

Pour ce qui est de son usage, on l'affecte aux problèmes de **clustering** et **d'association**. Le clustering se rapporte à une méthode regroupant des points de données par rapport à des conditions de similarité ou de distance. L'association va pareillement mener une classification de données dans des groupes, mais va analyser les règles et degrés d'association entre les données, ce qui permet de classer plus largement des ensembles de données. Outre ces deux usages principaux de classement des données, l'apprentissage non supervisé est aussi employé pour fluidifier le traitement de grands volumes de données avec la réduction dimensionnelle (algorithme PCA, UMAP, TSNE, etc.) et pour calculer approximativement la densité de distribution.

<u>Cas de clustering :</u>

Nous allons prendre un cas réel autour du domaine du sport, pour monter un cluster des joueurs avec des caractéristiques identiques, déployé lors du recrutement des joueurs.

1. Mise en forme du dataset avec seulement les données de caractéristiques (f(x)), car l'objectif n'est pas de prédire, mais de regrouper des joueurs aux profils identiques.

Caractéristiques (f(x))					
x1	x2	x3	x4	x5	x6

Nom Joueur	Nombre de match	But / match	Passe décisive / match	Kilomètre couru / match	Hors-jeu / match
Werner	44	0.39	0.16	5	3
Kolo Muani	39	0.33	0.13	6	5
Gameiro	36	0.31	0.08	4	3
Lukaku	46	0.41	0.02	4	2
Giroud	39	0.36	0.08	3	1
Aubameyang	43	0.53	0.09	7	4
Delort	40	0.43	0.12	5	5
Moreno	29	0.45	0.24	5	4
Benzema	51	0.94	0.29	7	2
Haaland	35	1.03	0.29	6	3

2. Dans le cas d'un grand ensemble de données, il est possible d'effectuer une réduction dimensionnelle, pour atténuer le temps de traitement computationnel. Par la suite, survient l'étape de décision de l'algorithme du modèle de clustering le plus en adéquation avec nos données. Ce modèle de clustering comprend les algorithmes du k-means, de décalage moyen, de classification hiérarchique, etc. Pour notre cas expérimental, nous avons utilisé l'algorithme du k-means[25] ou la demande était de regrouper les joueurs similaires dans 4 clusters distincts.

[25] Le jeu de données a été traité avec une version de l'algorithme K-means en ligne (voir annexe 3).

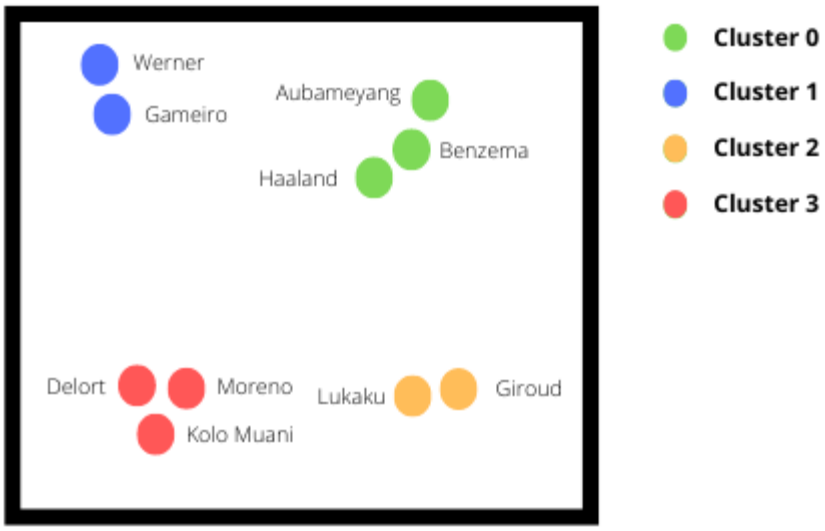

Figure 17 : Modélisation d'un cluster lors de l'apprentissage non supervisé | Clustering

Cas d'association :

Les règles d'association s'exploitent en majorité dans le cadre du comportement des utilisateurs, pour comprendre les habitudes et adapter des services ou produits (affichage de publicité, bon de réduction, etc.). De ce fait, nous allons prendre un exemple classique, celui du panier d'achat des ménages.

1. Création du dataset avec les listes (transaction) des produits (item) achetés par les clients d'un magasin alimentaire.

Caractéristiques (f(x))					
x1	x2	x3	x4	x5	x6

Identifiant	Liste de courses
Client 1	Lait, Pain
Client 2	Pâte à tartiner, Lait, Fromage, Bière, Sel, Pain
Client 3	Pâte, Fromage, Pâte à tartiner, Légume, Pain, Poivre
Client 4	Fromage, Pâte

Client 5	Bière, Pain
Client 6	Bière, Pâte, Pain
Client 7	Eau, Fromage, Pâte, Sel, Pain, Poivre
Client 8	Pâte, Légume, Pâte à tartiner, Fromage, Pain, Sel, Lait
Client 9	Eau, Fromage, Sel
Client 10	Bière, Pâte à tartiner, Sel, Pain

2. Comme lors du clustering il est possible de réduire la dimension pour faciliter les traitements. Le modèle d'association laisse le choix de plusieurs algorithmes tel qu'apriori, close, ocd, partition, dynamic itemset counting (DIC), etc. Pour notre jeu de données qui comporte des listes de courses, nous avons appliqué l'algorithme apriori avec un seuil de soutien[26] à 40 % et un intervalle de confiance[27] à 70 %. Ci-dessous[28], on aperçoit les combinaisons qui anticipent l'achat d'un produit avec au moins 70 % de véracité.

[26] Le seuil de soutien présent dans l'algorithme Apriori du modèle d'association correspond au nombre de fois qu'apparait une combinaison ou un produit (item) dans la totalité du jeu de données. Dans notre exemple, les combinaisons et produits se devait d'être présent à 40 % dans le jeu de données.

[27] L'intervalle de confiance dans l'algorithme Apriori du modèle d'association correspond au nombre de fois qu'une combinaison ou produit (item) va générer l'anticipation de l'achat d'un autre produit. Dans notre exemple, l'anticipation d'achat d'un produit par une combinaison ou un produit devait être assurée 70 % du temps.

[28] Le jeu de données a été traité avec une version de l'algorithme Apriori en ligne (voir annexe 4).

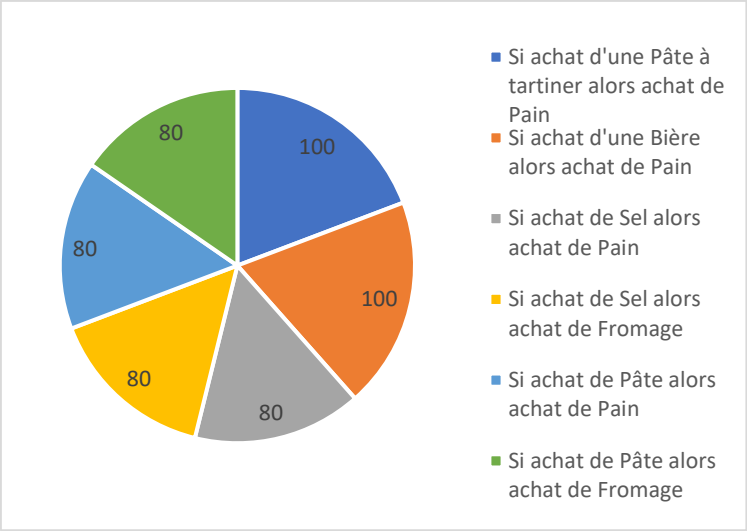

Figure 18 : Modélisation des règles d'association | Association

c. Apprentissage par renforcement

En comparaison avec les deux approches d'apprentissages précédentes, l'apprentissage par renforcement fonctionne bien différemment, en laissant la machine apprendre toute seule dans un environnement. Elle s'est inspirée de théorie de la psychologie animale, plus précisément sur le comportement qu'ils pouvaient avoir dans les décisions. Le modèle général utilisé est le « Processus de décision Markovien » (MDP), qui reprend le fonctionnement des « Chaines de Markov »[29] dans lequel une entité décisionnaire va se diriger d'état en état dans un environnement selon des probabilités. Ces chaines se sont vues améliorées par les ajouts d'action et de récompense par le mathématicien américain Richard Ernest BELLMAN.

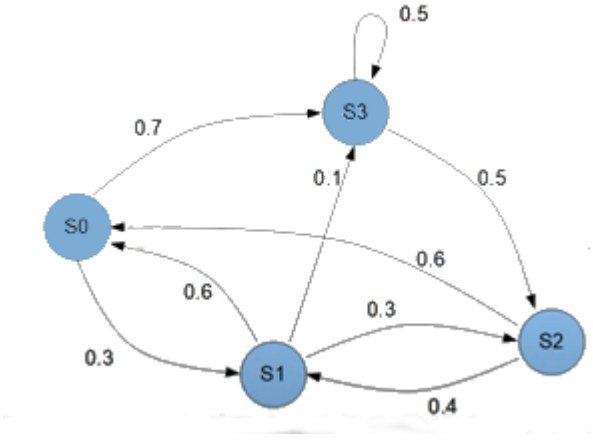

Figure 20 : Chaînes de MARKOV

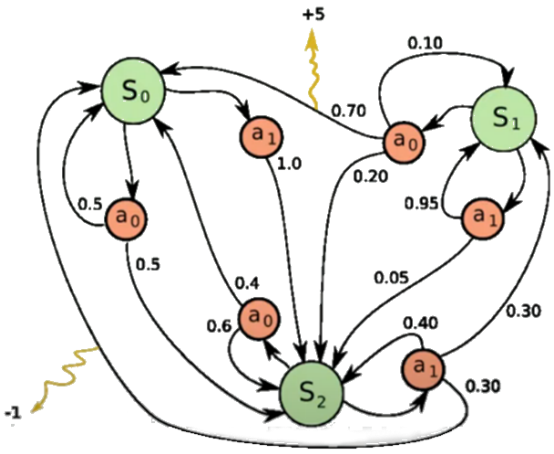

Figure 19 : Chaînes de MARKOV et équations de BELLMAN

[29] Les chaines de Markov ont été inventé par le mathématicien russe Andreï Andreïevitch MARKOV.

C’est un système Agent-Environnement ou un agent artificiel correspond à l’entité décisionnaire qui va mener des actions, le dirigeant vers divers états dans l’environnement. Selon les actions et les états atteints dans l’environnement, l’agent recevra des récompenses, bonus ou des malus par rapport à l’état souhaité. Le but principal est d’apprendre une stratégie/politique optimale à travers les essais et erreurs dans les actions et états, pour maximiser le gain de récompenses obtenues, ou à l’inverse dans les deux autres approches d'apprentissages, on cherchait à minimiser les erreurs. La décision du chemin optimal pour maximiser le gain est déterminée avec l’entrainement des algorithmes d’apprentissage par renforcement, qui diffèrent selon les caractéristiques de l’environnement (par exemple la connaissance des récompenses ou non entre les actions et états), on retrouve l’algorithme de Q-learning, différence temporelle, itératif, les réseaux neuronaux, etc. Une notion essentielle d’équilibre entre exploration et exploitation voit le jour, dès lors que l’agent apprend à se diriger dans un environnement. Suivant le cycle d’apprentissage, on demandera davantage d’exploration pour étudier tous les chemins possibles, ou d’exploitation pour se concentrer sur les chemins optimaux connus. L’algorithme d’Epsilon-Greedy offre la possibilité de contrôler cet équilibre. L’apprentissage par renforcement s’est vu largement appliqué dans les jeux vidéo, mais s’étend aujourd’hui dans les domaines de la robotique, de la santé ou la finance.

