ÉTHIQUE DE L'INFORMATIQUE : CAS À ÉTUDIER

Offre d'emploi : LLM spécialisé dans le code informatique

SOMMAIRE

1 CADRE TECHNIQUE : MODÈLES *TRANFORMERS*

1.1 Intelligence artificielle : deux grandes avancées

L'intelligence artificielle a connu dans la dernière décennie d'immenses progrès grâce aux réseaux neuronaux profonds (*Deep Neural Networks* ou DNN) en ce qui concerne deux aspects : la reconnaissance des images et le travail sur le langage.

Cela a été permis **grâce à trois facteurs** :

- une **quantité très importante de données** entre autres par le biais du *cloud* ;
- des **capacités de calcul bien plus importante** par le biais, entre autres, des cartes graphiques ;
- des **investissements financiers importants** – ainsi Microsoft qui avait investi 1 milliard de dollars dans OpenAI en 2019, compte en investir 10 de plus en 2023.

1.1.1 Les images

Jusqu'en 2012, ce qu'on appelle « intelligence artificielle », ce sont avant tout des systèmes experts, c'est-à-dire des programmes complexes intégrant de multiples règles pouvant produire une réponse en fonction des données qui lui sont soumises. En gros, pour leur apprendre à reconnaitre un chat dans une photo, il fallait les programmer pour reconnaitre les oreilles pointues, les moustaches, la queue et toute autre caractéristique pertinente.

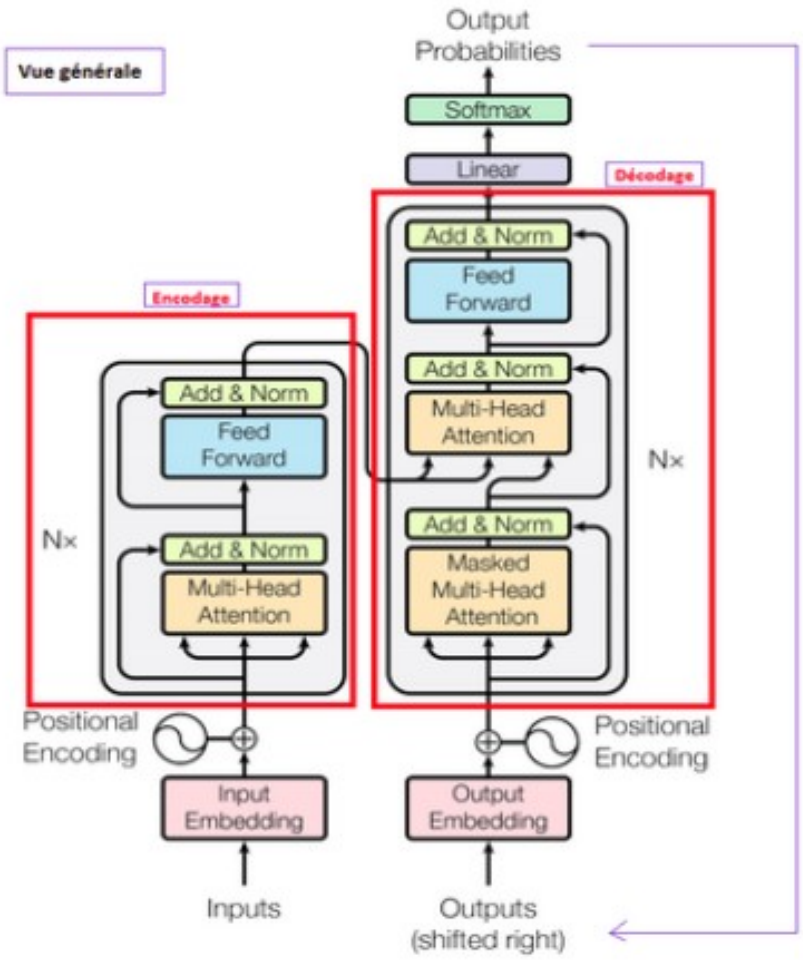
Depuis les années 80, il existait une autre branche, plus théorique, basée sur **« l'apprentissage profond », ou *Deep Learning***. Mais elle n'avait jamais produit de résultats probants à grande échelle. Et puis, en 2012, pour la première fois, grâce à la puissance de calcul monumentale des processeurs graphiques, un algorithme d'apprentissage profond arrive premier dans un concours de reconnaissance d'images, devant les meilleurs systèmes experts du monde. Ce **saut qualitatif sur les images** a été permis grâce à de vieux travaux concernant les réseaux convolutifs, repris entre autres par Yann le Cun[1], qui ont pu être exécutés grâce à la quantité de data à disposition depuis les clouds et à la puissance de calcul générée par les cartes graphiques (GPU) et les data centers.

1.1.2 Les textes[2]

Des chercheurs de Google ont publié **en 2017 un article révolutionnaire pour le traitement naturel des langues (NLP) : « *Attention is all you need* »**. Il concernait la traduction, mais ces recherches ont permis une avancée majeure dans le traitement du texte, des images, etc.

1 Yann Le Cun (né en 1960), professeur d'informatique et de neurosciences à l'université de New York, dirige la recherche fondamentale chez Facebook. Il a publié en 2019 *Quand la machine apprend. La révolution des neurones artificiels et de l'apprentissage profond* et a obtenu le Prix Turing 2019, conjointement avec Yoshua Bengio et Geoffrey Hinton, pour leurs travaux de recherche sur l'intelligence artificielle et le *deep learning*. **2023 : ces trois chercheurs divergent complètement concernant les LLM : le premier a peur, le second est réservé et appelle à un moratoire, le troisième incite à prendre le sujet à bras-le-corps**.

2 Partie créée grâce à « « BERT de Google AI sur le banc de test ! », à « BERT : Faire comprendre le langage naturel à une machine, en pré-entrainant des Transformers bi-directionnels profonds » et à **« Traitement automatique des langages : do you speak computer ? » du 5-02-2022 qui complète cette réflexion**.



Pour traduire, on va boucler et ajouter un mot à la traduction à chaque boucle.

Exemple : "Je me sens bien" à traduire en "I feel good" (et pas en "I me smell good" entraduction littérale).

Boucle 1 : Toute la phrase à traduire rentre dans l'encodeur, et le token <début de phrase> entre dans le décodeur. En sortie on a "<début de phrase> I"

Boucle 2 :
- Entrée du décodeur : La sortie de la boucle 1, à savoir "<début de phrase> I"
- En sortie on aura "<début de phrase> I feel"

Boucle 3 :
- Entrée du décodeur : La sortie de la boucle 2, à savoir "<début de phrase> I feel"
- En sortie on aura "<début de phrase> I feel good"

Boucle 4 :
- Entrée du décodeur : La sortie de la boucle 3, à savoir "<début de phrase> I feel good"
- En sortie on aura "<début de phrase> I feel good <fin de phrase>"

Ici on a une fin de phrase, on considère qu'on a fini de traduire la phrase

The Transformer – architecture, schéma provenant du document « BERT : Faire comprendre le langage naturel à une machine, en pré-entraînant des Transformers bi-directionnels profonds », d'après « Attention Is All You Need »

En 2018 : même saut pour le travail sur les textes avec l'apparition des modèles *Transformers* et en particulier BERT de Google. En traitement automatique du langage naturel, BERT[3] (acronyme anglais de *Bidirectional Encoder Representations from Transformers*) est un modèle de représentation du langage développé par Google en 2018[4] et lancé en octobre 2019. Les modèles *Transformer* (« transformeur » en français) sont d'autant plus importants qu'aujourd'hui **l'une des tâches les plus populaires en data science, c'est le traitement des informations présentées sous forme de texte**. Précisément, il s'agit de représenter du texte sous forme d'équations mathématiques, de formules, de paradigmes, de modèles afin de comprendre la sémantique (le contenu) du texte pour son traitement ultérieur : classification, fragmentation, etc. De manière générale, l'ensemble de ces pratiques sont réunies dans **le domaine du traitement du langage naturel (TNL ou NLP en anglais : *natural langage processing*)**.

3 D'après « BERT de Google AI sur le banc de test ! » et « BERT (modèle de langage) »

4 Voir « BERT : Pre-training of Deep Bidirectional Transformers forLanguage Understanding » par Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova de Google AI Language.
Résumé : « Contrairement aux récents modèles de représentation du langage, **BERT est conçu pour pré-entrainer les représentations bidirectionnelles profondes à partir de textes non étiquetés en conditionnant conjointement les contextes gauche et droit dans toutes les couches**. En conséquence, le modèle BERT préformé peut être affiné avec une seule couche de sortie supplémentaire afin de créer des modèles de pointe pour un plus large éventail de tâches, telles que la réponse à des questions et l'inférence linguistique, sans modification substantielle de l'architecture spécifique à la tâche. **BERT est conceptuellement simple et empiriquement puissant**.

1.2 « *Transformer* » : un modèle performant

1.2.1 Objectif

BERT (Google) ou GPT (OpenAI et Microsoft) s'appuient sur le modèle *Transformer* (ou modèle auto-attentif) apparu en 2017 qui est un modèle d'apprentissage profond utilisé principalement dans le domaine du traitement automatique des langues (TAL)**.** L'idée consiste à entrainer des modèles à prédire le mot suivant en fonction d'un début de texte et à le comparer à des textes déjà disponibles, notamment sur Internet. La force de ce modèle est de **construire des abstractions en partant de la place des mots et de leur proximité relative sur l'ensemble d'un corpus et de créer ainsi tout un réseau sémantique complexe qui permet ensuite de prédire le mot suivant à partir de l'intégralité du texte qui le précède**.

1.2.2 Fonctionnement

Les *Transformers* sont des composants qui s'appuient sur des méthodes d'attention[5]. Un *transformer* est construit sur la base d'un encodeur et d'un décodeur. Un encodeur se compose de couches d'attentions reliées à l'encodeur précédent (ou l'entrée pour le premier) et de couches denses. Un décodeur se compose, de manière similaire, de couches d'attentions reliées au décodeur précédent et de couches denses ainsi que d'une nouvelle couche d'attention supplémentaire intercalée entre ces deux couches et reliée à la sortie du dernier encodeur.

Avant *transformer*, la plupart des modèles de Traitement Naturel du Langage (NLP) étaient **entrainés pour des tâches spécifiques** comme la classification ou la traduction de texte. Ils utilisaient l'apprentissage supervisé. Or, **cette approche présente deux inconvénients**. Le premier est un manque de données annotées à disposition, et le second est l'incapacité à effectuer des tâches générales.