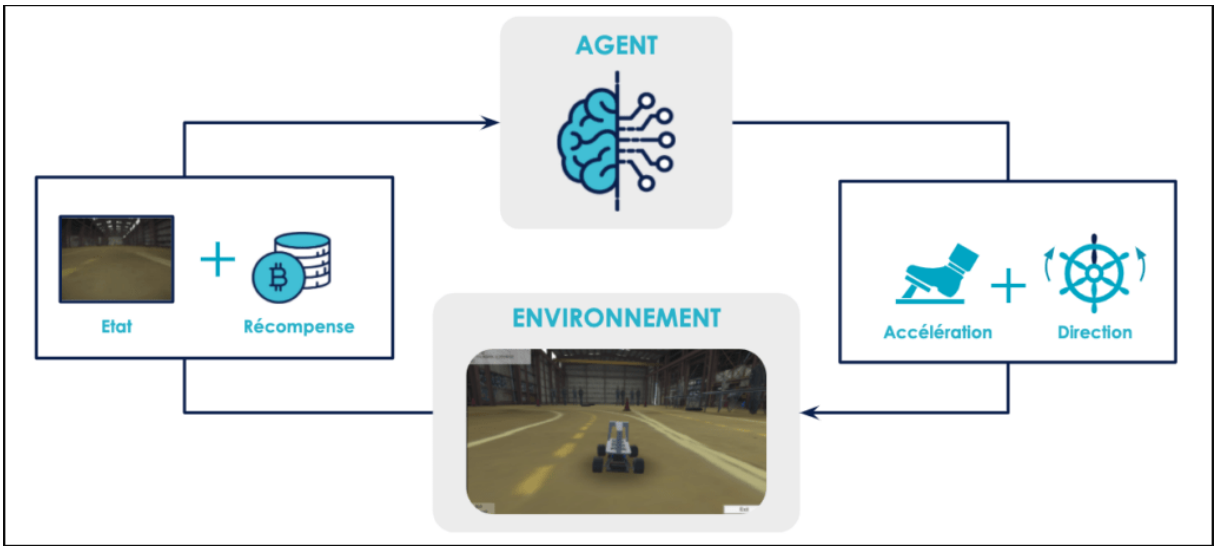

Figure 21 : Système de l'apprentissage par renforcement (source : Octo talks)

Comme exemple concret pour comprendre l'apprentissage par renforcement, on peut le transposer à la vie réelle. On va statuer que les Hommes sont les agents, qui vivent et tentent d'effectuer les meilleures décisions dans l'environnement qu'est la vie réelle. Ces choix d'actions vont mener à des conséquences ou états, positif ou négatif (la notion de récompense), ce qui résultera par l'amélioration de la compréhension de l'environnement et des actions choisies.

Cette approche a pour avantage de construire son propre jeu de données au fil du temps que l'agent s'entraine dans l'environnement, ne demandant aucune donnée de départ. Néanmoins, lors des processus d'apprentissage, l'agent prendra généralement davantage de temps pour tester tous les chemins possibles pour déceler le plus optimal.

Voici les différents concepts et métriques à considérer pour saisir le processus d'apprentissage par renforcement :

1. **L'agent**, c'est l'acteur décisionnaire des actions et états qu'il prendra dans l'environnement
2. **L'environnement**, c'est le monde ou figure l'agent, il peut être réel ou virtuel. Le cadre virtuel se montre plus flexible lors d'un apprentissage, offrant la possibilité par la suite de le transférer dans la réalité, même si ce changement reste encore complexe pour plusieurs projets.
3. **Les actions**, ce sont l'ensemble des possibilités susceptibles à être choisi par l'agent pour interagir avec l'environnement.

4. **Les états**, ce sont les caractéristiques ou paramètres qui définissent l'agent dans l'environnement, par exemple, sa position dans celui-ci ou sa taille.
5. **Une récompense**, c'est le résultat qui va indiquer la réussite ou l'échec de l'action menant l'agent vers un état. Un exemple simple, dans un jeu de voiture, si l'agent atteint la ligne d'arrivée, il aura un point et s'il tombe dans le vide, il aura un point en moins.
6. **La politique**, plan ou stratégie, c'est l'ensemble des décisions que sélectionnera l'agent pour évoluer dans l'environnement de manière optimale, en maximisant ses gains de récompenses. Pour arriver à la politique dite optimale, il y a deux possibilités de politique à suivre, soit la politique déterministe qui préconise de prendre une action précise lorsqu'on est dans tel ou tel état. Enfin, la politique stochastique, qui indique par probabilité les actions qui peuvent être choisies dans les différents états. Les actions optimales attachées aux états seront déterminées avec l'usage d'algorithmes.
7. **La fonction de valeur** correspond à la mesure de l'espérance du nombre de récompenses potentielles de l'état à l'instant t, jusqu'à l'état final ou prédéfini. Cette valeur sera importante pour l'entrainement des algorithmes d'apprentissage par renforcement.
8. **La fonction action et valeur, ou fonction Q**, représente la mesure de l'espérance du nombre de récompenses potentielles, de l'état à l'instant t et de l'action choisie, jusqu'à l'état final ou prédéfini. Cette valeur sera également primordiale pour l'entrainement des algorithmes d'apprentissage par renforcement.
9. **L'équilibre entre exploration et exploitation,** c'est une notion qui indique le comportement que l'agent doit suivre dans son entrainement. S'il doit explorer de nouveaux chemins ou exploiter, en gardant en mémoire les chemins optimaux, déjà acquis. Cet équilibre se règle avec des algorithmes, ou l'on peut indiquer le pourcentage d'exploration ou d'exploitation que l'agent doit adopter.

Maintenant que l'énumération des métriques indispensables à un processus d'apprentissage par renforcement est faite, nous pouvons alors en appliquer la majorité dans un exemple. Pour cela, nous allons nous fixer sur un cadre simple d'un robot qui doit trouver le chemin optimal pour sortir d'un labyrinthe, ce type d'exemple est fréquemment utilisé pour présenter l'apprentissage par renforcement.

Dans l'exemple ci-dessous, on retrouve une partie des métriques avec **l'agent** artificiel qui est le robot, **l'environnement** est le labyrinthe renfermant notre agent, **l'état** se rapporte à la position de l'agent, les **actions** sont les déplacements de l'agent, les **récompenses** surgissent lors des déplacements (-1 si déplacement dans un mur, -0,5 si déplacement sur une case vide et +1 si l'agent atteint la sortie) et **l'équilibre entre exploitation et exploration** qui évolue au fil de l'entrainement de l'agent. En début d'apprentissage, l'agent va avoir besoin de découvrir principalement pour se créer son jeu de données et trouver un chemin vers la sortie (attribution d'un équilibrage à 10 % d'exploitation et 90 % d'exploration). Après plusieurs tentatives, l'agent dans la seconde partie de l'exemple a repéré un chemin pour sortir, il obtient le bonus de +1 et enregistre le chemin. Cependant, la politique recherchée dite optimale, incite à continuer la découverte pour trouver un autre chemin plus efficace pour maximiser le gain de récompense, ainsi il faut de nouveau équilibrer l'exploitation et l'exploration (passage à 40 % d'exploitation et 60 % d'exploration, pour trouver un autre chemin, tout en gardant une base de connaissances du labyrinthe). Dans la dernière partie de l'exemple, l'agent a déniché un autre chemin qui se voit davantage optimal que le précédent, avec une récompense de – 8,5 (19 (nombre de cellules vides visitées) * (-0,5) + 1) pour le premier et – 6,5 (15 * (-0,5) + 1) pour celui-ci.

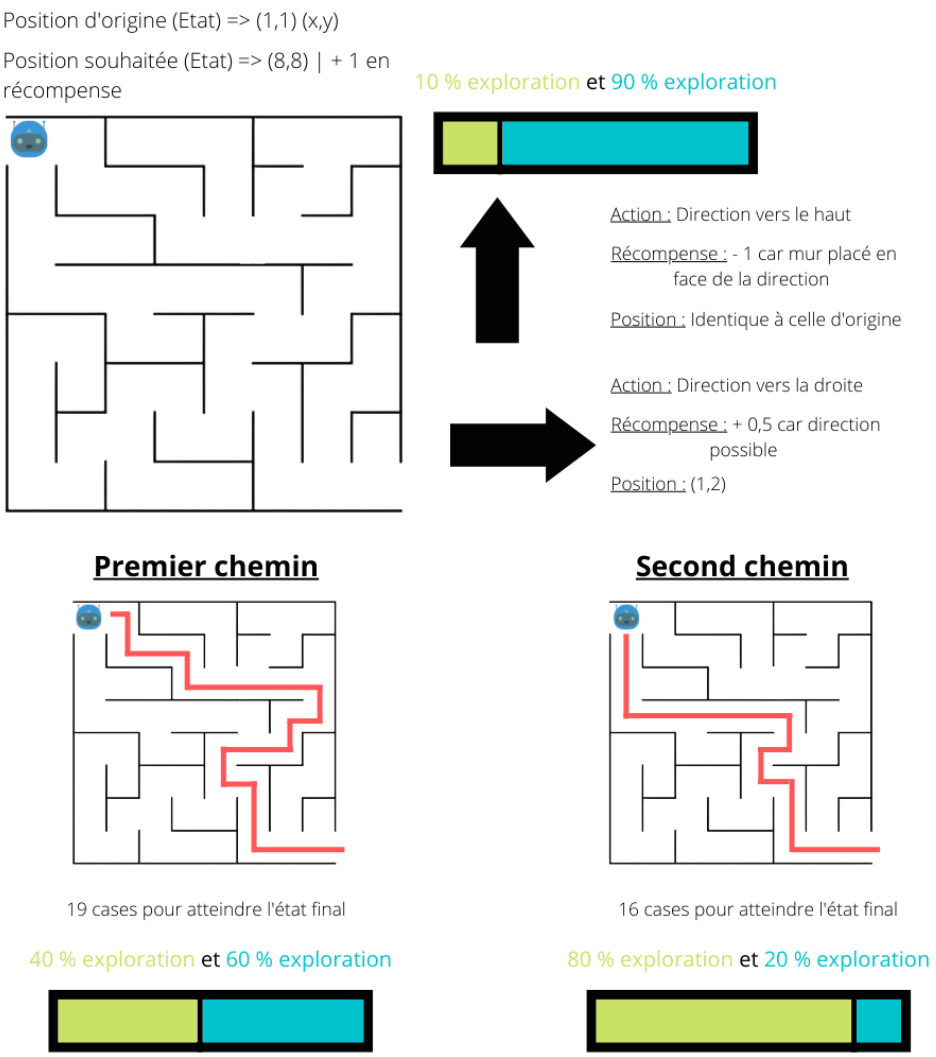

Figure 22 : Exemple de fonctionnement de l'apprentissage par renforcement

5.3.1.2 Les Réseaux de neurones

Le DL ou apprentissage profond est une composante du ML, qui offre une performance accrue pour le traitement des données et l'apprentissage des machines, en venant étendre les possibilités de tâches apprises. On le rencontre au quotidien dans les RSV, les logiciels, les voitures autonomes, etc. Le principe reste le même que les apprentissages supervisé et non supervisé, avec les démarches, que sont l'alimentation de grands ensembles de données à la machine (des volumes plus grands que pour les autres approches d'apprentissage avec un étiquetage des données), l'application d'un modèle, l'utilisation d'un algorithme d'apprentissage pour minimiser les erreurs puis la modification des paramètres du modèle pour affiner ses résultats. La différence provient dans la structure, car le DL n'utilise pas une simple fonction d'un modèle mathématique, mais un réseau de fonction connecté qui correspond au réseau de neurones artificiels/formels. Plus ces réseaux sont profonds et

contiennent des couches, plus ils contiennent de fonction et plus la machine affine sa précision et est capable d'apprendre à réaliser des tâches complexes, c'est pour cela que l'on nomme cette approche, l'apprentissage profond.

Pour comprendre convenablement le DL, il est essentiel de se pencher sur 4 dates clés qui ont amené cette approche à ce qu'elle est actuellement.

Le premier réseau de neurones artificiels[31] sur lequel se base encore le DL présentement a été inventé en **1943** par deux chercheurs, le spécialiste en neurologie Warren McCULLOCH et Walter PITTS mathématicien et psychologue cognitif. Ils se sont inspirés de la structure et l'activité du cerveau humain pour créer une abstraction du processus de fonctionnement des neurones.

À propos des neurones du cerveau humain, ce sont des cellules excitables interconnectées qui ont pour rôle de transporter l'information dans notre système nerveux, on n'en compte pas moins de 100 milliards dans le cerveau humain. Un neurone se constitue de synapses, de dendrites, d'un corps cellulaire, d'un axone et de terminaisons axonales. Des signaux de type excitateur (stimule le neurone, ce dont ils ont besoin pour déclencher une action) ou inhibiteur (équilibre la stimulation, empêchant l'activation du neurone) vont transiter du neurone précédent vers le suivant, en passant au niveau des synapses liés à la dendrite. Le neurone s'active lorsque la somme des signaux excitateur atteint un certain seuil, et va en suivant envoyer un signal électrique le long de l'axone en direction des terminaisons axonales, qui achemineront à leur tour le signal à un autre neurone.

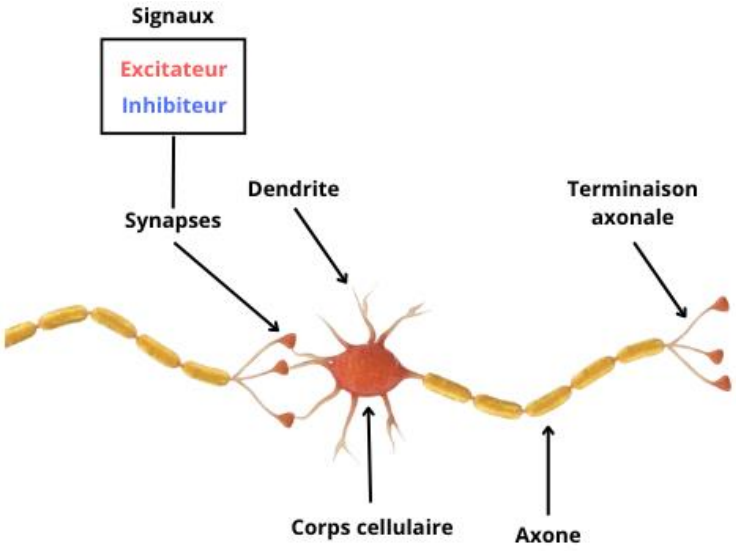

Figure 23 : Schéma d'un neurone

Dans leur modèle de réseau de neurones, les deux chercheurs sont venus reproduire ce système, en le répartissant en trois couches distinctes. La première couche se rapporte aux **entrées**, ou l'on reçoit plus des signaux, mais des valeurs. Ces valeurs vont être pondérées avec des poids et des valeurs de biais, en référence aux types de signaux provenant des synapses, pour en déduire l'utilité de la valeur. La valeur en sortie de ce calcul sera testée par une fonction d'activation (à cette époque l'algorithme adopté était heaviside ou marche), qui dans le cas d'un succès enverra la valeur dans la seconde couche, la **couche cachée** correspondant à une fonction mathématique. À la suite de l'activation, la seconde couche va transmettre la valeur vers la dernière couche, la **sortie**. Ce premier réseau de neurones artificiels se formait d'un neurone et proposait une structure forte encore employée de nos jours. Toutefois, son fonctionnement est perfectible, en ne proposant aucun algorithme d'apprentissage, laissant seul l'utilisateur pour trouver les paramètres adéquats pour affiner les résultats du réseau de neurone. Ce modèle offrait la possibilité de répondre à des problèmes logiques (binaires).

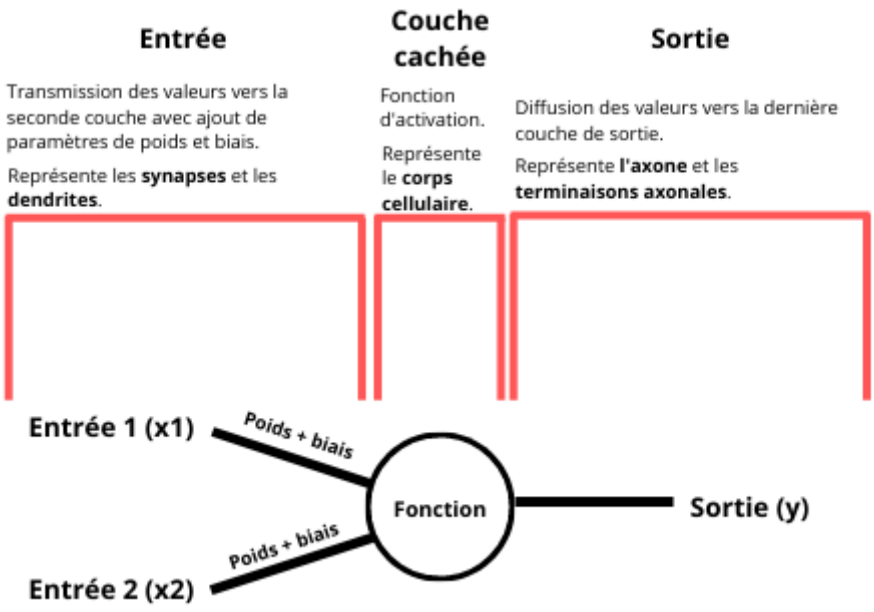

Figure 24 : Premier réseaux de neurones artificiels

En **1957**, le psychologue américain Frank ROSENBLATT a recyclé le modèle originel inventé en 1943, pour en intégrer un algorithme d'apprentissage, pour détecter par méthode mathématique les paramètres (poids et biais) optimaux à indiquer au réseau de neurones. Ce nouveau modèle de réseau de neurones porte le nom de **Perceptron monocouche**, qui dispose d'un neurone. La règle du neuropsychologue Donald HEBB et la notion de plasticité

synaptique a influencé Frank ROSENBLATT pour la construction de son algorithme d'apprentissage, en procédant au renforcement des liens synaptiques (poids des connexions) entre deux neurones quand ils sont excités/activés ensemble. Le défaut de cet algorithme d'apprentissage était son inefficacité sur un large ensemble de problèmes, s'expliquant par son statut de modèle linéaire.

Après une perte de vitesse et d'engouement autour de l'IA, le chercheur Canadien Geoffrey HINTON, spécialiste de l'IA, parvient à relancer le sujet en **1986**, avec la production du **Perceptron multicouche**. À l'inverse du Perceptron monocouche, celui-ci répond à une majorité des problèmes de par la structure se basant plus sur une, mais plusieurs neurones, ainsi qu'à l'aide de son algorithme d'apprentissage non linéaire. Ce modèle a évolué au fil des années avec la création de variantes tel que le réseau de neurones convolutifs[30] en 1990 ou le réseau de neurones récurrents[31] en 1997.

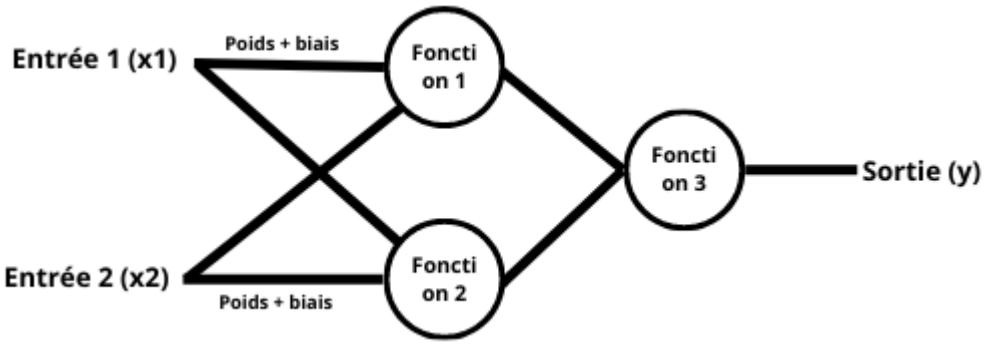

Figure 25 : Perceptron multicouche

L'instauration du DL dans notre quotidien a mis du temps, ce retard se justifie par l'impossibilité d'obtenir les variables indispensables de cette approche. Auparavant, on était en incapacité de pouvoir fournir les volumes de données importants dont la machine avait besoin, ce qui a changé avec la démocratisation du web, des smartphones et des objets connectés dans les foyers, décuplant les données. De plus, les avancées computationnelles n'offraient pas la puissance de calcul désirée pour mener convenablement du DL. En **2012**,

[30] Le réseau de neurones convolutifs est une variation du Perception multicouche, qui a adapté son fonctionnement au cortex visuel des animaux pour se spécialiser dans le traitement et la reconnaissance d'image.

[31] Le réseau de neurones récurrents est une variation du Perception multicouche, qui a arrangé son fonctionnement pour se spécialiser dans l'analyse temporelles, performant pour la reconnaissance vocale ou la lecture de textes.

lors de la compétition ImageNet mettant en concurrence des algorithmes de détection d'image, le DL a pris une tournure différente, grâce à une équipe de chercheurs dirigés par le créateur du modèle Perceptron multicouche, qui ont développé et entrainé un réseau de neurones. Leur modèle a remporté la compétition, en affichant une large performance par rapport aux autres algorithmes et approches d'apprentissage. C'est à ce moment-là que les entreprises ont pleinement eu confiance en cette approche, et l'ont appliqué dans notre société dans divers domaines et dispositifs.

À ce jour, nous appliquons le modèle du Perceptron multicouche, il se décompose en trois parties comme on a pu l'apercevoir, avec les **entrées**, les **couches cachées** (hidden layers) et le ou les **sorties**. Son processus de fonctionnement reprend quasiment le même principe que l'apprentissage supervisé et non supervisé. Les étapes ci-dessous sont répétées autant de fois qu'il le faut pour acquérir le résultat recherché.

1. Dans un premier temps, le modèle de réseau de neurones doit être alimenté en données et recevoir des paramètres aléatoires dans la première itération. Ces paramètres sont les poids de connexion entre les liaisons et les valeurs de biais. Enfin, les données vont circuler par la couche d'entrée, en passant par les fonctions d'activation (Actuellement, on manipule principalement l'algorithme de logistique, ReLU, TanH ou Softmax), puis en rejoignant la dernière couche pour produire une valeur de sortie. Cette étape se nomme le **feed forward ou forward propagation**.
2. Ce suit dans une seconde partie l'analyse de l'erreur entre la valeur en sortie et le résultat souhaité. La proportion de l'erreur sera traduite sous la forme d'une **fonction coût.**
3. Durant cette troisième partie, l'objectif est de mesurer la variation de la fonction coût à chaque paramètre des couches de notre modèle de réseau de neurones. La variation s'exécute avec une chaine de gradient, en calculant l'écart de la dernière couche à la première. Ce processus qui entreprend le chemin inverse porte le nom de **back propagation**.
4. C'est à ce moment-là qu'intervient **l'algorithme d'apprentissage de la descente de gradient**, pour procéder à la minimisation de la fonction coût grâce aux variations des erreurs inter couche retrouvées avec la chaine de gradient. Il nous indique les

paramètres les plus adéquats à attribuer aux couches du modèle, ce qui viendra affiner les résultats en sortie et minimiser la fonction coût.

Comme on a pu l'apercevoir, le DL à la possibilité de mener la plupart des tâches, cependant, il peut s'avérer plus complexe à mettre en place. Lors d'un projet qui demande du ML, il faut encore réfléchir vers quelles approches se tourner, en choisissant la plus adaptée à notre projet (le volume de données que l'on peut fournir, la durée du projet, le problème à résoudre, etc.).

5.3.2 Les problèmes

Au cours du paragraphe sur la modération manuelle, nous avons constaté que les Hommes étaient biaisés. Ces biais s'étendent logiquement aux IA, confectionnées et structurées elles aussi par des Hommes. Ce problème porte le nom de biais algorithmique, qui contrairement à l'Homme est plus complexe à comprendre, car son fonctionnement entre l'entrée et la sortie des données est opaque, comme une boite noire. Il existe deux sources expliquant l'apparition de ce biais, la première se situe au sein du groupe de programmeur, avec les biais de développement puis ceux liés aux biais cognitifs du programmeur lui-même. La seconde source se trouve à l'apprentissage des modèles, qui fait référence aux données présentées de façon inégale à celui-ci. En somme, ce biais algorithmique retranscrit plus ou moins les stéréotypes de la société, rappelant le problème des angles morts. Nous avons un exemple clair en 2016 qui atteste de ce biais, avec l'intégration du chatbot « Tay » de Microsoft sur Twitter, sa présence sur ce RSV se voulait purement expérimentale, avec l'objectif d'échanger et de développer son modèle en direct avec les jeunes américains. Pour mener à bien les premiers dialogues, elle s'établissait sur un corpus de langages accessibles publiquement et sur un corpus de phrases comiques, ajoutés par des humoristes en partenariat avec Microsoft. Sa présence sur Twitter aura été de courte de durée (16 h), supprimé par Microsoft pour cause de propos haineux (raciste, mysogine, etc.) dû aux utilisateurs lui envoyant une quantité importante de messages de ce type.