1.2.2.1 Fonctionnement de BERT[6]

Avant BERT, les mots-clés étaient pris individuellement sans forcément créer des liens entre eux. Désormais, les phrases sont prises dans leur ensemble tout en tenant compte du contexte. Ce qui inclut les prépositions, les expressions longues et complexes, les mots de liaison, les nuances polysémiques, les pronoms… tout ce qui peut faire partie du langage naturel qu'un internaute peut inclure dans sa requête.

Google BERT se base sur un **réseau de neurones qui traite le langage naturel (NLP). Le *deep learning*, associé au NPL, analyse les requêtes** en examinant les termes qui précèdent ou suivent un mot-clé pour comprendre l'intention de l'internaute. **Cette intelligence artificielle évalue ensuite les mots les plus importants et les plus déterminants pour le sens d'une phrase, afin de ne présenter que des résultats pertinents et précis**.

---

5 **Les mécanismes d'attention sont les moyens de faire passer au décodeur l'information de quelles étapes de l'encodeur sont les plus importantes au moment de générer un mot de sortie**. Quels sont les mots de la séquence d'entrée qui se rapportent le plus au mot qu'il est en train de générer en sortie, soit qu'ils s'y rapportent comme contexte pour lui donner un sens, soit qu'ils s'y rapportent comme mots "cousins" de signification proche.
Les mécanismes d'auto-attention (*self-attention*) sont similaires, sauf qu'au lieu de s'opérer entre les éléments de l'encodeur et du décodeur, ils s'opèrent sur les éléments de l'*input* entre eux (le présent regarde le passé et le futur) et de l'*output* entre eux aussi (le présent regarde le passé, vu que le futur est encore à générer).

6 D'après « Google BERT : comprendre et s'adapter à ce nouvel algorithme »

1.3 BERT et GPT

1.3.1 Année décisive : 2018

GPT-1 : un modèle génératif de langage à 117 millions de paramètres, entrainés sur des données sans label et *fine-tuned* (raffiné) pour des tâches spécifiques comme la classification et l'analyse de sentiment ;

BERT est constitué de 340 millions de paramètres à optimiser. C'est un modèle pré-entrainé prêt à l'emploi. Son apprentissage s'est fait sur la totalité du corpus anglophone de Wikipédia. Quatre jours de calcul ont été nécessaires sur 64 TPU (*Tensor Processing Unit*) suivi d'environ trois semaines de calcul pour améliorer l'apprentissage[7].

Il se compose d'une suite d'encodeurs seulement (N = 12 ou 24 suivant la version : Base ou Large)[8].

Cet algorithme permet d'intégrer en quantité impressionnante non seulement du vocabulaire, mais également des tournures de phrase, etc. il appartient à la catégorie des « modèles contextuels de langue », capables de maitriser grammaire, syntaxe ou vocabulaire et également contexte.

Son modèle possède trois avantages sur les autres :

- il est **plus performant** que ses prédécesseurs **en termes de résultats** ;
- il est plus performant que ses prédécesseurs **en termes de rapidité d'apprentissage** ;
- une fois pré-entrainé, de façon non supervisée, il possède une « représentation » linguistique qui lui est propre. Il est ensuite possible, sur la base de cette représentation initiale, de le customiser pour une tâche particulière. **Il peut être entrainé en mode incrémental (de façon supervisée cette fois) pour spécialiser le modèle rapidement et avec peu de données**.

1.3.2 Nombreux domaines d'application

L'objectif de BERT est, comme souvent chez Google, de fournir des résultats toujours plus pertinents aux recherches des internautes[9]. En effet, cette méthode a permis d'améliorer significativement les performances en traitement automatique des langues.

BERT essaye donc de « comprendre » les intentions formulées derrière les expressions naturelles. **Google utilise de tels modèles pour mieux répondre aux requêtes adressées à son moteur de recherche**[10] – comme celui-ci lui rapporte environ 80 % de ses revenus, l'investissement dans de telles recherches est très rentable à terme. Une **meilleure « compréhension » du contexte (plus fine et sur une fenêtre bi-directionnelle très large)** et une **meilleure prise en compte de l'importance de l'ordre des mots** sont les points forts mis en avant par Google dans le cadre d'analyse de requêtes. De quoi alimenter ensuite les applications de traduction, de simplification de texte, de synthèse de grands corpus, d'analyse sémantique, de recommandation, de réponses sensées à des questions…

D'où d'importants progrès dans de nombreux domaines :

7 Le TPU est un type de processeur dédié au calcul d'apprentissage des réseaux de neurones.

8 Pour plus d'informations techniques, lisez « BERT : Faire comprendre le langage naturel à une machine, en pré-entrainant des Transformers bi-directionnels profonds ».

9 De la même façon, Microsoft tente d'améliorer Bing grâce à GPT-3.

10 Cette nouvelle mise à jour permet à Google Search de mieux comprendre le langage naturel des internautes qui préfèrent taper des mots-clés plutôt que des questions construites quand ils effectuent une recherche.

- **pour les requêtes**, évidemment – Google estime que cela pourrait impacter et améliorer le résultat d'une requête sur dix. Pour les requêtes de deux ou trois mots, BERT ne sera d'aucune utilité. **Seules les requêtes longues et complexes se retrouvent affectées**. L'algorithme devra aider à :
  - ✗ résoudre les problèmes d'homonymie ;
  - ✗ replacer les phrases dans leur contexte sémantique ;
  - ✗ déterminer les nuances dans les requêtes ;
  - ✗ retrouver la référence précise d'une longue requête conversationnelle ;
  - ✗ prédire la phrase suivante ;
  - ✗ répondre directement aux requêtes des internautes dans les résultats de recherche, notamment à travers les featured snippets[11] ou extraits enrichis ;
- **pour la traduction** – il peut même une fois pré-entrainé pour traduire par exemple : [français/anglais – anglais/français] puis [anglais/allemand – allemand/anglais], ensuite il peut traduire du français vers l'allemand sans entrainement ;
- **pour classifier les textes** (l'agressivité des tweets ou le repérage des messages haineux sur Facebook par exemple) ;
- **pour les générateurs de textes** (qui créent par exemple des articles de presse). Il est d'ailleurs très créatif comme le montre Łukasz Kaiser – chercheur chez Google Brain – dans sa présentation des modèles de réseaux neuronaux attentionnels (de 3min18 à 6min50) ;
- **ou bien pour donner une réponse à une question** ;
- etc.

Google a rendu libres les modèles BERT, comme souvent pour les technologies que l'entreprise découvre, même si sa découverte a couté cher (plus de 3 millions de dollars pour faire tourner les machines de calcul).

L'Inria[12], grâce à ce partage, a fait même chose en français et a créé les modèles CamemBERT (novembre 2019) avec Facebook (110 millions de paramètres). FlauBERT est un autre modèle francophone développé par le CNRS[13] (même nombre de paramètres).

1.3.3 De GPT-2 à GPT-3

En 2019, est présenté **GPT-2 – 1,5 milliard de paramètres** – qui a été entrainé sur un jeu de données bien plus large, afin d'accroitre ses performances. Il exploite aussi les techniques de conditionnement de tâches, d'apprentissage *zero-shot*, et de transfert de tâches *zero-shot*.

Puis, en 2020, OpenAI introduit **GPT-3 et ses 175 milliards de paramètres** (présenté dans l'article « *Language Models are Few-Shot Learners* ») qui a été entrainé sur un *dataset* encore plus large pour maximiser ses résultats. Il repose aussi sur les techniques de *learning in-context*, *one-shot*, few-shot et *zero-shot*.

11 Les *Featured Snippets* sont des encarts particuliers qui se trouvent en haut des résultats lors d'une recherche sur Google. Pour les créer, Google utilise des sites de confiance et le contenu des liens de la première page de résultats, sur laquelle sont extraites les données. Le moteur de recherche affiche alors, dans le *Featured Snipped*, le lien qui lui semble le plus pertinent.

12 L'Institut national de recherche en informatique et en automatique (Inria) est un établissement public français spécialisé en mathématiques et informatique, créé en 1967.

13 Centre national de la recherche scientifique.

La technologie mise en œuvre n'est « *pas tellement une nouveauté* », indique à l'AFP[14] Amine Benhenni, expert IA chez Dataswati. « *La recherche sur les apprentissages de séquences se développe depuis quelques années.* ***La grosse différence, c'est la taille du modèle*** ». GPT-3 a en effet été nourri du contenu de milliards de pages web accessibles librement en ligne et d'innombrables ouvrages en tout genre. **L'intégralité de Wikipédia ne représente au final que 3 % du contenu total de son apprentissage – 300 000 milliards de mots pour produire un modèle de langue**. Il est également capable de programmer en CSS, JSX, Python, entre autres.

Avec 175 milliards de paramètres, **il peut désormais être utilisé tel quel, alors que les modèles précédents nécessitaient toujours un « ré-entrainement » sur des tâches spécifiques**. GPT-3 « *est incroyablement puissant si vous savez comment bien l'amorcer* », a témoigné Shreya Shankar, une ingénieure et chercheuse spécialisée sur l'IA. « *Cela va changer le paradigme de l'apprentissage machine : au lieu de construire des immenses jeux de données pour ré-entrainer les modèles, il sera possible d'extrapoler à partir de quelques exemples* », a-t-elle poursuivi.

Toutefois, GPT-3 lancé en 2020 est encore loin d'être parfait. Des problèmes tendent à survenir quand on lui demande de générer de longs textes, en particulier sur des sujets complexes nécessitant des connaissances précises. GPT-4, arrivé en mars 2023, a apporté des améliorations à ces problèmes.

1.4 Importance relative des paramètres[15]

1.4.1 Des paramètres de plus en plus nombreux

Il faut voir un modèle d'apprentissage comme une **monumentale collection de paramètres, c'est-à-dire des variables de configuration interne au modèle IA dont les valeurs peuvent être extraites depuis les données fournies, qu'il faut réussir à régler le plus finement possible** pour obtenir le bon résultat en fonction des données qu'on lui fournit en entrée. Ce réglage se fait automatiquement, petit à petit, grâce à des formules mathématiques complexes, par exemple, grâce à l'algorithme de rétropropagation du gradient d'erreur.

Le nombre de paramètres d'un modèle IA est couramment utilisé comme métrique de performances. Selon la « *Scaling Hypothesis* », les performances de la modélisation de langage augmentent proportionnellement à la taille du langage, au volume de données et à la puissance de calculs. C'est pourquoi de nombreux créateurs d'IA se focalisent sur l'augmentation du nombre de paramètres. Ainsi, en 2020, Microsoft a présenté sa génération de langage naturel Turing (T-NLG) comme le « *plus grand modèle de langage jamais publié avec 17 milliards de paramètres* ». À ce moment-là, il a donné de meilleurs résultats que les autres modèles linguistiques dans une variété de tâches.

Pareillement, **GPT-3 a cent fois plus de paramètres que GPT-2** ; le modèle présenté par Google en avril 2022, **PalM, en compte 540 milliards**. Le CNRS a annoncé fin 2022 la livraison de Bloom, issu du projet international *BigScience*, capable de gérer 176 milliards de paramètres. Et les résultats sont impressionnants.

14 L'Agence France-Presse (AFP), créée en 1944, est chargée de collecter, vérifier, recouper et diffuser l'information, sous une forme neutre, factuelle et utilisable directement par tous types de médias (radio, télévision, presse écrite, sites internet) ainsi qu'auprès des grandes entreprises et administrations.

15 Partie créée d'après des extraits de l'article « GPT-4 : une IA 100 fois plus puissante que ChatGPT en 2023 ? Tout savoir » paru sur *Le Big Data* le 18/01/23.

1.4.2 La qualité plutôt que la quantité

Toutefois, **GPT-4 n'est pas beaucoup plus large que GPT-3** : il aurait 8 x 220 milliards de paramètres soit 1,7 trillions. Le "8x" est important car les paramètres ne sont pas tous utilisés en même temps.

Il existe depuis quelques années des modèles bien plus larges que GPT-3 et GPT-4. Toutefois, leurs performances sont similaires et **des modèles de moindre taille peuvent offrir des performances supérieures**. En d'autres termes, **le nombre de paramètres n'est que l'un des facteurs impactant les performances d'un modèle**. C'est pour cela qu'à présent, OpenAI cherche à **créer des modèles plus petits et plus performants**.

En effet, les **modèles larges requièrent un vaste jeu de données, des ressources informatiques massives et leur implémentation est complexe**. Même le déploiement de ces modèles peut s'avérer **trop onéreux pour de nombreuses entreprises**.

1.4.2.1 Une paramétrisation optimale

La plupart des larges modèles de langage sont sous-optimisés. Compte tenu du cout massif de leur entrainement, les entreprises doivent souvent **sacrifier la précision pour faire des économies**. Par exemple, **GPT-3 n'a été entrainé qu'une seule fois malgré la présence d'erreurs**. Les chercheurs n'avaient pas le budget requis pour une optimisation des hyperparamètres[16]. Or, Microsoft et OpenAI ont prouvé que **GPT-3 aurait pu être amélioré s'il avait été entrainé sur des hyperparamètres optimaux**.

Les deux firmes ont compris que **les meilleurs hyperparamètres pour les plus petits modèles sont les mêmes que les meilleurs pour les plus larges avec la même architecture**. Cette découverte permet aux chercheurs d'**optimiser les modèles larges pour une fraction du cout**.

1.4.2.2 Des tokens[17] d'entrainement de plus en plus nombreux ?

Récemment, DeepMind a découvert que **le nombre de tokens d'entrainement influence autant les performances du modèle que sa taille**. Les chercheurs ont entrainé un modèle

16 Dans l'apprentissage automatique, un hyperparamètre est un paramètre dont la valeur est utilisée pour contrôler le processus d'apprentissage. En revanche, les valeurs des autres paramètres (généralement la pondération de nœuds) sont obtenues par apprentissage.
Les hyperparamètres peuvent être classifiés comme étant des hyperparamètres de modèle, qui ne peuvent pas être déduits en ajustant la machine à l'ensemble d'entrainement parce qu'ils s'appliquent à la tâche de la sélection du modèle, ou des hyperparamètres d'algorithmes, qui en principe n'ont aucune influence sur la performance du modèle mais affectent la rapidité et la qualité du processus d'apprentissage. Un exemple d'hyperparamètre de modèle est la topologie et la taille d'un réseau de neurones. Des exemples d'hyperparamètres d'algorithme sont la vitesse d'apprentissage et la taille des lots.
Les différents hyperparamètres varient en fonction de la nature des algorithmes d'apprentissage. Compte tenu de ces hyperparamètres, l'algorithme d'apprentissage apprend les paramètres à partir des données.

17 Un **token** désigne une *entité* (ou *unité*) *lexicale*, dans le contexte de l'apprentissage automatique et du traitement du langage naturel, est une représentation numérique d'un élément de texte. Ce processus de conversion du texte en nombres est appelé **tokenisation**. Avec GPT, les tokens peuvent être des lettres, des groupes de lettres, ou des mots entiers. Par exemple, la phrase « Ceci est un test » est encodée en utilisant un seul token pour chaque mot, tandis que le mot « entremêlé » est encodé en trois tokens : le préfixe « entre- », le radical « mêl » et le suffixe « -é » qui indique un participe passé ou un adjectif.
La façon dont le texte est tokenisé peut avoir un impact significatif sur les performances du modèle, rendant le processus de tokenisation crucial pour l'apprentissage automatique et le traitement du langage naturel.
La tokenisation est utilisée dans tous les domaines impliquant le traitement du langage naturel : traduction automatique, génération de texte, analyse des sentiments, réponse automatique aux questions…

Chinchilla 70B quatre fois plus petit que Gopher, mais capable de rivaliser avec lui grâce à un entrainement sur un nombre de données quatre fois supérieur.

Ainsi, **GPT-4 Turbo, sorti en novembre 2023, a une fenêtre de contexte de 128 000 tokens, soit quatre fois plus que le plus grand modèle de GPT-4**[18]. **Cela implique que GPT-4 Turbo peut traiter environ 96 000 mots d'un coup**. C'est plus que ce que contiennent la plupart des romans. De plus, **ceci permettra au chatbot de tenir des conversations plus longues sans oublier le sujet** comme c'était le cas jusqu'à présent.

Cette augmentation du nombre de tokens d'un LLM à un autre induit que l'entrainement des modèles nécessite toujours plus de FLOPS[19] **pour atteindre la perte minimale**.

1.4.2.3 Le premier modèle parcimonieux d'OpenAI ?

Les modèles parcimonieux utilisent le calcul conditionnel pour réduire les couts. Cela signifie que tous les neurones ne sont pas actifs simultanément. Un tel modèle peut facilement **dépasser le milliard de paramètres sans impliquer de couts massifs**. Ceci permet d'entrainer de larges modèles de langage[20] avec un nombre réduit de ressources. De plus, **les modèles parcimonieux comprennent davantage le contexte**.

2 MISSION À ACCOMPLIR

OpenAI ouvre un poste d'ingénieur·e en *Deep Learning* spécialisé·e dans l'entrainement et l'optimisation de *Large Language Models* (LLM). Plus particulièrement, il s'agira de travailler sur un LLM spécialisé dans le domaine médical : chatbot pour médecins et patients, aide au diagnostic…

3 QUESTION ÉTHIQUE

Acceptez-vous ce travail ? Pour quelles raisons (donnez au minimum deux raisons pour et deux contre) ?

18 Par comparaison, GPT-3 a une limitation de seulement 4096 tokens par interaction.

19 Le nombre d'opérations en virgule flottante par seconde (*floating-point operations per second*) est une unité de mesure de la rapidité de calcul d'un système informatique et donc d'une partie de sa performance.

20 Un *language model* est un modèle créé pour prédire les mots à venir, comme GPT-3.

4 POUR ALIMENTER LA TENSION ÉTHIQUE

Lisez les points suivants[21] pour mieux comprendre les enjeux éthiques liés à cet emploi.

4.1 OpenAI

4.1.1 D'une association à but non lucratif aux recherches publiées en open source…

Les modèles de langage GPT-3 et GPT-4 (*Generative Pre-trainer Transformer*), tout comme le chatbot ChatGPT, le générateur d'images DALL-E et le générateur de code Codex, sont des programmes développés par **OpenAI**, une **entreprise spécialisée dans le raisonnement artificiel**.

Elle a été **fondée en 2015** par Elon Musk[22] et **Sam Altman** dont la vision de l'IA est empreinte de transhumanisme et le long-termisme[23] : si on peut améliorer l'homme par la technologie et par l'IA, il est nécessaire de prendre en compte des risques existentiels menaçant des milliards de vies humaines, comme ceux posés en théorie par l'IA. Le dessein d'OpenAI est donc de « ***créer une intelligence artificielle générale, de façon sûre et bénéfique pour l'humanité*** ». OpenAI a ainsi cherché à s'associer à quelques startups utilisant le raisonnement artificiel pour **avoir un effet transformateur**, par exemple dans les domaines des soins de santé, du changement climatique et de l'éducation et « *où les outils d'IA peuvent autonomiser les gens en les aidant à être plus productifs* ». Afin d'éviter de voir un projet jugé aussi crucial soumis aux logiques financières, l'entreprise est **à but non lucratif** et **toutes les recherches devaient être publiées en accès libre open source** pour éviter que l'IA reste aux mains d'un monopole ou d'un oligopole.

4.1.2 … À une structure à but lucratif aux recherches fermées

Mais un **tournant majeur a lieu en 2019** : Sam Altman devenu directeur général affirme que l'amélioration des modèles d'IA passe par leur taille, nécessitant une puissance de calcul phénoménale, donc des milliards de dollars. Il se tourne alors vers **Microsoft qui investit 1 milliard, sous la forme d'accès à ses data centers** pour développer conjointement de nouvelles technologies de supercalculateurs. Sam Altman crée pour cela une **structure à but lucratif et renonce à publier en open source le modèle de traitement du langage GPT-2**. Ce changement a été très critiqué.

Fin 2022, OpenAI lance ChatGPT qui connait un succès fulgurant. **Microsoft annonce plus de 10 milliards de dollars d'investissement**. **L'objectif est d'accélérer l'intégration des solutions d'OpenAI dans les différents produits de Microsoft** : le cloud, Word, Outlook, PowerPoint, Bing. Ce nouveau financement prévoit que Microsoft obtienne 75 % des bénéfices d'OpenAI jusqu'à la récupération de son investissement initial. Parallèlement, Altman devenu un patron star de la tech déborde sur le terrain politique : il plaide entre autres pour une régulation

21 Vous trouverez les principales sources ayant permis de créer cette partie dans la sitographie.

22 Elon Musk a quitté la société en 2018 à la suite d'un conflit de direction.

23 Long-termisme : position éthique née dans les années 2010 qui donne la priorité à l'amélioration de l'avenir à long terme. De ce fait, la survie de l'humanité est le seul objectif, éclipsant tous les autres problèmes, en particulier ceux que vit la population au temps présent. Selon ses partisans, l'intelligence humaine est peut-être, dans l'Univers, un phénomène d'une rareté extraordinaire. Il faut donc tout faire pour éviter son extinction. « *Cela justifie donc tous les projets sur la conquête de l'espace, tous les projets en biotechnologie, sur l'immortalité…* », critique Asma Mhalla, spécialiste des enjeux géopolitiques du numérique. Et la lutte contre l'autonomisation d'une « superintelligence » non contrôlée que les long-termistes redoutent.

de l'IA, partage ses peurs qu'elle n'anéantisse l'humanité et assure agir « *avec en tête l'intérêt de [cette dernière]* ».