Selon les sujets traités, les IA sont susceptibles d'être plus ou moins fiables et performantes. Présentement, les conditions qui assurent ce résultat à une IA sont la présence de grands corpus de données, de contenus précis avec des variables fixes et claires, de règles de modération universelle applicable dans les pays et communautés. La modération

automatique des contenus violant les droits d'auteurs demeure fiable, de par la précision, le volume du corpus de données et l'universalisation des règles de modération du domaine. Son fonctionnement se base sur une technique de comparaison (Hachage Numérique, Empreinte Digitale) du contenu. L'IA prend en compte le contenu à vérifier et compare les différents contenus présents dans la base de données qui stocke tous les contenus considérés comme protégés. De surcroît, sa fiabilité se voit quasiment infaillible aux multiples manipulations (redimensionnement, altération de la qualité, changement des couleurs, réduction ou augmentation de la durée des vidéos et audio, etc.) possibles des contenus du fait de son entrainement à les détecter. Cependant, la performance et la fiabilité diminue fortement lorsque l'on constate l'apparition de nuance et de contexte. On retrouve ce type d'IA en baisse de performance dans les cas de la reconnaissance d'image ou le comportement et la parole humaine. La complexité réside dans la limitation de l'IA de comprendre les nuances de langage, de comportement, ainsi que les variations contextuelles qui diffèrent selon le lieu et la communauté ou figure le contenu. Par exemple, certains mots de vocabulaire ou d'argot peuvent être interprétés comme légitime ou insultant selon le pays, la plateforme ou la communauté (Le mot « Negro » en Espagne correspond à la couleur noire, mais en Français c'est une manière péjorative et raciste de citer un individu de couleur noir). En plus de cela, se rajoute une évolution constante du discours et de l'argot dans les communautés, ce qui exige au développeur d'actualiser régulièrement le modèle des IA. Le manque de performance de ces IA dans ce domaine, provoque de la censure abusive. Pour la contrer, les communautés inventent des langages alternatifs pour continuer de s'exprimer librement sur des sujets sensibles pour les plateformes. Les communautés de la plateforme Tik Tok, ont élaboré un langage alternatif du nom de « algospeak » pour cause de modération abusive (suppression ou déréférencement des contenus). Ce langage se compose d'alternative pour le mot lesbien devenue « le$bean », pandémie changée en « panini » et « panda express » ou suicide transformé en « becoming unalive »[32].

La reproche de transparence des géants du web provient en partie dans la conception et l'activité des IA de MDC sur notre quotidien. On rencontre des problèmes de transparence sur la globalité du processus de modération automatique, du début avec les données utilisées pour former le modèle, au milieu avec le traitement de la modération qui se situe dans la « boite noire » et la fin avec le voile mis sur le volume de contenus supprimés et déréférencés.

Suite à des demandes de transparence, ces géants du web ont répondu par un refus, en adoptant le secret commercial pour conserver leur avantage concurrentiel sur le marché. Néanmoins, les plateformes modifient depuis peu leur politique de modération, en instaurant de la transparence avec un mécanisme de contestation donnant la possibilité de responsabiliser les utilisateurs, avec l'intégration de fonctionnalités d'avis et d'appels des contenus modérés.

5.3.3 Les IA existantes

Le marché de l'IA est florissant, pas seulement dans la MDC, on retrouve une multitude d'IA performante dans des domaines variés :

- Le milieu de la santé a vu l'apparition en 2018 d'une IA pour repérer les cancers de la peau[33], les résultats étaient probants avec 95 % de bonnes réponses pour elle et 89 % pour les médecins. On enregistre également plusieurs IA pour détecter les cancers du sein[34], mais DeepMind une société d'Alphabet (Google) s'est fait remarquer en 2020 avec son IA qui a réussi à réduire le taux d'erreur de 5,7 % au Royaume-Unis et de 9,4 % aux États-Unis. Le système d'apprentissage automatique s'est construit avec 29 000 mammographies de femmes.
- Le marché du commerce de détail a vu l'expérimentation des magasins autonomes sans caisses. C'est Amazon qui a lancé publiquement ses magasins Amazon Go/Amazon Fresh[35] (prise de petit-déjeuner, tous types de collations) et Amazon Go Grocery (Produits d'hygiènes, produits ménagers, plats cuisinés). Plusieurs IA sont combinées pour mener à bien le processus « Just Walk out ». Dans cette combinaison d'IA, il en existe une pour suivre le mouvement des objets sur les étagères, une pour surveiller le mouvement des clients et une pour l'analyse du comportement et du physique des clients pour recommander ou prédire des achats, etc.
- Dans le milieu de la musique, le laboratoire de recherche Sony CSL Paris a élaboré Flow Machines[36], une IA permettant de composer des musiques dans les styles pop, jazz et brésilien. En 2016, apparait deux musiques, « Mr Shadow » et la plus connu « Daddy's Car » qui se repose sur le style des Beatles. Le système d'apprentissage de l'IA s'est formé sur 13 000 partitions dans les 3 styles ciblés.

6. Les solutions automatiques de la modération

Comme nous l'avons aperçu, la MDC automatique est indispensable pour traiter les grands volumes de données présents de nos jours, que ce soit dans une approche exclusivement automatisée ou hybride. Selon les plateformes de RSV, la diffusion de contenus des utilisateurs s'opère sous différents supports (Audio, Vidéo, Texte, Image). Nombreux sont les exemples, Snapchat spécialisé dans les contenus de type vidéo et image, YouTube dans la vidéo ou Facebook qui à travers les posts laisse le choix du support à l'utilisateur. Ces supports comportent des spécificités liées aux caractéristiques humaines (l'ouïe, la vision, la réflexion, etc.), rendant la compréhension et l'analyse complexe pour les outils automatiques. Plusieurs techniques permettent de répondre à ce besoin, en intégrant les connaissances suffisantes pour classer ou modérer automatiquement les contenus des multiples supports.

Le support vidéo utilise toutes les caractéristiques des autres supports, avec le visuel en référence à l'image, le son lié à l'audio et le texte dans le cas d'une présence d'un quelconque langage à l'image, sollicitant une réflexion pour en analyser le sens. En somme, la modération des vidéos requiert une combinaison des différentes techniques de modérations de l'image, l'audio et le texte.

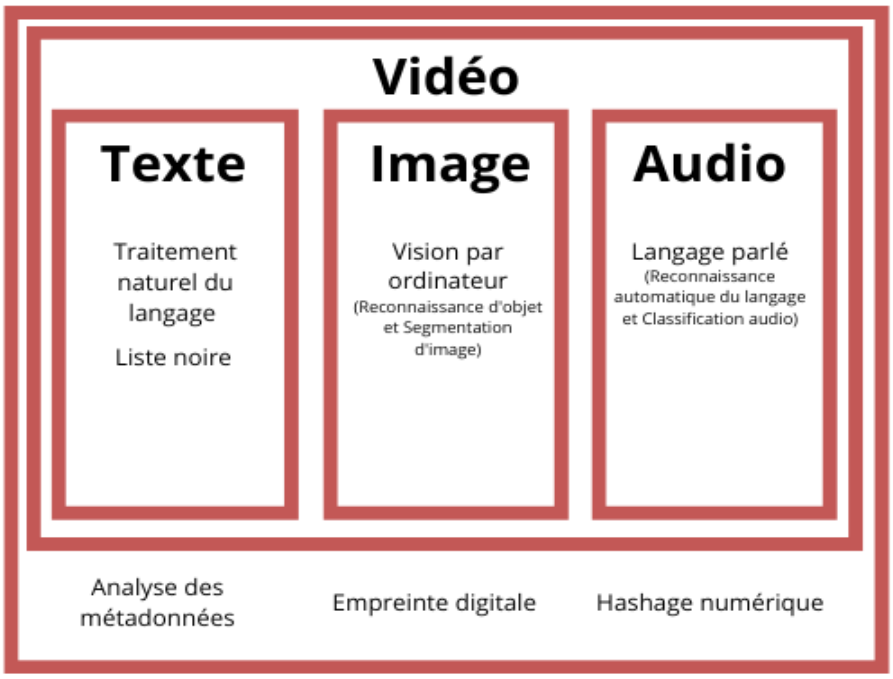

Figure 26 : Les solutions utilisées pour la modération automatique des contenus

6.1 Modération des contenus sous forme de texte

La MDC sous forme de texte comprend tous les types de textes, allant du texte littéraire au commentaire sous un post. Selon l'objectif de modération, on peut adapter le dispositif. On peut dans des cas simples, appliquer une liste noire de mots ou d'expression à bannir, ou à l'inverse avec l'IA analyser la langue et le sens des phrases. Lorsque l'on parle de droit d'auteur, on comparera les contenus en leur entièreté pour détecter les violations.

6.1.1 Liste noire (Blacklist)

Le but de la liste noire est de monter et actualiser régulièrement les mots ou expression indésirables. Pour détecter les contenus indésirables, les textes sont comparés à la liste noire pour trouver des correspondances. Lorsqu'il y a une concordance, le contenu se voit signalé ou supprimé.

L'action sur le contenu de cette technique de modération est facilement contournable par les utilisateurs. La recherche d'une concordance exacte est un problème, laissant la possibilité aux utilisateurs de corrompre la liste noire avec l'écriture de variante des mots, ou avec l'ajout d'émojis dans la composition des mots.

En ce qui concerne l'accessibilité de cette technique, les plateformes mettent à disposition aux utilisateurs la possibilité de filtrer par une liste noire les contenus de ses posts ou pages.

6.1.2 Traitement automatique du langage naturel

À propos du TALN, c'est un des sous-domaines de l'IA qui s'inspire de méthodes de diverses disciplines, de la linguistique à la science des données. Il rend possible l'interaction entre l'Homme et la machine, grâce à la création d'un langage commun nommé « langage naturel ». C'est avec ce langage commun que les machines sont en mesure de lire et interpréter le sens des phrases du langage humain.

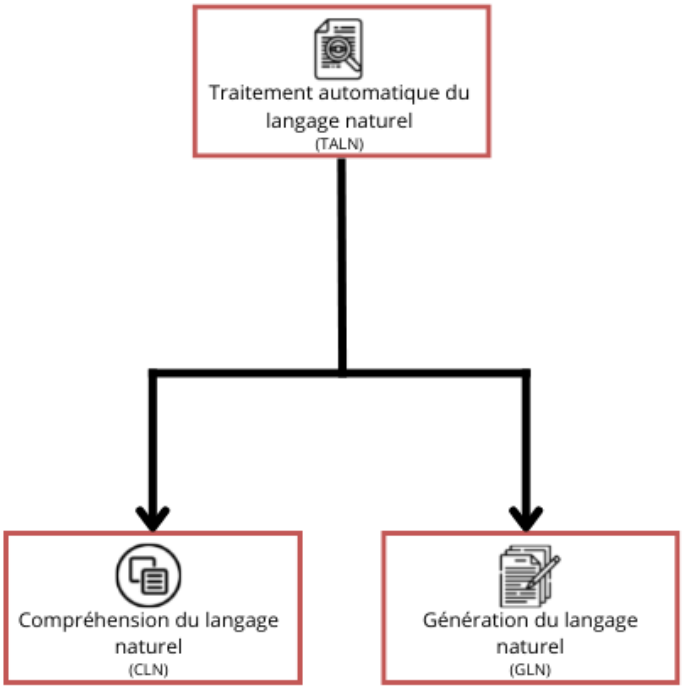

Figure 27 : Composition de la sous-branche de l'IA TALN

Le TALN permet aux machines de comprendre le langage humain sous forme écrite et verbale. À l'aide du processus de « reconnaissance d'entités nommées »[32] et d'identifiants de modèle de mots (tokenisation, radicalisation, lemmatisation, etc.), elle est capable de convertir littéralement les données non structurées du langage humain, en des données structurées. Deux composantes se sont vues intégrées à cette sous-branche de l'IA, pour venir la compléter et l'approfondir. La compréhension du langage naturel (CLN ou NLU en anglais) qui se concentre sur la compréhension de la lecture par ordinateur, en considérant la syntaxe, la sémantique et diverses méthodes telles que la détection du sarcasme, l'analyse des sentiments, etc. Et la génération du langage naturel (GLN ou NLG en anglais) qui permet aux machines de produire des réponses en langage humain sous des formats textuels ou audios.

[32] Processus de traitement et d'organisation des données non structurées provenant du langage humain. Il permet de catégoriser les mots qui composent les phrases pour en comprendre le but et le sens. Exemple : « Carlo a rejoint Nestle pour un stage à Paris » -> Carlo = Personne, Nestle = Entreprise, Localisation = Paris.

L'approche du DL est généralement favorisée pour l'apprentissage des modèles de TALN, exposant davantage de flexibilité et d'intuitivité.

Une large variété de domaine s'appuie sur TALN et ses composantes, et c'est également le cas de la MDC. Dans la situation de contrôle de contenu textuel, TALN et CLN peuvent être combinés ou utilisés indépendamment.

Malgré des avancées conséquentes sur ce sujet, les modèles sont encore perfectibles, dû à la grande complexité et pluralité du langage humain, contenant des différences grammaticales, syntaxique et des ambiguïtés liées au sarcasme ou à la culture.

6.2 Modération des contenus sous forme d'image

Le traitement des images peut s'effectuer différemment selon les cas de figure. L'usage de l'IA se voit intéressant pour des besoins de reconnaissance d'image pour détecter certains types de contenu. Enfin, il existe des situations plus précises, tel que la violation de droits d'auteur ou le sujet de la pédopornographie, que l'on résout avec des programmes qui comparent les images à une base de données répertoriant les contenus de ce genre.