À l'automne 2023, la société OpenAI est valorisée à plus de 90 milliards de dollars, bien que non rentable – en effet, Sam Altman, directeur général d'OpenAI, estime que le cout moyen de chaque réponse de ChatGPT est de l'ordre de « *quelques centimes* ». L'entreprise débourserait donc environ 3 millions de dollars par mois[24] pour son robot conversationnel, sans oublier qu'OpenAI a versé 120 millions de dollars en 2022 à Google Cloud.

4.1.3 Des visions de l'IA divergentes

Néanmoins, **fin novembre 2023, Sam Altman a été limogé, à la surprise générale, puis remplacé, avant d'annoncer son départ chez Microsoft, puis de faire un retour triomphal à OpenAI, le tout en 5 jours**.

Ce scénario pourrait être expliqué par des divergences de vue. **Une première critique fondamentale s'attaque au but de l'entreprise et au concept-même « *d'intelligence artificielle générale* » : la nature et la possibilité d'une telle superintelligence sont contestées dans la communauté de l'IA**. L'idée sert « *une techno-utopie dangereuse* », solutionniste[25] et élitiste, voire eugéniste car visant à dépasser l'homme, dénoncent certains chercheurs en éthique de l'IA. **Se centrer sur les risques hypothétiques à long terme de l'IA empêcherait de se concentrer sur la régulation de ses risques immédiats et réels, comme les biais sexistes ou racistes, les discriminations, le pillage des œuvres ou la surveillance**, dénoncent les ONG AI Now Institute ou Distributed AI Research Institute (DAIR).

Surtout, un autre front s'est ouvert, créant progressivement un schisme au sein même des partisans d'une « *superintelligence bénéfique* » et d'OpenAI : **les plus « *catastrophistes* » (« *doomers* ») s'opposent aux plus « *techno-optimistes* » et les « *accélérationnistes* » (partisans de hâter les recherches) aux « *décélérationnistes* ».**

Il y avait en outre des tensions personnelles, des débats sur l'accès aux capacités de calcul, ainsi que le fait que Sam Altman s'investisse dans d'autres start-ups, au risque de conflits d'intérêts. Aux yeux de plusieurs membres du conseil d'OpenAI, dont Ilya Sutskever – débauché de Google par OpenAI, une jeune star qui avait accompli un bond décisif dans la reconnaissance d'images – **Sam Altman cherchait à avancer et à lancer des produits à un rythme trop rapide, dans un but jugé trop commercial et contraire à la politique de sureté d'OpenAI**. «*[En renvoyant M. Altman], le conseil n'a fait que remplir sa mission, qui est de s'assurer qu'OpenAI construit une IA qui bénéficie à toute l'humanité* », a justifié M. Sutskever lors d'une réunion interne. Helen Toner, autre membre du conseil, a rédigé un article universitaire où elle regrettait que le lancement de ChatGPT, en bêta et « *sans évaluation complète des risques* », ait déclenché une « *course au moins-disant* » et une vague de chatbots concurrents. Selon *Reuters* et *The Information*, les avancées récentes de chercheurs d'OpenAI sur un modèle baptisé Q*, apparemment capable de résoudre certains problèmes mathématiques, auraient aussi renforcé les inquiétudes du conseil, de même que la réduction par M. Altman d'une équipe de recherche à long terme de M. Sutskever.

Dernier signe du divorce : **Emmett Shear, le PDG nommé par le conseil en remplacement de M. Altman, estimait dans un podcast de juin « *entre 5 % et 50 %* » la probabilité de voir une superintelligence s'autoépliquer et anéantir l'humanité**.

24 Estimations de Tom Goldstein, professeur associé au département d'informatique de l'université du Maryland.

25 Solutionniste : partisan du solutionnisme qui est une façon de voir ou d'espérer la solution à des problèmes sociaux ou écologiques, etc. uniquement grâce à la technique.

Ailleurs que chez OpenAI, les débats sur la régulation de l'IA se multiplient. *Le Future of l Life Institute*, ONG à l'origine de la lettre signée au printemps par des chercheurs demandant une « *pause* » des recherches en IA[26], accuse OpenAI et d'autres entreprises du secteur de double discours sur la régulation. Elle **reproche à Sam Altman d'avoir tenté d'adoucir le règlement AI Act**[27] en menaçant de se retirer d'Europe, avant de se rétracter.

« *Sam Altman va avoir plus de mal à convaincre qu'il travaille pour le bien de l'humanité et pas pour ses actionnaires* », a écrit l'éditorialiste de *Bloomberg* Dave Lee, qualifiant l'**autorégulation sur l'IA de « *mascarade* »**. Des ONG voient d'ailleurs dans la crise d'OpenAI une raison de ne surtout pas assouplir les projets de règlement comme l'AI Act européen. D'autres, comme les fondations à but non lucratif Signal, Mozilla ou Wikipédia, y voient une **interrogation sur la gouvernance de l'IA**. « ***Les promesses parlant "d'IA bénéfique", "pour l'humanité" ou "alignée sur nos valeurs" omettent que ces questions sont politiques, avec des désaccords et des compromis***. *Ne faisons pas semblant que l'IA change quoi que ce soit à cela* », a écrit sur X l'ex-députée européenne Marietje Schaake, en réaction au feuilleton d'OpenAI.

4.2 Impacts des LLM sur le climat

4.2.1 Énergie nécessaire pour les modèles *Transformer*

Les grands modèles de langage, tels que GPT-3, ont été **critiqués par des chercheurs en éthique de l'IA de Google pour l'impact environnemental de la formation et du stockage des modèles**, détaillé dans un article co-écrit par Emily Bender, Timnit Gebru, Angelina McMillan-Major et Margaret Mitchell publié en 2021 : « Sur les dangers des perroquets stochastiques[28] : les modèles de langage peuvent-ils être trop grands ? »[29]. En effet, **le pré-entrainement de tous les modèles *transformers* est très lourd en ressource, notamment en électricité et donc en émission de CO_2.** Les chercheurs d'OpenAI ont expliqué que « *le pré-entrainement pratique à grande échelle nécessite de grandes quantités de calcul, qui sont gourmandes en énergie : l'entrainement du modèle*

26 **Le but de cette « *pause* » est d'élaborer de meilleurs garde-fous pour ces logiciels, jugés « *dangereux pour l'humanité* »**. « *Ces derniers mois ont vu les laboratoires d'intelligence artificielle s'enfermer dans une course incontrôlée pour développer et déployer des cerveaux numériques toujours plus puissants, que* ***personne – pas même leurs créateurs – ne peut comprendre, prédire ou contrôler de manière fiable***. *[…] Devons-nous laisser les machines inonder nos canaux d'information de propagande et de mensonges ? Devrions-nous automatiser tous les emplois, y compris ceux qui sont gratifiants ? Devons-nous développer des esprits non humains qui pourraient un jour être plus nombreux, plus intelligents, et nous remplacer ?* », ont écrit les signataires.
Selon la lettre, il faut créer des « *systèmes de gouvernance* », avec une surveillance des grands modèles, un « *système robuste d'audit externe et de certification* », ainsi qu'une « *responsabilité pour les dommages causés par l'IA* ». Il faut aussi pouvoir « *distinguer* » les contenus réels ceux créés par l'IA. OpenAI y travaille mais ces solutions ne sont ni obligatoires ni infaillibles. Les auteurs appellent enfin à **« *tracer les fuites* » de modèles**, c'est-à-dire la publication non autorisée de logiciels d'intelligence artificielle, une mésaventure arrivée à Meta.

27 L'AI Act est un règlement européen voté le 8 décembre 2023. Il introduit une législation commune et un cadre légal à l'IA. Il s'étend à tous les secteurs (excepté militaire) et tous les types d'IA. Il **régule les fabricants de systèmes d'IA (imposant aux grands modèles d'avoir été entrainés sur données de bonne qualité, sans biais discriminatoires, sans violation de la législation sur le droit d'auteur…) et les entités les utilisant professionnellement (obligation d'afficher sur les créations qu'elles l'ont été de manière artificielle…)**. Il doit encourager l'innovation en Europe mais tout en limitant les éventuelles dérives de l'IA par le biais de sanctions économiques : des amendes pouvant aller jusqu'à 7 % du chiffre d'affaires, et au minimum 35 millions d'euros pour les infractions les plus graves.

28 Probabilistes.

29 Fin 2020, Google s'est séparé de Gebru, qui était co-responsable technique de l'équipe d'intelligence artificielle éthique, sans doute à cause de cet article critiquant la façon de produire des modèles de langue.

GPT-3 de 175 B de paramètres a consommé ***plusieurs milliers de pétaflop/s-jours[30] de calcul pendant le pré-entrainement****, contre des dizaines de pétaflop/s-jours pour un modèle GPT-2 de 1,5 B de paramètres.* ***Cela signifie que nous devons être attentifs au cout et à l'efficacité de tels modèles****.*

L'utilisation d'un pré-entrainement à grande échelle permet également d'envisager l'efficacité des grands modèles sous un autre angle *: nous devons non seulement* ***tenir compte des ressources nécessaires à leur entrainement, mais aussi de la manière dont ces ressources sont amorties tout au long de la durée de vie d'un modèle****, qui sera ensuite utilisé à diverses fins et affiné pour des tâches spécifiques.* ***Bien que des modèles comme le GPT-3 consomment des ressources importantes pendant la formation, ils peuvent être étonnamment efficaces une fois formés*** *: même avec le GPT-3 175B complet, la génération de 100 pages de contenu à partir d'un modèle formé peut couter de l'ordre de 0,4 kW-h, soit quelques centimes seulement en couts énergétiques. En outre,* ***des techniques telles que la distillation de modèles peuvent encore réduire le cout de ces modèles****, ce qui nous permet d'utiliser des modèles plus efficaces, en nous permettant d'adopter un paradigme consistant à former des modèles uniques à grande échelle, puis à créer des versions plus efficaces de ces modèles pour les utiliser dans des contextes appropriés.* ***Les progrès algorithmiques peuvent également augmenter naturellement l'efficacité de ces modèles au fil du temps****, à l'instar des tendances observées dans la reconnaissance d'images et la traduction automatique neuronale****.*** »

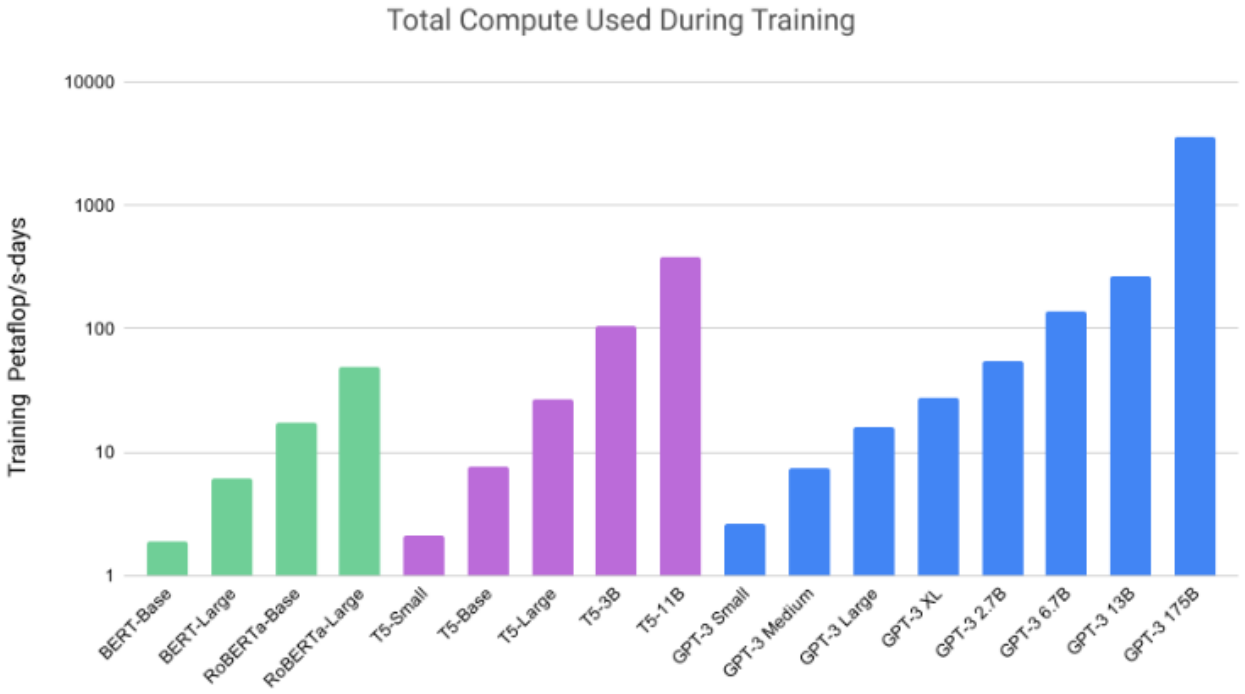


Figure 2.2: Total compute used during training. Based on the analysis in Scaling Laws For Neural Language Models [KMH+20] we train much larger models on many fewer tokens than is typical. As a consequence, although GPT-3 3B is almost 10x larger than RoBERTa-Large (355M params), both models took roughly 50 petaflop/s-days of compute during pre-training. Methodology for these calculations can be found in Appendix D.

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

30 1 pétaflop = 1 million de milliards d'opérations par seconde.

Un problème, c'est que la publication de ChatGPT fin novembre 2022 a déclenché une **course à l'IA. Cela a donc forcément augmenté les émissions de CO_2 liées à l'IA générative**. En revanche, difficile de savoir quel est son véritable impact climatique puisque les géants de la tech et les start-ups du secteur font preuve d'une grande discrétion sur leurs modèles.

4.2.2 Utilité pour le climat

Cela ne signifie pas que l'IA ne présente aucun intérêt pour le climat. **Elle peut contribuer à résoudre certains problèmes liés à la transition énergétique**. Des modèles d'IA sont spécialisés dans les **prédictions météorologiques à très court terme, ce qui permet de prévoir la production d'énergie d'origine éolienne ou solaire**. Et donc de maintenir l'équilibre entre l'offre et la demande d'électricité sur le réseau électrique.

Cet été, la Californie a, par ailleurs, entrainé un modèle d'IA à r**epérer les incendies avant qu'ils se transforment en fournaises**. « *Cela a absolument amélioré notre temps de réponse* », affirme Phillip SeLegue, le chef du renseignement de *Cal Fire*. Dans deux douzaines de cas, l'IA a signalé des débuts d'incendie qui seraient passés inaperçus. Ce qui a permis de les éteindre sans dommage.

En revanche, il n'est **pas certain que les larges modèles de langage soient très utiles dans la lutte contre le changement climatique**. Des start-ups, dont la française Ekimetrics, ont certes lancé des chatbots visant à répondre aux questions du public sur la crise climatique. Mais leurs réponses ne sont pas toujours fiables à 100 %, ce qui pourrait contribuer à semer le doute.

4.3 Hacking des robots conversationnels

C'était à prévoir, tant ChatGPT est sollicité un peu partout dans le monde depuis sa sortie : il a subi un certain nombre d'attaques. Ainsi,

- dès janvier 2023, des **hackers russes ont déclaré s'être emparés de ChatGPT** ;
- en juin 2023, le spécialiste cyber singapourien Group-IB a révélé avoir identifié un peu **plus de 100 000 comptes ChatGPT d'OpenAI volés par des hackers** ayant utilisé l'*info stealer* Raccoon, un cheval de Troie souvent diffusé par des liens dans des campagnes de phishing. L'attaque s'est étendue de juin 2022 à mai 2023 et a touché toutes les régions du monde. **Les données se sont retrouvées sur le dark web**. Selon Group-IB, ces piratages représentent un **risque majeur pour les données d'entreprise**. En effet, de plus en plus de professionnels utilisent le chatbot dans le cadre de leur travail. Par défaut, ChatGPT conserve l'historique des requêtes des utilisateurs, **exposant ainsi les organisations à la divulgation de données potentiellement sensibles qui peuvent être exploitées pour des attaques ciblées à leur encontre** ;
- en novembre 2023, il y a eu une attaque par DDos ; etc.

4.4 LLM : transformation du processus d'écriture

Les grands modèles de langage d'OpenAI GPT-3.5 et GPT-4 font partie du *natural language processing* (NPL) qui est une forme d'IA permettant le dialogue entre les humains et les machines. **Sa particularité c'est que son interface à usage général « *text in, text out* » peut compléter presque n'importe quelle tâche**, au lieu du cas d'utilisation unique habituel.

4.4.1 Impacts de la confusion entre écrits humains et écrits robotiques

Dans une approche sans précédent, les chercheurs d'OpenAI dans un long article, « *Language Models are Few-Shot Learners* » (2020), expliquent que « *GPT-3 a le potentiel de faire progresser les applications bénéfiques et néfastes des modèles de langage* ». En effet, la qualité du texte généré par ces grands modèles de langage est si élevée qu'**il peut être difficile de déterminer s'il a été écrit ou non par un humain, ce qui présente à la fois des avantages et des risques**. Dans la conclusion de leur article de novembre 2020, « GPT-3 : sa nature, sa portée, ses limites et ses conséquences », Luciano Floridi et Massimo Chiriatti expliquent que **malgré ses lacunes GPT-3 écrit mieux que beaucoup de gens** – et GPT-4 a encore progressé. Sa disponibilité représente l'**arrivée d'une nouvelle ère dans laquelle nous pouvons maintenant produire en masse des artefacts sémantiques de qualité et bon marché** : traductions, résumés, procès-verbaux, pages Web, articles de journaux, guides, publicités – en particulier les pièges à clics de toutes sortes –, etc. L'inconvénient, c'est qu'**une bonne partie de ces textes – articles, romans, etc. – seront de piètre qualité. Cela risque donc de noyer les textes pertinents et de qualité**.

4.4.1.1 Aide apportée par ces LLM

- les personnes dont le travail consiste encore à écrire seront **de plus en plus soutenues par des outils tels que GPT-3 et GPT-4** ;
- les chatbots et la gestion des connaissances pourront **améliorer les relations entre les consommateurs et les producteurs, les clients et les entreprises**.

4.4.1.2 Risque de confusion informationnelle

- ces chatbots **facilitent les arnaques** en permettant de créer efficacement spams, phishing et ingénierie sociale avec faux-semblant ou d'aider des hackers à optimiser leurs attaques en produisant du code malveillant ;
- la **rédaction frauduleuse d'essais universitaires, y compris scientifiques, risque de saper la confiance du public dans l'entreprise scientifique et de rendre plus difficile la distinction entre les vraies et les fausses informations, jetant ainsi le doute sur les contenus légitimes**[31] ;
- il existe le même problème pour les informations politiques, géopolitiques... C'est pour cela que **selon le *Global Risks Report* (« Rapport sur les risques mondiaux ») publié le 10 janvier 2024, en amont du Forum économique mondial de Davos, la désinformation est actuellement l'un des plus grands risques pour l'humanité** : « *alors que la polarisation s'accentue et que les risques technologiques ne sont pas maitrisés, la "vérité" sera mise à rude épreuve* » surtout en 2024 où quasi la moitié de l'humanité doit voter pour choisir un nouveau président ou un nouveau Parlement. « *L'**utilisation généralisée de mésinformations et de désinformations***[32], *ainsi que des*

31 Guillaume Cabanac, enseignant-chercheur en informatique, a débusqué en septembre 2023 dans des revues scientifiques reconnues des articles où se trouvaient les expressions « *Regenerate Response* » ou « *en tant qu'IA, je ne peux pas répondre à votre question* ». Cela signifie que **certains scientifiques ayant utilisé ChatGPT pour rédiger leur article scientifique n'ont pas relu ce qu'avait généré l'IA avant de demander à le faire publier. Pire, que l'intégralité de la chaine de publication a failli**. **Ce qui pose la question du contenu même des articles scientifiques** comme l'explique Guillaume Cabanac : « ***Lorsqu'on trouve une de ces erreurs dans l'article, que peut-on penser par la suite des expérimentations ?*** *[...] Y a-t-il vraiment eu une cohorte de personnes qui avaient telle ou telle maladie,* ***on ne peut qu'en douter****. On ne peut que penser que, si les auteurs ont fabriqué du texte avec une machine en introduisant des erreurs consternantes, leur méticulosité n'est pas au rendez-vous, on ne peut que penser qu'ils ont peut-être aussi trafiqué, arrangé la partie méthode, la partie résultat et c'est pour ces raisons-là que ces articles sont bien souvent rétractés par les maisons d'édition* ».