6.2.1 Vision par ordinateur

Pour ce qui de la MDC des images, nous allons utiliser la vision par ordinateur, sous-branche de l'IA. L'objectif est de donner à la machine la faculté d'acquérir une compréhension de haut niveau au moyen de vidéo ou d'image dans notre cas. Comme pour la précédente sous-branche de l'IA examinée, la vision par ordinateur se forme de plusieurs composantes de reconnaissance d'image qui fonctionnent différemment.

La composante primaire se nomme la **classification**, elle se fonde sur l'idée du classement avec des étiquettes qui donnent la possibilité de catégoriser les images qu'on lui transmet. Les images utilisées pour l'entrainement ne doivent pas contenir différents éléments. Dans le cas d'une image avec plusieurs éléments, il faut mettre en place la notion de position, pour distinguer les éléments à reconnaitre. Les repères de position des éléments peuvent s'afficher par une forme rectangulaire/carré autour d'eux, on l'appelle le cadre de sélection, ou également par coloration des pixels formant les éléments à discerner. Concernant le positionnement des éléments avec un cadre de sélection, on retrouve la composante de **localisation**, employée pour la détection d'un élément seulement, puis la composante de

reconnaissance d'objets, utile afin de repérer plusieurs éléments différents. Pour ce qui du repérage par coloration des pixels, elle permet de segmenter les éléments de l'image. Il en existe deux modèles, qui vont tous deux créer un masque de pixels pour chaque élément de l'image, manipulable de manière totalement autonome pour la détection ou pour venir en complément du positionnement par cadre de sélection, clarifiant les images et éléments, prenant en compte seulement ce qui apparait comme essentiel à détecter. La composante et premier modèle est la **segmentation sémantique**, elle va colorer et classifier les éléments par grandes catégories, sans nuances (exemple : la catégorie véhicule est appliquée sur tous les éléments ressemblant à une voiture, un camion, un scooter, etc.). Enfin, la composante et second modèle se voit être la **segmentation par instance**, qui va nuancer sa distinction des éléments, en attribuant des sous-catégories, précisant la nature des éléments et l'itération d'apparition (exemple : pour la catégorie véhicule, il existera des sous-catégories comme voiture, vélo, scooter, camion, etc.).

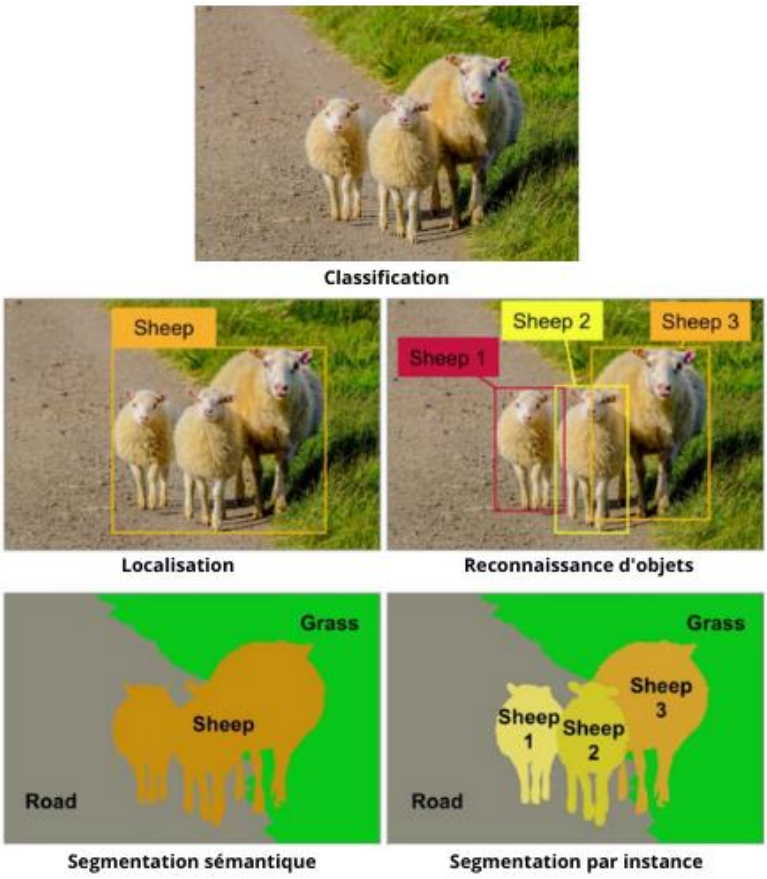

Figure 28 : Différentes composantes pour la reconnaissance des images (source : towardsdatascience)

6.3 Modération des contenus sous forme d'audio

Concernant la modération des contenus audios, une nouvelle sous-branche de l'IA voit son utilité, c'est le langage parlé. À travers les composantes de cette sous-branche, la machine a la possibilité d'obtenir une compréhension des enregistrements et des pistes audios qui figurent dans des vidéos. Généralement sur les RSV, la MDC audio se traite avec la composante de la reconnaissance automatique de la parole (Speech-to-text en anglais) et la classification audio. L'apprentissage des modèles de ces deux composantes, se réalise principalement par le DL, pour des raisons de complexité de traitement des fréquences audios et d'analyse du langage. De plus, l'usage de l'IA est évitable lorsque l'on a besoin de détecter des contenus audios précis, comme pour le cas de la violation des droits d'auteurs.

6.3.1 Langage parlé

La **reconnaissance automatique de la parole** est une composante qui se combine avec la sous-branche TALN, développée précédemment. L'audio est récupéré et converti sous forme de texte grâce à la reconnaissance automatique de la parole, puis la sous-branche TALN intervient avec sa composante de compréhension du langage naturel pour en analyser le texte et en ressortir les entités, l'intention et le sens. La **classification audio** est une composante qui procède au classement des sons dans diverses catégories. Contrairement à la reconnaissance automatique de la parole, son emploi ne se tourne pas seulement vers la classification du langage naturel, mais également vers 3 autres types de sons, la classification des évènements acoustiques[33], les sons environnementaux[34] et la musique[35].

6.4 Solutions de modérations communes

Nous venons d'examiner les multiples solutions de modérations automatiques spécifiques aux formats textes, image et audio. Ceci-dit, il en existe également des communes à tous, avec l'analyse des métadonnées, le hashage numérique, l'empreinte digitale (fingerprinting en

[33] La classification d'évènements acoustiques permet d'identifier et classer l'endroit d'origine d'où provient l'enregistrement audio (exemples : dans une maison, une voiture, dans les hauteurs d'une montagne, etc.).

[34] La classification environnementale permet d'identifier et classer les différents sons de l'audio (exemples : une sonnette, un klaxon, une sirène, etc.)

[35] La classification musicale permet d'identifier et classer les différents audios dans des catégories musicales (exemples : jazz, rock, pop, etc.)

anglais). Ces solutions ont la possibilité d'être déployé sur tout type de fichier numérique. On aperçoit leur emploi dans les missions de détection de la violation des droits d'auteurs, ou de pédopornographie pour les images. Par exemple, il est possible de reconnaitre des contenus de pornographie juvéniles avec la solution PhotoDNA ou de détecter d'autres contenus illégaux, comme le terrorisme avec la solution eGLYPH.

6.4.1 Analyse des métadonnées

Tout contenu numérique se compose de métadonnées, ce sont les informations liées au contenu. Ces informations peuvent se retrouver sous la forme du nom du propriétaire, de l'auteur, du titre, de l'éditeur, du numéro de page, de la durée d'une vidéo ou audio, du format du fichier, etc. Principalement utilisé pour la violation de droits d'auteur, avec la comparaison des métadonnées identiques avec les fichiers d'origines. Toutefois, c'est une solution comportant des failles, avec des résultats imprécis liés à aucune unicité des informations et une facilité pour les modifier.

6.4.2 Hashage numérique

Le hashage numérique propose une solution davantage robuste que l'analyse des métadonnées, en récupérant le contenu et en créant un enregistrement lié de manière unique à celui-ci dans une base de données. Lors de la modération, les contenus sont comparés à la base de données, pour découvrir une concordance avec un enregistrement. C'est une solution efficace, appliqué pour un large domaine de détection des contenus, allant de la violation des droits d'auteur aux contenus terroristes. Néanmoins, son fonctionnement est perfectible, produisant un nouvel enregistrement unique pour chaque modification effectuée sur un contenu, ce qui laisse des failles.

6.4.3 Empreinte digitale

Enfin, l'empreinte digitale est une solution qui reprend le fonctionnement du hashage numérique, en ajoutant également un enregistrement unique du contenu. Cependant, la différence se voit dans le processus d'enregistrement du contenu, qui offre plus de robustesse contre la modularité. Des caractéristiques (pour une image les mouvements et objets présents, pour un audio la fréquence sonore, l'environnement acoustique et le contenu textuel) viennent accompagner l'enregistrement pour anticiper les éventuelles modifications

du contenu d'origine. C'est la solution la plus robuste des 3, qui demande la mise en place d'un environnement complexe et couteux.

Conclusion

L'objectif de ce mémoire était de mettre en lumière les clés d'analyse et de compréhension, pour en déduire si la modération des RSV pouvait se perfectionner à travers l'automatisation des processus grâce à l'IA.

Il était nécessaire de débuter par la présentation des notions fondamentales de la sociologie des relations humaines, pour examiner le fonctionnement des réseaux sociaux réels et virtuels. Lorsque les bases sociologiques se sont vues exposées, il a paru possible de mener l'état des lieux des RSV, en passant par son histoire, les méthodes d'encadrement et de modération des communautés, puis les dispositifs instaurés pour la MDC actuellement et dans un futur potentiel.

Durant l'observation des RSV alt-tech, nous avons constaté que la liberté d'expression à outrance sur ce type de plateforme laissée apparaître des débordements, liés à des communautés en majorité extrémistes, incitant à la haine. Ce sont des cas concrets justifiant le besoin de modération.

En réponse à la problématique, l'automatisation totale des processus de modération n'est toujours pas privilégiée, pour des raisons de faille dans la compréhension des différents types de contenus. Pour des questions de fiabilités des résultats, la modération hybride se voit préférée, combinant la modération automatique et manuelle, pour faciliter le travail des humains tout en continuant l'apprentissage des machines. Ceci dit, la modération hybride comme manuelle comporte des problématiques qu'il faut résoudre pour pouvoir maintenir ce genre de modération. Cela passera par une réelle prise de conscience des plateformes sur les conditions des modérateurs humains. De plus, les plateformes doivent jouer la carte de la transparence, pour éduquer et responsabiliser leurs utilisateurs autour des processus de modération.

Le changement d'angle de ces plateformes autour du design pourrait grandement diminuer la quantité de contenus à modérer. Cela exigerait d'appliquer un design éthique, et du design persuasif positif, pour altérer l'addiction effective en venant répondre seulement aux besoins

de ces derniers et non ceux des plateformes. En somme, ça demanderait un changement du business model.

Ce travail de mémoire s'est tourné principalement sur le web d'aujourd'hui, correspondant à un web centralisé. Cependant, de multiples services décentralisés voient le jour, tel que les cryptomonnaies, ce qui nous fait poser des questions sur la forme du web de demain. Il se voudra peut-être décentraliser, comme le projet du web3 porté par Gavin Wood, souhaitant un web décentralisé établit sur la blockchain. Ce changement de forme du web apporterait des modifications sur les processus de modération.

Bibliographie

[1]

Pierre MERKLE, intervention sur les « Réseaux sociaux », 2011

[2]

Mark GRANOVETTER, « La force des liens faibles », 1973

[3]

Karl MARX, « Le capital », 1867

[4]

Pierre BOURDIEU, « Le capital social », *Actes de la Recherche en Sciences Sociales*, p. 2-3, 1980

[5]

Pierre BOURDIEU, « Les trois états du capital culturel », *Actes de la Recherche en Sciences Sociales*, p 3-6, 1979

[6]

David RIESMAN, « La foule solitaire », 1950

[7]

Robert PUTNAM, « Bowling Alone », 2000

[8]

Nathalie BLANPAIN et Jean-Louis PAN KE SHON, Etude sur la sociabilité en France entre 1983 – 1997, 1998

[9]

Anabel QUAN-HAASE et Alyson YOUNG, « Uses gratifications of social media: a comparison of Facebook and Instant Messaging », *Bulletin of Science, Technology & Society*, volume 30, n°5, p. 350-361, 2010

[10]

Bénédicte AFFO et Olivier ROQUES, « Réseaux sociaux virtuels : Un relais à la sociabilité et au capital social ? », 2015

[11]

Jean CARON et Stéphane GUAY, « Soutien social et santé mentale : concept, mesures, recherches récentes et implications pour les cliniciens », *Santé mentale au Québec*, volume 30, n°2, p. 15-41, 2005

[12]

Monica C. HIGGINS et Kathy E. KRAM, « Reconceptualizing Mentoring at work : A Develeopmental Network Perspective », *The Academy of Management Review*, volume 26, n°2, p.264-288, 2001

[13]

Lee SANG-HOON et Kim YO-HAN, « L'expression de soi et les réseaux sociaux », *Société*, n°133, p.49-60, 2016

[14]

Simon KEMP, « Digital 2022 : Une nouvelle année de croissance exceptionnelle ! », We are social et Hootsuite, 2022

[15]

Burrhus Frederic SKINNER, « Expérience de la boite de SKINNER », début des années 1930

Il a étudié le comportement des souris sur le système de récompense avec la nourriture. Il a démontré que lorsque la nourriture n'était pas distribuée de manière fixe, les souris actionnaient plus souvent le bouton pour en obtenir. Concluant que les cobayes étaient bien plus motivés par une récompense variable que fixe.