32 La mésinformation consiste à transmettre des informations fausses qui ne sont pas créées dans l'intention de nuire alors que la désinformation transmet des informations fausses délibérément créées pour nuire.

outils permettant de les diffuser, pourrait saper la légitimité des gouvernements nouvellement élus » **en trompant les gens et en jetant le doute sur les contenus légitimes peut créer des troubles, de la violence, de la censure, etc.**

4.4.1.3 Risque d'anthropomorphisme

Un des ingénieurs de Google qui travaillait sur LaMDA[33] a déclaré dans la presse en juin 2022 que cette IA était douée de sensibilité et réellement consciente. Google l'a licencié. Mais cela montre que « ***nous avons maintenant des machines qui peuvent générer des mots sans réfléchir, mais nous n'avons pas appris à cesser d'imaginer un esprit derrière eux*** », a déclaré Emily M. Bender, professeure de linguistique. La terminologie utilisée pour les grands modèles de langage, comme « apprentissage » ou même « réseaux neuronaux », crée une « *fausse analogie* » avec le cerveau humain, a-t-elle ajouté. Brian Gabriel, porte-parole de Google, ne nie pas la possibilité de l'existence future d'une IA sensible, mais selon lui « *cela n'a pas de sens de le faire en anthropomorphisant les modèles conversationnels actuels, qui ne sont pas sensibles. Ces systèmes imitent les types d'échanges que l'on trouve dans des millions de phrases* ». **GPT-3, comme les autres IA génératives n'est pas capable de raisonnement ; en effet, « *il n'a pas de modèle interne du monde, ni d'aucun monde* »**, explique Melanie Mitchell, professeure et auteure en 2019 de *Artificial Intelligence : A Guide for Thinking Humans*.

Le problème, c'est que « *nos esprits sont très, très bons pour construire des réalités qui ne sont pas nécessairement fidèles à un ensemble plus large de faits qui nous sont présentés* », a déclaré Margaret Mitchell, chercheuse travaillant sur l'IA éthique. « *Je suis vraiment inquiète de ce que cela signifie pour les gens d'être de plus en plus affectés par l'illusion* », ajoute-t-elle. Si les IA génératives sont largement disponibles, mais pas comprises, « ***cela peut être profondément préjudiciable à la compréhension de ce que les gens vivent sur Internet*** », a-t-elle précisé. C'est à cause de ce problème que nombre de personnes demandent à ChatGPT des diagnostics pédiatriques alors qu'il n'a pas été entrainé spécifiquement dans le domaine médical et qu'en janvier 2024, une nouvelle étude a montré qu'il donne 72 % de réponses erronées et 11 % de réponses incomplètes lorsqu'il est confronté à des cas théoriques d'enfants malades par une invite.

Dans une tribune de Timnit Gebru et Margaret Mitchell, anciennes codirectrices de l'IA éthique chez Google, ont déclaré que « ***les leaders de la soi-disant IA alimentent la propension du public à voir l'intelligence dans les systèmes actuels***, *vantant qu'ils pourraient être "légèrement conscients", tout en décrivant mal ce qu'ils font* réellement *[…]. Les entreprises technologiques ont affirmé que ces grands modèles ont des capacités de raisonnement et de compréhension, et montrent des capacités apprises « émergentes". […]* ***Ces récits ont pour motif le profit : les objectifs déclarés de nombreux chercheurs et entreprises de recherche en IA sont de construire une "intelligence artificielle générale"***, *un système imaginé plus intelligent que tout ce que nous avons jamais vu, qui peut accomplir n'importe quelle tâche qu'un humain peut faire sans relâche et sans salaire. Bien qu'****un tel système n'ait pas été réellement démontré comme faisable, sans parler d'un bien net****, les entreprises qui y travaillent amassent et étiquettent déjà de grandes quantités de données, souvent sans consentement éclairé et par le biais de pratiques de travail abusives.*

La poussée vers cette fin ***balaie les nombreux préjudices potentiels non traités des systèmes LLM****. Et* ***attribuer de la "conscience" à un produit implique que tout acte répréhensible est l'œuvre d'un être indépendant, plutôt que de l'entreprise*** *– composée de personnes réelles et de leurs décisions, et soumise à une réglementation – qui l'a créé* ». Une façon d'éviter d'être poursuivi pour atteinte au droit d'auteur[34].

33 LaMDA (pour *Language Model for Dialogue Applications*) est une famille de modèles de langage neuronal conversationnel développé par Google.

34 Depuis 2023, les ayants-droit multiplient les procédures contre OpenAI. En effet, **les informaticiens, les éditeurs et les auteurs affirment que leur travail est utilisé sans crédit ni compensation**. En revanche, **la start-up juge que la création des outils d'IA générative est irréalisable sans l'utilisation de contenus**

Par conséquent, « ***il faut examiner attentivement les implications éthiques de ces modèles et prendre des mesures pour s'assurer que les informations qu'ils génèrent sont exactes et fiables*** » explique David Bikard, microbiologiste, chercheur à l'Institut Pasteur. De ce fait, « ***nous avons désespérément besoin de moyens de différencier le texte écrit par l'homme et l'IA afin de contrer les utilisations abusives potentielles de la technologie*** », déclarait en 2022 Irene Solaiman, directrice des politiques de la start-up AI Hugging Face, anciennement chercheuse en IA chez OpenAI.

4.5 Deux inconvénients inhérents au *deep learning*

4.5.1 Boites noires et hallucinations

Yoshua Bengio, un des pionniers de l'IA, explique que **les LLM sont des systèmes qui s'apparentent à des « *boites noires imprévisibles***, *parce que [...] les réseaux de neurones qui font appel à l'apprentissage profond sont capables d'apprendre à partir des données qu'on leur a fournies, et sont entrainés à imiter la façon dont les humains conversent entre eux. Les résultats restent imprévisibles, et* ***nous ne comprenons pas parfaitement le fonctionnement de ce processus très complexe. Cela pose des questions de sécurité, de transparence, de fiabilité. On s'aperçoit très vite que ces systèmes inventent parfois des réponses, qu'ils peuvent aller dans une direction que n'ont pas choisie les programmeurs. La seule façon de les réorienter, c'est de les récompenser lorsque la réponse est pertinente. Depuis des mois, les concepteurs de ChatGPT travaillent à parer les coups, mais ce n'est pas suffisant*** ». Hugues Bersini, directeur du laboratoire d'IA de l'université libre de Bruxelles renchérit : « ***nous avons un devoir de comprendre, ne serait-ce que pour nous assurer qu'un jour, il ne parte pas en vrille*** ». De façon plus pragmatique, **la compréhension du fonctionnement des LLM limiterait voire éliminerait leurs hallucinations**.

C'est d'ailleurs **à cause de ces hallucinations que certains chercheurs réfléchissent déjà aux successeurs des IA génératives basées sur des LLM**. En novembre, Yann LeCun, qui travaille pour Meta, s'est emporté : « ***Les LLM, ça craint ! Il faut les faire entrainer par des humains et pourtant ils continuent à faire des erreurs ; l'IA générative est moins intelligente qu'un chat !*** » Pour lui, les LLM et les IA génératives ne constituent qu'une étape : « ***L'avenir de l'IA, ce ne sont pas les LLM, ce sera autre chose, une IA capable d'apprendre en regardant des vidéos.*** »

4.5.2 Biais des IA génératives

L'apprentissage s'effectuant sur un très large jeu de données – GPT-3 a scanné presque tout ce qui se trouve sur l'internet –**, il emporte avec lui tous les biais qui y sont présents et ils ne sont pas tous facilement identifiables**[35]. Ainsi, lors d'un apprentissage à partir d'une large base de données textuelle, s'il est envisageable de supprimer les propos explicitement racistes ou sexistes, il est par exemple moins évident de garantir l'absence de préjugés genrés concernant les professions d'aide à la personne ou celles considérées comme prestigieuses :

Traduction proposée par DeepL (25/01/24) – si on écrit juste « a nurse and a doctor », il propose également « infirmier », mais pas « une médecin » ni « une doctoresse »...

protégés pour les former.

35 Tariq Krim, spécialiste des questions d'éthique et de vie privée sur Internet, explique dans « ChatGPT : les nouveaux enjeux de l'IA » que ChatGPT a un modèle cognitif pensé essentiellement pour une cible américaine. Cela se perçoit quand on lui demande de répondre à une question comme le ferait un étudiant ou sur des questions de géopolitique, il donne les éléments de langage du département d'État américain. Il faut donc **se demander vers quel référentiel idéologique ces intelligences artificielles renvoient**.

Les chercheurs d'OpenAI ont attiré l'attention sur ces dangers pour demander des recherches sur l'atténuation des risques.

En médecine : problème de l'origine des données

Dans la médecine, l'IA aide. Mais parfois elle **se trompe quand il s'agit de faire des diagnostics pour les femmes**. En effet, **les algorithmes ont été programmés avec des données dont les médecins ne savent rien. Or, elles proviennent souvent d'hommes jeunes blancs. Donc ces IA ne sont pas efficaces pour les femmes qui peuvent avoir des métabolismes et des symptômes différents des hommes**. Ainsi, jusqu'à très récemment, une IA était utilisée par le service de santé britannique pour calculer les risques d'insuffisance rénale. Or il n'y a pas les mêmes symptômes chez les hommes et les femmes et les données provenaient des hôpitaux militaires américains. Elles étaient donc à 94 % d'origine masculine. L'IA ne pouvait donc pas calculer correctement le risque d'insuffisance rénale pour les femmes alors que cette maladie peut être mortelle.

Certains chercheurs demandent donc une **protection des données moins stricte** pour avoir autant de données venant de femmes que d'hommes.

En outre, beaucoup de médecins ne savent pas que les IA peuvent être sexistes et donc avoir des conséquences négatives sur la santé des femmes. **Il faudrait que leur formation change et qu'on y ajoute un module sur les données utilisées et les biais possibles**.

4.5.3 Deux attitudes face à ces biais

4.5.3.1 Limiter les biais grâce aux humains

C'est ce que font toutes les entreprises qui créent un chatbot. Problème : **cela peut être fait dans des conditions peu éthiques en employant des travailleurs du clic[36].** Ainsi, une enquête du *Time Magazine* publiée en janvier 2023 dévoile qu'Open AI alimente ChatGPT d'exemples signalés de discours haineux et de violences sexuelles, afin qu'elle sache détecter ces formes de toxicité et ne les laisse pas passer. Pour ce faire, OpenAI a fait appel à Sama, une entreprise sise à San Francisco mais qui emploie des travailleurs au Kenya. Ceux-ci doivent lire des textes sexistes et racistes ou décrivant automutilations, incestes ou contenus pédopornographiques et les classer selon leur type (racisme, violence…). Sur une journée de neuf heures, chaque travailleur doit ainsi lire entre 150 et 250 textes faisant chacun de 100 à 1 000 mots, et y signaler les passages sensibles, et sont pour cela payés entre 1,32 et 2 dollars de l'heure. D'autre part, **psychiquement marqués** par ce travail, plusieurs employés expliquent qu'il est très difficile de parler avec des « conseillers en bien-être » ou des médecins du travail[37]. **Ces faits contredisent la déclaration de principe d'OpenAI : *« la mission d'OpenAI est de s'assurer que l'intelligence artificielle générale profite à toute l'Humanité*** ».

4.5.3.2 Ouvrir les modèles

Ouvrir son modèle linguistique aux universitaires, à la société civile et aux organisations gouvernementales, c'est ce qu'a décidé de faire en mai 2022, Meta, la société mère de Facebook. Joëlle Pineau, directrice générale de Meta AI, a déclaré qu'il est impératif que les entreprises technologiques améliorent la transparence au fur et à mesure que la technologie est construite : « ***l'avenir des grands modèles linguistiques ne doit pas être uniquement entre les mains des grandes entreprises ou des laboratoires*** ». Yann Le Cun rappelle d'ailleurs que **« *jusqu'au [lancement de***

36 Les travailleurs du clic font de petites tâches répétitives comme trier ou annoter une base de données.

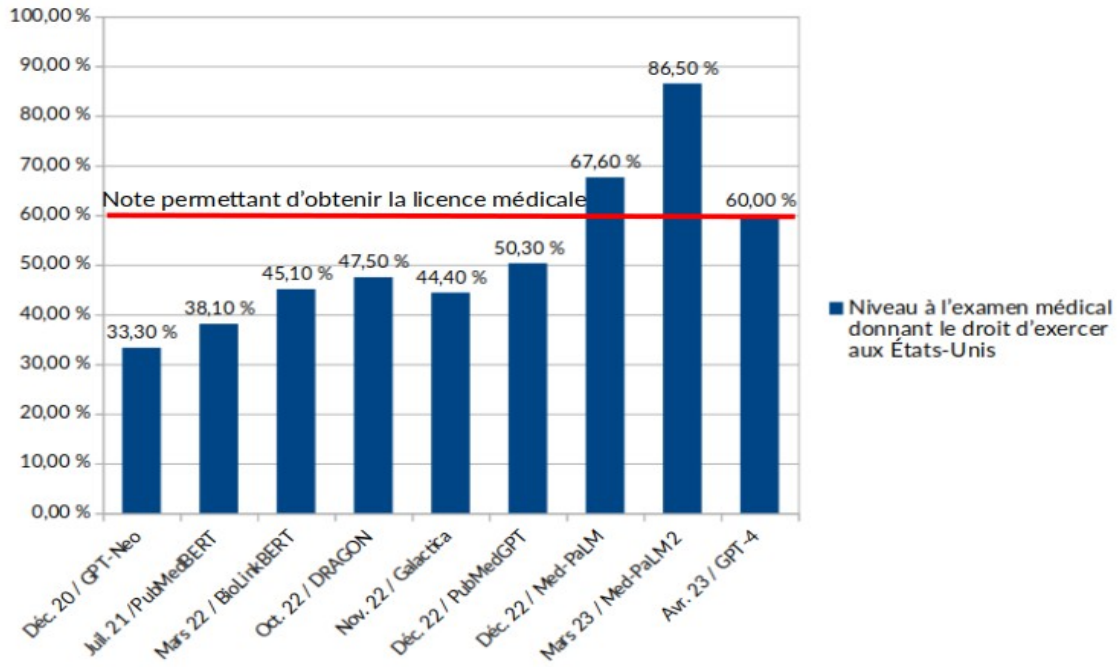
37 En octobre 2023, certains anciens employés de Sama ont déposé une requête devant le Parlement kényan pour tenter de réguler le secteur des entraineurs d'IA en pleine croissance mais parfois traumatisant.

ChatGPT], le domaine de la recherche et développement en IA était plutôt collaboratif. Les grandes entreprises comme Meta, Google ou Microsoft publiaient leurs travaux ».

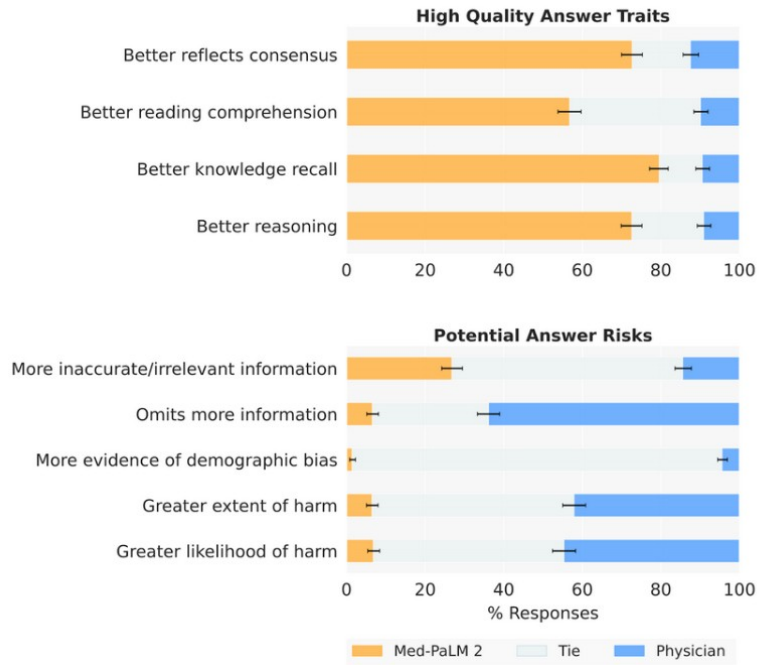
4.6 LLM et médecine : entre progrès et inquiétude

4.6.1 Progrès considérables

Les LLM peuvent aider les médecins et faire faire des progrès à la médecine, surtout, ceux entrainés spécifiquement dans le domaine de la médecine. Les progrès sont fulgurants.



Graphique élaboré d'après Google – il manquait GPT-4.



Dans une étude par pairs, les réponses Med-PaLM 2 ont été préférées aux réponses des médecins dans huit des neuf axes considérés.
(Schéma provenant de Google)

Ainsi, en janvier 2023, des chercheurs de Google et de DeepMind ont présenté un grand modèle de langage, appelé Med-PaLM, destiné à répondre à des questions médicales non spécialisées. Med-PaLM, basé sur le modèle de langage PaLM de Google avec 540 milliards de paramètres, a été entrainé sur le domaine médical et évalué à l'aide d'examens médicaux, de recherches médicales et de requêtes des consommateurs. Il a montré **des performances équivalentes, voire supérieures, à celles des médecins humains dans le diagnostic des affections respiratoires et cardiovasculaires**. En effet, **en plus de répondre avec précision aux questions à choix multiples et ouvertes, il fournit également un raisonnement et est capable d'évaluer ses propres réponses**.

Les capacités de génération des grands modèles de langage leur permettent de produire des réponses détaillées aux questions médicales des consommateurs. Cependant, **s'assurer que les réponses des modèles sont exactes, sûres et utiles a été un défi de recherche crucial, en particulier dans ce domaine critique pour la sécurité**. Les chercheurs de Google écrivent : « ***une attention particulière devra être accordée au déploiement éthique de cette technologie, y compris une évaluation rigoureuse de la qualité dans différents contextes cliniques avec des garde-fous pour atténuer les risques. Par exemple, les méfaits potentiels de l'utilisation d'un LLM pour diagnostiquer ou traiter une maladie sont beaucoup plus importants que l'utilisation d'un LLM pour obtenir des informations sur une maladie ou un médicament. Des recherches supplémentaires seront nécessaires pour évaluer les LLM utilisés dans les soins de santé pour l'homogénéisation et l'amplification des biais – la compréhension actuelle de la santé, de la maladie et de la physiologie humaines sont souvent teintées de discrimination fondée sur la race ou l'origine ethnique, le sexe, l'âge et les capacités – et des vulnérabilités de sécurité hérités des modèles de base*** ».

Depuis mi-2023, **« *l'IA est parfois capable d'identifier des maladies rares, peu ou pas connues des médecins, à partir de la description des symptômes.*** *On commence à voir aux États-Unis des familles qui ont obtenu un diagnostic de leur maladie rare par ChatGPT. Je pense que cela va contribuer à réduire l'errance médicale* », assure le Pr Jean-Emmanuel Bibault.

En janvier 2024, l'**IA de Google** « AMIE » (*Articulate Medical Intelligence Explorer*) **spécifiquement entrainée pour mener des entretiens médicaux afin de recueillir les antécédents du patient et en déduire un diagnostic a elle aussi mieux mené ces tâches que les médecins**. Pour les besoins de l'expérience, les chercheurs de Google ont fait appel à des acteurs formés pour se mettre dans la peau de vrais malades en simulant des scénarios cliniques prédéterminés. Chacun avait une liste de symptômes à présenter à l'IA et aux médecins, au cours de l'entretien via une interface de messagerie. Ces échanges montrent que

- **le robot obtient autant d'informations sur les patients que les médecins humains** ;
- **il fait preuve d'une plus grande empathie** selon les acteurs ayant joué les malades ;
- **l'IA seule est plus performante que le médecin pour établir un diagnostic** ;
- **si le médecin collabore avec elle, le duo sera meilleur que le médecin seul** ;
- en revanche, **le duo médecin plus IA est moins bon que l'IA toute seule**. Cela signifie que quand l'humain se fait aider par la machine, il dégrade les performances de cette dernière. « *Cela peut sembler surprenant, mais cela a déjà été observé dans des expériences de diagnostic radiologique.* ***Le médecin ne fait pas toujours confiance à l'IA. Quand elle arrive au même résultat, certains praticiens changent leur diagnostic*** », explique le Pr Jean-Emmanuel Bibault, cancérologue à l'hôpital européen Georges-Pompidou et chercheur en IA à l'Inserm.

Ces résultats sont particulièrement importants pour l'avenir de la médecine assistée par l'intelligence artificielle. Ils montrent qu'**il faut évaluer non pas les performances de l'IA seule, mais celles du médecin aidé par l'IA. Car la décision finale sera à priori toujours prise par lui**.

4.6.2 Ouverture des modèles pour lutter contre la méfiance

Des scientifiques de l'EPFL (École polytechnique fédérale de Lausanne) ont lancé en novembre 2023 **Meditron, un grand modèle de langage open source et adapté au domaine médical pour aider à la prise de décision clinique**.

Le fait qu'**il cible un domaine de connaissance spécifique rend ce modèle plus petit et plus accessible**. **Il peut potentiellement démocratiser l'accès à des informations fondées sur la science**.

Contrairement à MedPaLM et GPT-4, Meditron 7B et 70B est en open source. Ils comportent respectivement 7 et 70 milliards de paramètres. S'appuyant sur le modèle Llama-2 en

libre accès lancé par Meta, avec la contribution continue de cliniciennes et cliniciens ainsi que de biologistes, Meditron a été entrainé avec des sources de données médicales soigneusement sélectionnées. Ces dernières incluaient la littérature médicale évaluée par des pairs et issue de référentiels en libre accès comme PubMed, et un ensemble unique de directives de pratiques cliniques diverses, couvrant de nombreux pays, régions, hôpitaux et organisations internationales.