[16]

Charles DARWIN, « L'expression des émotions chez l'homme et les animaux », 1872

Il a démontré qu'il y avait 6 catégories d'émotions universelles, la peur, la joie, la tristesse, le dégout, la surprise et la colère.

[17]

Source des données provenant de l'application de visualisation des données WISQARS du Centers for Disease Control and Prevention (CDC). Le filtre : Year Range : 2001 – 2018 ; Intent of Injury : Self-Harm ; Mechanism of Injury : Cut/Pierce ; Disposition : All Cases ; Sex : Males et Females ; Ages : From 10 to 14 To 20 to 24

[18]

Source des données provenant du Centers for Disease Control and Prevention (CDC)

[19]

Victor PAPANEK, « Design for the real world, Human Ecology and Social change », 1970

[20]

Assemblée Générale des Nations Unies, « Le Pacte International relatif aux Droits Civils et Politiques (PIDCP) », 1966

[21]

Assemblée Générale des Nations Unies, « La Déclaration Universelle des Droits de l'Homme (DUDH) », 1948

[22]

Chiffres des votants lors des élections législatives américaine

[23]

Emmanuel PIERRAT, « La privatisation de la censure », *Constructif*, n°56, p. 32-35, 2020

[24]

Laetitia AVIA, Loi Avia, 2020

[25]

Commission Européenne, Digital Services Act, 2020

[26]

Christophe ASSELIN, « Instagram, les chiffres incontournables en 2022 en France et dans le monde », Digimind, 2022

[27]

Thomas COEFFE, « Chiffres Facebook – 2021 », Blog du modérateur, 2021

[28]

« Tuerie de Pittsburgh : Gab, le réseau social utilisé par le tireur, se dit contraint de fermer », BFM, 2018

[29]

Damien LELOUP, « Ecofascisme : comment l'extrême droite en ligne s'est réapproprié les questions climatiques », 2019

[30]

Adrian CHEN, « The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed », WIRED, 2014

[31]

Warren McCULLOCH et Walter PITTS, « A logical calculus of the ideas immanent in nervous activity », *The Bulletin of Mathematical Biophysics*, volume 5, p. 115-133, 1990

[32]

Taylor LORENZ, « Internet 'algospeak' is changing our language in real time, from 'nip nops' to 'le dollar bean' », The Washington Post, 2022

[33]

« Une intelligence artificielle repère les cancers de la peau à partir de photos, et mieux qu'un dermatologue », Huggingtonpost, 2018

[34]

Bastien L, « L'IA du MIT permet de prédire le cancer du sein beaucoup plus tôt », lebigdata.fr, 2019

[35]

Maggie TILLMAN, « Amazon Go et Amazon Fresh : comment fonctionne la technologie « Just walk out », Pocket-lint, 2022

[36]

Elsa FERREIRA, « Comment l'IA de Sony Flow Machines se prend pour les Beatles », Makery, 2016

Suite de la liste des ressources exploitées pour monter ce mémoire :

Sociabilité et réseaux sociaux :

https://www.cairn.info/revue-reseaux-2016-1-page-165.htm

https://www.franceculture.fr/sociologie/lien-social-et-reseaux-sociaux

http://ses.ens-lyon.fr/articles/les-reseaux-sociaux-138014

http://www.couplesfamilles.be/index.php?option=com_content&view=article&id=339:reseaux-sociaux-entre-reel-et-virtuel-la-sociabilite-en-evolution&catid=6&Itemid=108

https://www.cairn.info/revue-societes-2016-3-page-49.htm

https://www.rse-magazine.com/Mark-Granovetter-et-la-force-des-liens-faibles_a3736.html

https://journals.openedition.org/sociologies/2902

https://media-animation.be/Facebook-echanges-sociaux-faibles.html#nb20

https://www.cairn.info/revue-d-economie-du-developpement-2010-4-page-97.htm

https://halshs.archives-ouvertes.fr/halshs-02468843/document

https://www.cairn.info/revue-management-et-avenir-2011-5-page-179.htm

https://ses.webclass.fr/notions/capital-culturel/

https://www.persee.fr/doc/arss_0335-5322_1979_num_30_1_2654

https://www.cairn.info/revue-l-economie-politique-2001-4-page-32.htm

https://www.erudit.org/fr/revues/mi/2013-v17-n2-mi0560/1015398ar/

https://hal.archives-ouvertes.fr/hal-03271144/document

https://ses.webclass.fr/notions/capital-social/

https://www.persee.fr/doc/arss_0335-5322_1979_num_30_1_2654

https://ses.webclass.fr/notions/capital-culturel/

https://www.rse-magazine.com/Pierre-Bourdieu-et-les-formes-de-Capital_a3583.html

https://www.persee.fr/doc/arss_0335-5322_1980_num_31_1_2069

https://laurentcombaud.typepad.fr/blog_de_laurent_combaud/2008/01/tout-le-monde-a.html

Emergence des réseaux sociaux virtuels :

https://www.arturin.com/infographie-evolution-reseaux-sociaux-1997-2019/

https://histoire-internet.vincaria.net/2020/04/1997-peer-to-peer-p2p.html

https://arxiv.org/abs/2001.02611

https://www.antevenio.com/fr/une-breve-histoire-des-reseaux-sociaux/

https://www.mindfruits.biz/blog/quelle-est-lhistoire-des-reseaux-sociaux-voici-les-dates-importantes/

https://www.ledroitautravail.fr/differents-reseaux-sociaux/

https://www.futura-sciences.com/tech/definitions/informatique-reseau-social-10255/

https://www.elaee.com/2015/08/27/23907-la-messagerie-instantanee-enjeu-majeur-des-reseaux-sociaux

https://www.webmarketing-conseil.fr/liste-reseaux-sociaux/

https://knowledgeone.ca/du-web-1-0-au-4-0/?lang=fr

https://www.youtube.com/watch?v=8ZxTtMYo9Mc

https://www.ionos.fr/digitalguide/sites-internet/developpement-web/arpanet/

https://c-marketing.eu/du-web-1-0-au-web-4-0/

https://flatworldbusiness.wordpress.com/flat-education/previously/web-1-0-vs-web-2-0-vs-web-3-0-a-bird-eye-on-the-definition/

https://siecledigital.fr/2021/07/26/usage-internet-et-mobile-en-2021/

https://www.agence90.fr/chronologie-innovations-reseaux-sociaux/#2010_Le_premier_reseau_social_mobile_avec_Instagram

https://www.blogdumoderateur.com/30-chiffres-internet-reseaux-sociaux-mobile-2022/

https://www.cairn.info/revue-reseaux1-2002-2-page-276.htm

https://www.neodia.fr/strategie-digitale/109-facebook-business-model-perspectives-et-web-marketing

https://status200.net/what-is-web-5-0-a-brief-introduction-to-world-wide-web-1-0-5-0/

https://fr.wikipedia.org/wiki/M%C3%A9dia_social

https://www.notboring.co/p/own-the-internet?s=r

https://readwrite.com/the-next-evolution-of-the-internet-is-closer-than-it-seems/

https://fredcavazza.net/2021/06/03/les-differents-stades-devolution-du-web/

Le flux sur les réseaux sociaux virtuels :

https://www.revuepolitique.fr/les-maitres-de-la-manipulation-un-siecle-de-persuasion-de-masse/

https://www.kibo.ac/tinder-strategie-recompense-aleatoire/

https://medium.com/@onurkarapinar/comment-la-technologie-pirate-lesprit-des-gens-e8bd041adb4c

https://brandnewsblog.com/2018/05/06/reseaux-sociaux-et-plateformes-apres-des-annees-de-manipulation-cognitive-vers-une-veritable-ethique-de-lattention/

https://numerique-investigation.org/snapchat-comment-les-filtres-nous-piegent-ils/2407/

https://www.windtopik.fr/quest-ce-que-leffet-zeigarnik-et-comment-le-maitriser/?msclkid=acc0154cc61111ec996f080c2ba7d9cd

https://journals.openedition.org/sdj/286#quotation

https://www.beedeez.com/fr/gamification?msclkid=d3187b95c61211eca80a1cab3aa32d70

https://mad-monkeys.fr/la-gamification-3-exemples-outre-atlantique/?msclkid=4a2a50ccc61211ec87863b74aff69861

https://www.youtube.com/watch?v=KJScBxSnelY

http://webchronique.com/economie-de-l-attention/?msclkid=6a39f868c62011ec9164471e7aeeaf48

https://www.lecho.be/opinions/general/nous-vivons-dans-un-monde-d-autopropagande/10327630.html

https://www.ionos.fr/digitalguide/web-marketing/analyse-web/la-bulle-de-filtres/?msclkid=dfde088dc63411ec9e6f9171b9847c0a

https://www.youtube.com/watch?v=mEuokfY0EH0

https://france.filgoodhealth.com/fr/dossiers/les-addictions-comportementales-9

https://www.inserm.fr/dossier/addictions/

https://sproutsocial.com/insights/new-social-media-demographics-fr_fr/

https://www.cdc.gov/nchs/products/databriefs/db362.htm

https://www.cdc.gov/nchs/data/databriefs/db362-tables-508.pdf#2

https://wisqars.cdc.gov/data/non-fatal/explore/trends?nf=eyJpbnRlbnRzIjpbIjIiXSwibWVjaHMiOlsiMzA4MCJdLCJ0cmFmZmljIjpbIjAiXSwiZGlzcCI6WyIxIiwiMiIsIjQiLCI1Il0sInNleCI6WyIxIl0sImFnZUdyb3Vwc01pbiI6WyIxMC0xNCJdLCJhZ2VHcm91cHNNYXgiOlsiMjAtMjQiXSwiY3VzdG9tQWdlc01pbiI6WyIwIl0sImN1c3RvbUFnZXNNYXgiOlsiMTk5Il0sImZyb21ZZWFyIjpbIjIwMDEiXSwidG9ZZWFyIjpbIjIwMTgiXSwiYWdlYnV0dG4iOiI1WXIiLCJncm91cGJ5MSI6IkFHRUdQIn0%3D

Ethics by Design - 2020

dgitags.io | Le design éthique : qu'est-ce que c'est ?

L'éthique de la persuasion dans la conception | Idées Adobe XD

Qu'est-ce que le design éthique ? La réponse de Use.Design, agence de designer spécialisée dans l'expérience et l'interface utilisateur.

Les limites éthiques du design persuasif | par Amoli | Collectif UX (uxdesign.cc)

Qu'est-ce que le design éthique ? Définition Design éthique (usabilis.com)

Design éthique : quelles pratiques pour un design responsable ? (lebondigital.com)

Les principes du design éthique (et comment s'en servir) - 99designs

Les types de modérations :

https://www.20minutes.fr/high-tech/2897883-20201106-reseaux-sociaux-faut-trouver-juste-equilibre-entre-regulation-liberte-expression-explique-romain-badouard

https://la-rem.eu/2021/11/les-enjeux-de-la-moderation-des-contenus-sur-le-web/#:~:text=La%20mod%C3%A9ration%2C%20c'est%20l,sur%20les%20fils%20de%20discussion.

https://urbania.ca/article/quand-la-moderation-sur-les-reseaux-sociaux-renforce-la-discrimination

https://la-rem.eu/2021/11/les-enjeux-de-la-moderation-des-contenus-sur-le-web/

https://newmediaservices.com.au/fundamental-basics-of-content-moderation/

https://www.lemonde.fr/pixels/article/2020/01/11/sarah-t-roberts-les-geants-du-web-ont-choisi-de-rendre-le-processus-de-moderation-invisible_6025491_4408996.html

https://larevuedesmedias.ina.fr/reseaux-sociaux-moderateurs-web-sarah-t-roberts

https://www.lefigaro.fr/secteur/high-tech/2018/09/25/32001-20180925ARTFIG00270-moderation-tri-des-donnees-l-onu-s-inquiete-des-conditions-de-travail-des-ouvriers-du-clic.php

https://usbeketrica.com/fr/article/les-eboueurs-du-web-moderateurs-invisibles-des-reseaux-sociaux

https://www.letemps.ch/societe/reseaux-sociaux-vertige-moderation

https://blog.digimind.com/fr/tendances/fake-news-reseaux-sociaux-a-la-man%C5%93uvre-ou-pas

https://www.numerama.com/politique/15013-les-geants-du-web-en-france-demandent-une-neutralite-du-net-pas-vraiment-neutre.html

https://newmediaservices.com.au/automated-and-live-moderation/

https://imagga.com/blog/automated-content-moderation/

https://apro-software.com/artificial-intelligence-moderate-social-networks/

https://static1.squarespace.com/static/571681753c44d835a440c8b5/t/58d058712994ca536bbfa47a/1490049138881/FilteringPaperWebsite.pdf

https://www.lemonde.fr/pixels/article/2019/10/04/ecofascisme-comment-l-extreme-droite-en-ligne-s-est-reappropriee-les-questions-climatiques_6014255_4408996.html

https://www.lemonde.fr/pixels/article/2019/03/08/comment-un-jeu-video-fde-viol-a-pu-se-retrouver-sur-la-plate-forme-steam_5433258_4408996.html

https://www.bfmtv.com/tech/tuerie-de-pittsburgh-gab-le-reseau-social-utilise-par-le-tireur-se-dit-contraint-de-fermer_AN-201810280027.html

https://www.bfmtv.com/tech/ces-reseaux-sociaux-anti-censure-qui-ont-tout-a-gagner-de-la-moderation-renforcee-sur-facebook_AN-202101080185.html

https://www.ladn.eu/tech-a-suivre/gab-network-twitter-fachosphere-americaine-effacer-web/

https://www.renaissancenumerique.org/ckeditor_assets/attachments/571/la_moderation_des_contenus.pdf

https://www.wired.com/2014/10/content-moderation/

https://www.lemonde.fr/pixels/article/2016/03/24/a-peine-lancee-une-intelligence-artificielle-de-microsoft-derape-sur-twitter_4889661_4408996.html?msclkid=a92b4b0ecf9c11ec995959eb3bb300fb

https://www.francetvinfo.fr/internet/reseaux-sociaux/twitter/tay-le-robot-de-microsoft-quitte-twitter-apres-des-derapages-racistes_1374963.html?msclkid=510fb213cf9c11ecb6a7577c34217de2

https://www.elementai.com/fr/news/2019/le-pourquoi-de-ia-explicable#:~:text=Les%20mod%C3%A8les%20d%E2%80%99IA%20sont%20souvent%20utilis%C3%A9s%20pour%20automatiser,qui%20peuvent%20contenir%20des%20pr%C3%A9jug%C3%A9s%20de%20notre%20soci%C3%A9t%C3%A9.?msclkid=401fdd1ccf9b11ec83e6808f03203958

https://siecledigital.fr/2021/05/11/intelligence-artificielle-quelle-approche-des-biais-algorithmiques/

https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/the-limitations-of-automated-tools-in-content-moderation/

https://www.washingtonpost.com/technology/2022/04/08/algospeak-tiktok-le-dollar-bean/

https://www.youtube.com/watch?v=w_pbwWNvsns

https://www.lesnumeriques.com/vie-du-net/deepmind-google-a-developpe-une-ia-pour-detecter-le-cancer-du-sein-n145427.html

https://www.liberation.fr/france/2020/01/05/selon-google-l-ia-depiste-le-cancer-du-sein-mieux-que-l-humain_1771593/#:~:text=Dans%20un%20article%20publi%C3%A9%20dans,qui%20lui%20ont%20%C3%A9t%C3%A9%20soumises.

https://www.lesechos.fr/tech-medias/intelligence-artificielle/cancer-du-sein-lintelligence-artificielle-de-google-meilleure-que-les-medecins-1159917

https://www.lebigdata.fr/cancer-du-sein-ia-google

https://www.sciencesetavenir.fr/sante/une-ia-de-google-surpasse-les-radiologues-pour-detecter-le-cancer-du-sein_140225

https://siecledigital.fr/2018/06/01/lia-surpasse-les-dermatologues-pour-detecter-les-cancers-de-la-peau/

https://www.huffingtonpost.fr/2018/05/28/une-intelligence-artificielle-repere-les-cancers-de-la-peau-a-partir-de-photos-et-mieux-quun-dermatologue_a_23445515/

https://www.usine-digitale.fr/article/le-premier-supermarche-amazon-go-grocery-ouvre-a-seattle.N933659

https://www.pocket-lint.com/fr-fr/gadgets/actualites/amazon/139650-quest-ce-que-amazon-aller-ou-est-il-et-comment-ca-marche

https://www.zdnet.fr/actualites/le-magasin-sans-caisses-amazon-go-fait-son-entree-en-europe-39918905.htm

https://www.youtube.com/watch?v=qgImjYWY9OQ

https://www.lemagit.fr/conseil/Apprentissage-supervise-et-non-supervise-les-differencier-et-les-combiner

https://machinelearnia.com/machine-learning-introduction/

https://datascientest.com/apprentissage-non-supervise

https://www.lebigdata.fr/data-labeling#:~:text=Les%20donn%C3%A9es%20%C3%A9tiquet%C3%A9es%20sont%20des%20donn%C3%A9es%20marqu%C3%A9es%2C%20annot%C3%A9es%2C,diff%C3%A9rentes%20t%C3%A2ches%20en%20plus%20de%20l%E2%80%99annotation%20de%20donn%C3%A9es.

https://analyticsinsights.io/le-clustering-definition-et-implementations/

https://helios2.mi.parisdescartes.fr/~vincent/siten/Publications/theses/pdf/chouaib.pdf

https://www.quora.com/What-are-the-key-differences-between-the-data-mining-method-prediction-association-and-clustering

https://fr.slideshare.net/ssuserab1db8/les-algorithmes-de-gnration-des-rgles-d-association

https://www.ibm.com/docs/fr/spss-modeler/18.0.0?topic=nodes-association-rules

https://www.youtube.com/watch?v=a4WUL_KZeZo

https://www.youtube.com/watch?v=t3m4MnZa_FY

https://www.youtube.com/watch?v=mxHhrP4fVmM

https://www.youtube.com/watch?v=Rgxs8lfoG4I

https://datascientest.com/q-learning-le-machine-learning-avec-apprentissage-par-renforcement

https://www.simplilearn.com/tutorials/machine-learning-tutorial/what-is-q-learning

https://www.simplilearn.com/tutorials/machine-learning-tutorial/reinforcement-learning#what_is_markovs_decision_process

https://bigdatablog.skapane.com/apprentissage-par-renforcement/

https://towardsdatascience.com/policy-based-reinforcement-learning-the-easy-way-8de9a3356083

https://www.baeldung.com/cs/ml-policy-reinforcement-learning

https://machinelearningknowledge.ai/beginners-guide-to-what-is-policy-in-reinforcement-learning/#:~:text=%20Types%20of%20Policy%20in%20Reinforcement%20Learning%20,given%20state.%20When...%202%20Stochastic%20Policy%20More%20

https://larevueia.fr/apprentissage-par-renforcement/#:~:text=La%20fonction%20de%20valeur%20correspond%20au%20cumul%20des,changeant%20parfois%20de%20strat%C3%A9gie%20pour%20am%C3%A9liorer%20son%20score.

https://medium.com/intro-to-artificial-intelligence/relationship-between-state-v-and-action-q-value-function-in-reinforcement-learning-bb9a988c0127

https://datascientest.com/reinforcement-learning

https://hal.archives-ouvertes.fr/hal-02301161/document#:~:text=Une%20politique%20d%C3%A9terministe%20stationnaire%20est,selon%20la%20politique%20%CB%9C%CF%80.

https://www.veryfrog.com/reseau-neuronal-et-deep-learning/

https://www.futura-sciences.com/tech/actualites/intelligence-artificielle-5-choses-vous-ignorez-deep-learning-74568/

https://deeplylearning.fr/cours-theoriques-deep-learning/fonction-dactivation/

https://datascientest.com/perceptron

https://www.lebigdata.fr/perceptron-machine-learning

https://deepomatic.com/fr/difference-entre-la-computer-vision-la-reconnaissance-dimages

https://www.ibm.com/fr-fr/topics/computer-vision

https://www.jedha.co/blog/la-vraie-difference-entre-machine-learning-deep-learning#:~:text=Il%20est%20souvent%20expliqu%C3%A9%20que,son%2C%20le%20texte%2C%20l%27

https://networkcultures.org/nofun/2021/01/28/full-automation-full-fantasy/

https://journals.openedition.org/quaderni/2049?lang=fr

https://www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf

https://www.forbes.com/sites/kalevleetaru/2019/03/19/the-problem-with-ai-powered-content-moderation-is-incentives-not-technology/?sh=72c2aa055b7b

https://apro-software.com/guide-to-deep-learning/

https://apro-software.com/machine-learning-for-newbies/

https://www.forbes.com/sites/anniebrown/2021/10/27/understanding-the-technical-and-societal-relationship-between-shadowbanning-and-algorithmic-bias/?sh=36cf6fc36296

https://techxplore.com/news/2021-01-exploring-underpinnings-shadowbanning-twitter.html

https://moncoachdata.com/blog/modeles-de-machine-learning-expliques/

https://www.lemagit.fr/conseil/Machine-Learning-les-9-types-dalgorithmes-les-plus-pertinents-en-entreprise

https://www.youtube.com/watch?v=K9z0OD22My4

https://www.youtube.com/watch?v=wg7-roETbbM

https://www.youtube.com/watch?v=SfOoRsUj9kQ&t=8s

https://www.youtube.com/watch?v=XUFLq6dKQok&list=RDCMUCmpptkXu8iIFe6kfDK5o7VQ&start_radio=1

https://www.youtube.com/watch?v=K9z0OD22My4&t=99s

https://www.youtube.com/watch?v=EUD07IiviJg

https://www.youtube.com/watch?v=esiKN7k2IBI&t=491s

https://www.youtube.com/watch?v=RC7GTAKoFGA&t=14s

https://www.youtube.com/watch?v=jTF2aLBHA80

https://www.youtube.com/watch?v=PKNxUF9CGn8

https://www-igm.univ-mlv.fr/~dr/XPOSE2014/Machin_Learning/C_Comparaison.html

https://www.youtube.com/watch?v=EovbRqiVpS8

https://www.youtube.com/watch?v=gPVVsw2OWdM

https://www.ionos.fr/digitalguide/web-marketing/search-engine-marketing/quest-ce-quun-reseau-neuronal-artificiel/

https://www.europarl.europa.eu/RegData/etudes/STUD/2020/657101/IPOL_STU(2020)657101_EN.pdf

https://www.edureka.co/blog/backpropagation/

AI Pricing | How Much Does Artificial Intelligence Cost in 2020? (webfx.com)

How Much Does Artificial Intelligence Cost in 2021-2022? — ITRex (itrexgroup.com)

How Much Does It Cost to Run AI Infrastructure? - Pandio

Cost of Implementing AI into Modern Software Projects (indatalabs.com)

How much does artificial intelligence (AI) cost in 2021? - Azati: Uniting experts to fulfil important projects

Les solutions automatiques de modération

https://blogs.manageengine.com/fr/2022/03/30/comment-fonctionne-le-traitement-du-langage-naturel-tal.html

https://www.lebigdata.fr/traitement-naturel-du-langage-nlp-definition

https://datascience.eu/fr/traitement-du-langage-naturel/traitement-des-langues-naturelles-nlp/

https://www.youtube.com/watch?v=CTXn5YMdkec

https://www.youtube.com/watch?v=1I6bQ12VxV0

https://www.sciencedirect.com/topics/social-sciences/sentiment-analysis

https://www.unite.ai/what-is-natural-language-understanding/

https://www.ibm.com/blogs/watson/2020/11/nlp-vs-nlu-vs-nlg-the-differences-between-three-natural-language-processing-concepts/

https://blog.clevy.io/conversationnel/nlp-nlu-comment-extraire-le-sens-des-enonces/

https://monkeylearn.com/blog/natural-language-understanding/

https://zaion.ai/ressources/actualites/quelle-est-la-difference-entre-nlp-et-nlu/

https://www.lemagit.fr/conseil/Intelligence-Artificielle-quelle-difference-entre-NLP-et-NLU

https://viadialog.com/asr-nlu-nlp-tts-terminologie-ia-simplifiee/

https://fr.shaip.com/blog/named-entity-recognition-and-its-types/#:~:text=Qu%27est-ce%20que%20la%20reconnaissance%20d%27entit%C3%A9%20nomm%C3%A9e%20%3F%20La,classer%20ces%20entit%C3%A9s%20nomm%C3%A9es%20dans%20des%20cat%C3%A9gories%20pr%C3%A9d%C3%A9finies.

https://datascientest.com/computer-vision#:~:text=La%20Computer%20Vision%20ou%20Vision,m%C3%AAme%20mani%C3%A8re%20qu%27un%20humain.

https://cs.stackexchange.com/questions/51387/what-is-the-difference-between-object-detection-semantic-segmentation-and-local#:~:text=%22Object%20detection%22%20is%20localizing%20%2B,is%20basically%20per%2Dpixel%20classification.

https://www.quora.com/What-is-the-difference-between-semantic-segmentation-and-object-detection

https://openaccess.thecvf.com/content_ICCV_2019/papers/Takikawa_Gated-SCNN_Gated_Shape_CNNs_for_Semantic_Segmentation_ICCV_2019_paper.pdf

https://keymakr.com/blog/semantic-segmentation-uses-and-applications/

https://medium.com/analytics-vidhya/image-classification-vs-object-detection-vs-image-segmentation-f36db85fe81

https://towardsdatascience.com/what-is-the-difference-between-object-detection-and-image-segmentation-ee746a935cc1

https://www.finsliqblog.com/ai-and-machine-learning/natural-language-processing-nlp-speech-to-text-technical/#:~:text=Natural%20Language%20Processing%20%28NLP%29%20speech%20to%20text%20is,to%20act%20and%20react%2C%20as%20usual%2C%20humans%20do.

https://www.telusinternational.com/articles/what-is-audio-classification#:~:text=Audio%20classification%20is%20the%20process,and%20text%20to%20speech%20applications.

https://towardsdatascience.com/audio-deep-learning-made-simple-sound-classification-step-by-step-cebc936bbe5

https://www.analyticsvidhya.com/blog/2021/06/introduction-to-audio-classification/

https://appen.com/blog/an-introduction-to-audio-speech-and-language-processing/

https://www.techopedia.com/what-is-the-difference-between-speech-to-text-and-chatbots/7/33228

https://www.kardome.com/blog-posts/difference-speech-and-voice-recognition#:~:text=The%20simple%20definition%20of%20speech,learning%20to%20translate%20human%20speech.

https://www.alibabacloud.com/blog/598438

https://wonderfall.space/apple-csam/

https://www.counterextremism.com/french/eglyph-combattre-lextr%C3%A9misme-sur-le-net

https://www.anishathalye.com/2021/12/20/inverting-photodna/

Annexes

Annexe 1 : Les jeux de données sur les mutilations et suicides des jeunes Américains

Affichage des jeux de données manipulé pour modéliser les graphiques. Toutes les données présentes proviennent du Centers for Disease Control and Prevention (CDC).

a. Jeu de données sur les mutilations des jeunes hommes Américain

Mutilation des jeunes hommes Américain de 2001 à 2018				
Années	10 - 14 ans	15-19 ans	20-24 ans	Par 100 k Hommes
2001	--	49,45	85,56	
2002	--	55,73	59,12	
2003	--	43,11	71,4	
2004	16,73	57,99	72,14	
2005	15,37	68,87	74,97	
2006	11,77	57,03	73,37	
2007	12,85	58,46	63,96	
2008	--	44,81	65,22	
2009	--	43,06	67,82	
2010	16,18	46,07	70,59	
2011	15,98	71,34	79,04	
2012	11,47	72,2	70,87	
2013	26,85	59,99	62,26	
2014	22,02	62,06	55,05	
2015	--	78,13	59,32	
2016	16,95	69,24	71,95	
2017	23,55	73,73	69,46	
2018	30,57	71,57	78,68	
	-3,28750747	-12,9221436	-20,7339878	Total augmentation de 2001 à 2009
	88,9369592	55,3505535	11,4605468	Total augmentation de 2010 à 2018

b. Jeu de données sur les mutilations jeunes femmes Américaine

Mutilation des jeunes femmes Américaine de 2001 à 2018				
Années	10 - 14 ans	15-19 ans	20-24 ans	Par 100 k Femmes
2001	18,75	85,09	59,36	
2002	26	79,86	56,35	
2003	22,34	81,06	53,36	
2004	43,31	116,79	67,92	
2005	36,48	109,24	68,26	
2006	49,38	97,46	76,46	
2007	50,53	91,15	80,81	
2008	35,56	93,43	78,22	
2009	32,11	101,67	73,02	
2010	66,54	117,91	82,62	
2011	71,32	133,73	106,53	
2012	73,92	165,09	91,52	
2013	108,04	138,82	85,91	
2014	143,52	191,75	100,75	
2015	110,24	190,81	106,1	
2016	112,61	203,36	86,66	
2017	114,31	187,96	109,76	
2018	156,12	228,58	105,21	
	71,2533333	19,4852509	23,0121294	Total augmentation de 2001 à 2009
	134,625789	93,8597235	27,3420479	Total augmentation de 2010 à 2018

c. Jeu de données sur les suicides des jeunes hommes Américain

Suicides des jeunes hommes Américain de 2000 à 2018			
Années	10 - 14 ans	15 - 24 ans	Par 100 k Hommes
2000	2,3	17,1	
2001	1,9	16,5	
2002	1,8	16,4	
2003	1,7	15,9	
2004	1,7	16,7	
2005	1,9	16,1	
2006	1,4	16	
2007	1,2	15,7	
2008	1,4	16	
2009	1,6	16,1	
2010	1,7	16,9	
2011	1,9	17,6	
2012	2,1	17,4	
2013	2,3	17,3	
2014	2,6	18,2	
2015	2,4	19,4	
2016	2,5	20,5	
2017	3,3	22,7	
2018	3,7	22,7	
	-30,43478261	-5,847953216	Total augmentation en 2000 jusqu'à 2009
	117,6470588	34,31952663	Total augmentation en 2010 à 2018

d. Jeu de données sur les suicides des jeunes femmes Américaine

Suicides des jeunes femmes Américaine de 2000 à 2018			
Années	10 - 14 ans	15 - 24 ans	Par 100 k Femmes
2000	0,6	3	
2001	0,6	2,9	
2002	0,6	2,9	
2003	0,5	3	
2004	0,9	3,5	
2005	0,7	3,5	
2006	0,6	3,2	
2007	0,5	3,1	
2008	0,7	3,5	
2009	0,9	3,6	
2010	0,9	3,9	
2011	0,8	4	
2012	0,8	4,5	
2013	1,4	4,5	
2014	1,5	4,6	
2015	1,6	5,3	
2016	1,7	5,4	
2017	1,7	5,8	
2018	2	5,8	
	50	20	Total augmentation en 2000 jusqu'à 2009
	122,222222	48,7179487	Total augmentation en 2010 à 2018

Annexe 2 : Les coûts de l'IA

Le secteur de l'IA est en forte croissance depuis de nombreuses années, de multiples entreprises misent aujourd'hui sur l'instauration de systèmes d'IA. Il existe deux types de solution d'IA, **la solution IA préconçue** et **la solution IA personnalisée.**

Concernant le tarif de la conception et l'intégration d'une IA, cela varie selon des facteurs et des décisions prises. Présentation des facteurs qui font varier le prix d'une solution d'IA préconçue et personnalisé

1. Le type d'IA (Analyse de texte, reconnaissance d'image, Assistants virtuels)
2. Le type de projet (Si le type d'IA souhaité existe déjà avec une IA préconçue ou non, alors se diriger vers une IA personnalisée)
3. Fonctionnalité de l'IA (différentes tâches voulues)
4. Niveau de précision des prédictions ciblé dans la/les tâches
5. Stockage et la structure de données utilisées pour l'entrainement (structuré, non structuré ou à créer (par renforcement))

Dans le cas d'une IA préconçue, le prix augmentera dans la formule selon la taille de stockage idéale pour le stockage des données. La prise en compte de la structure des données est essentielle, elle influera sur le prix (Plus complexe de traiter des données non structurées que des données structurées, ce qui augmentera le prix) et la solution d'IA préconçue (si la solution peut traiter les données non structurées par exemple).

Dans le cas d'une IA personnalisée, il faut compter tous les coûts liés à la mise en place du stockage (Data Lake, Data Warehouse, etc.) et processus d'intégration des données (ETL, ELT). De plus, la structure des données importe pour la jauger la complexité de l'IA dans l'acquisition et le traitement des données pour son entrainement.

6. La visualisation et le pilotage

Pour une solution d'IA préconçue, une interface web avec un tableau de bord sera mis à disposition pour visualiser les résultats et la contrôler.

Pour une solution d'IA personnalisée, il faudra monter cette interface web, ce qui engendre des couts supplémentaires.

7. Maintenance interne et externe

Lorsque l'on parle de solution d'IA préconçue et personnalisée, la maintenance peut s'effectuer en interne ou externe. La maintenance en interne demande de former ses employés et/ou embaucher. La maintenance en externe permet de déléguer ce poids à une entreprise tiers.

8. Durée du projet

Maintenant que nous avons examiné les processus qui affecte le cout lors d'un projet IA, nous pouvons donner des fourchettes de prix pour ces deux types de projets. Pour l'entrainement et l'intégration d'une IA préconçue, le cout pourrait être de 0 à 37 500 euros. Il est possible d'intégrer des IA préconçues Open Source, gratuites, faisant appel à des connaissances personnelles pour réaliser les processus d'entrainement, intégration et maintenance. Enfin, le projet de création d'une IA personnalisé, pourrait se trouver de 5500 à 282 500 euros.

Annexe 3 : Jeu de données Algorithme K-means

Simulation d'un cluster avec l'algorithme K-Means en ligne Online K Means Clustering (revoledu.com). Ci-dessous, vous pouvez retrouver le tableau résultant du passage de l'algorithme.

Label	Vector	Cluster id	Cluster centroid
Werner	0.39,0.16,5,3	1	0.35,0.12,4.5,3
Kolo Muani	0.33,0.13,6,5	3	0.4033333333333333,0.16333333333333333,5.333333333333333,4.666666666666666
Gameiro	0.31,0.08,4,3	1	0.35,0.12,4.5,3
Lukaku	0.41,0.02,4,2	2	0.385,0.05,3.5,1.5
Giroud	0.36,0.08,3,1	2	0.385,0.05,3.5,1.5
Aubameyang	0.53,0.09,7,4	0	0.8333333333333333,0.2233333333333333,6.666666666666666,3
Delort	0.43,0.12,5,5	3	0.4033333333333333,0.16333333333333333,5.333333333333333,4.666666666666666
Moreno	0.45,0.24,5,4	3	0.4033333333333333,0.16333333333333333,5.333333333333333,4.666666666666666
Benzema	0.94,0.29,7,2	0	0.8333333333333333,0.2233333333333333,6.666666666666666,3
Haaland	1.03,0.29,6,3	0	0.8333333333333333,0.2233333333333333,6.666666666666666,3

Annexe 4 : Jeu de données Algorithme Apriori

Simulation d'une situation de règle d'association avec l'algorithme apriori en ligne Algorithme Apriori - codeding.com. Ci-dessous, vous pouvez retrouver les pourcentages de soutien et les règles d'associations résultantes.

Pourcentage de soutien des listes d'items dans le jeu de données :

13 Large Itemsets (by Apriori):

{Pain} (support: 80%)

{Pâte à tartiner} (support: 40%)

{Fromage} (support: 60%)

{Bière} (support: 40%)

{Sel} (support: 50%)

{Pâte} (support: 50%)

{Pain, Pâte à tartiner} (support: 40%)

{Pain, Fromage} (support: 40%)

{Pain, Bière} (support: 40%)

{Pain, Sel} (support: 40%)

{Fromage, Sel} (support: 40%)

{Pain, Pâte} (support: 40%)

{Fromage, Pâte} (support: 40%)

Règles d'associations :

6 Association Rules

{Pâte à tartiner} => {Pain} (Support: 40.00%, Confidence: 100.00%)

{Bière} => {Pain} (Support: 40.00%, Confidence: 100.00%)

{Sel} => {Pain} (Support: 40.00%, Confidence: 80.00%)

{Sel} => {Fromage} (Support: 40.00%, Confidence: 80.00%)

{Pâte} => {Pain} (Support: 40.00%, Confidence: 80.00%)

{Pâte} => {Fromage} (Support: 40.00%, Confidence: 80.00%)