« *Après avoir développé Meditron, nous l'avons évalué par rapport à quatre points de référence médicaux majeurs, montrant que* ***ses performances dépassent celles de tous les autres modèles open source disponibles, ainsi que celles des modèles fermés GPT-3.5 et Med-PaLM. Meditron-70B est même à moins de 5 % de GPT-4 et 10 % de Med-PaLM-2, les deux modèles les plus performants, mais fermés, actuellement adaptés aux connaissances médicales*** », explique Zeming Chen, principal auteur de l'étude.

Dans un monde où la plupart des gens se méfient, voire redoutent, les progrès rapides de l'intelligence artificielle, le professeur Martin Jaggi, responsable du Laboratoire d'apprentissage machine et d'optimisation (MLO) de l'EPFL, souligne l'importance de la particularité open source de Meditron, y compris le code de sélection du corpus médical de préentrainement et les poids des modèles.

« ***Il y a une transparence sur la manière dont Meditron a été entrainé et les données qui ont été utilisées. Nous souhaitons que les chercheuses et chercheurs testent notre modèle et le rendent plus fiable et plus robuste grâce à leurs améliorations, en renforçant sa sécurité dans la validation en conditions réelles, un processus long et nécessaire. Rien de tout cela n'est disponible avec les modèles fermés développés par les grandes entreprises technologiques*** », explique-t-il.

La professeure Mary-Anne Hartley, médecin et responsable du Laboratory for intelligent Global Health Technologies, dirige les aspects médicaux de l'étude. « ***La sécurité était au centre de nos préoccupations dès le début de la conception de Meditron****. Ce qui est unique, c'est qu'il code les connaissances médicales à partir de sources transparentes d'informations de qualité. Il s'agit maintenant de* ***s'assurer que le modèle est capable de fournir ces informations de manière appropriée et en toute sécurité*** ».

« ***Nous avons développé Meditron car l'accès aux connaissances médicales devrait être un droit universel***, conclut Antoine Bosselut. *Nous espérons qu'il sera un point de départ utile pour les chercheuses et chercheurs qui souhaitent adapter et valider cette technologie en toute sécurité dans leur pratique.* »

4.6.3 Médecins décisionnaires

La machine n'est pas prête à remplacer les médecins :

- **elle souffre encore de limitations. Par exemple, dans cette expérience, les entretiens ont été menés sur de faux malades entrainés pour énoncer de manière claire et précise leurs supposés symptômes. Cela ne se passe pas comme cela dans une consultation réelle, où l'art du médecin est de mener un interrogatoire qui permet au patient de verbaliser au mieux ses symptômes.** « *Il faudrait maintenant tester cette IA avec de vrais patients. Par ailleurs, le diagnostic ne repose pas toujours sur ce seul interrogatoire. Il peut exiger un examen clinique, des examens biologiques, de l'imagerie, etc. C'est donc une première étape. Mais les prochaines couvriront probablement l'ensemble des données du patient* », prédit Jean-Emmanuel Bibault ;
- selon les chercheurs de Google, **ces LLM ne devraient pas remplacer les interactions avec les médecins, car la médecine va au-delà de la collecte d'informations, impliquant des relations humaines**.

Néanmoins, les auteurs estiment que **ces LLM pourraient éventuellement contribuer à la démocratisation des soins de santé**.

4.7 Qu'est-ce ce qu'une donnée de santé ?

Le Règlement Général de Protection des Données (RGPD) encadre le traitement des données de manière égalitaire sur tout le territoire de l'Union Européenne. Il est entré en application le 25 mai 2018. Il a été mis en place pour offrir aux usagers une meilleure protection de leurs données personnelles. Certaines de ces dernières jouissent d'une protection plus accrue, il s'agit des **données sensibles**[38]**, dont font partie les données de santé** que le RGPD définit. **Elles concernent les données relatives à la santé physique ou mentale, passée, présente ou future, d'une personne physique (y compris la prestation de services de soins de santé) qui révèlent des informations sur l'état de santé de cette personne**. Cette définition comprend par exemple :

- **les informations relatives à une personne physique** collectées lors de son inscription en vue de bénéficier de services de soins de santé ou lors de la prestation de ces services : un numéro, un symbole ou un élément spécifique attribué à une personne physique pour l'identifier de manière unique à des fins de santé ;
- **les informations obtenues lors du test ou de l'examen d'une partie du corps** ou d'une substance corporelle, y compris à partir des données génétiques et d'échantillons biologiques ;
- **les informations concernant une maladie,** un handicap, un risque de maladie, les antécédents médicaux, un traitement clinique ou l'état physiologique ou biomédical de la personne concernée (indépendamment de sa source, qu'elle provienne par exemple d'un médecin ou d'un autre professionnel de santé, d'un hôpital, d'un dispositif médical ou d'un test de diagnostic in vitro).

Cette définition permet d'englober **certaines données de mesure à partir desquelles il est possible de déduire une information** sur l'état de santé de la personne. Il existe **trois catégories de données :**

- **celles qui sont des données de santé par nature** : antécédents médicaux, maladies, prestations de soins réalisés, résultats d'examens, traitements, handicap, etc.
- **celles, qui du fait de leur croisement avec d'autres données, deviennent des données de santé en ce qu'elles permettent de tirer une conclusion sur l'état de santé ou le risque pour la santé d'une personne** : croisement d'une mesure de poids avec d'autres données (nombre de pas, mesure des apports caloriques…), croisement de la tension avec la mesure de l'effort, etc.
- **celles qui deviennent des données de santé en raison de leur destination, c'est-à-dire de l'utilisation qui en est faite au plan médical**.

La loi ne s'applique pas aux traitements qui comporteraient des données de santé à l'usage exclusif de la personne. À titre d'exemple, la loi ne s'applique pas aux applications mobiles en santé qui proposent dans leurs fonctionnalités, la collecte, l'enregistrement ou la conservation de données à condition que ces opérations s'effectuent localement sur un ordinateur, un ordiphone ou une tablette, sans connexion extérieure et à des fins exclusivement personnelles.

De plus, n'entrent pas dans la notion de données de santé celles à partir desquelles aucune conséquence ne peut être tirée au regard de l'état de santé de la personne concernée (exemple : une application collectant un nombre de pas au cours d'une promenade sans croisement de ces données avec d'autres).

---

38 Elles révèlent des informations liées à une personne concernée :

- l'origine ethnique ou raciale ;
- le traitement des données génétiques ;
- l'appartenance syndicale ;
- la vie sexuelle ou l'orientation sexuelle ;
- les données biométriques qui permettent l'identification d'une personne.
- les convictions religieuses ou philosophiques ;
- les opinions politiques ;
- la santé ;
- les données sur les infractions ou condamnations ;

Données de santé très encadrées

Les données de santé retenues comme telles sont soumises à un régime juridique soutenu, justifié par la sensibilité de ces informations personnelles. Le Règlement Général sur la Protection des Données **interdit le traitement des données sensibles, car leur traitement présente un risque plus important pour les droits et libertés des personnes concernées**. Malgré tout, cette notion est limitée : le RGPD prévoit dans ses articles 9 et 10 **des exceptions à cette interdiction de traitement**, entre autres :

- **si la personne concernée a donné son consentement explicite** au traitement de ces données à caractère personnel pour une ou plusieurs finalités spécifiques ;
- si le traitement est nécessaire aux fins de l'exécution des obligations et de l'exercice des droits propres au responsable du traitement ou à la personne concernée en matière de droit du travail, de la sécurité sociale et de la protection sociale ;
- si le traitement est nécessaire aux fins de la médecine préventive ou de la médecine du travail, de l'appréciation de la capacité de travail du travailleur, de diagnostics médicaux, de la prise en charge sanitaire ou sociale, ou de la gestion des systèmes et des services de soins de santé ou de protection sociale ;
- si le traitement est nécessaire pour des motifs d'intérêt public dans le domaine de la santé publique ;
- *si **le traitement est nécessaire à des fins archivistiques dans l'intérêt public, à des fins de recherche scientifique ou historique ou à des fins statistiques**, conformément à l'article 89, paragraphe 1, sur la base du droit de l'Union ou du droit d'un État membre qui doit être proportionné à l'objectif poursuivi, respecter l'essence du droit à la protection des données et prévoir des mesures appropriées et spécifiques pour la sauvegarde des droits fondamentaux et des intérêts de la personne concernée.*

4.8 Données de santé : contours d'une controverse[39]

Les données de santé sont l'objet d'une convoitise intense : gisement de connaissances aux yeux des chercheurs, opportunité de création de valeur pour des industriels du numérique, source de transparence pour les associations de patients… Mais les risques sont à la hauteur des promesses.

À l'heure où l'économie numérique ne parle que *big data* et intelligence artificielle, les données de santé sont l'objet d'une convoitise particulièrement intense. **Gisement sous-utilisé de connaissances aux yeux des chercheurs, vecteur de thérapies inventives pour une partie du corps médical, opportunité de nouvelles créations de valeur pour des industriels du numérique, innovations de services pour les *start-ups*, source de transparence pour les associations de patients…** Si le monde numérique est coutumier des promesses enflammées, elles prennent une coloration particulière dans le champ de la santé, en raison du caractère unique des données impliquées : **particulièrement sensibles, les données de santé parlent de notre intimité, de nos souffrances, de nos fragilités et appellent une protection particulière contre de possibles mésusages. Les risques sont à la hauteur des promesses, un contexte qui oblige tous les acteurs, à commencer par la puissance publique, à avancer sur un chemin étroit, entre enthousiasme et prudence.**

39 Extraits de « Données de santé : contours d'une controverse » de Valérie Peugeot paru dans *L'Économie Politique* le 7novembre 2018.

4.8.1 La mise en données du monde

La santé, comme tant d'autres domaines de l'activité humaine, s'appuie chaque jour un peu plus sur des dispositifs numériques et, ce faisant, génère massivement des données. La numérisation de **l'hôpital** n'est pas un phénomène récent, les débuts du programme de médicalisation des systèmes d'information – PMSI – remontent à 1982. Mais elle connait depuis quelques années une accélération, et tend vers le zéro papier : de l'aide-soignante au médecin chef, du biologiste au radiologue, tous sont ou seront demain outillés. En ville, les médecins s'équipent également : 96 % des **médecins généralistes** déclarent disposer d'un logiciel pour la gestion des patients et 86 % des **spécialistes** interrogés déclarent avoir accès au dossier patient informatisé au sein de leur établissement. Des outils et logiciels métiers qui sont autant de sources produisant des informations. Les **officines pharmaceutiques** n'échappent pas au phénomène, l'intégralité de la gestion de stocks étant informatisée. **Le patient devient également producteur de données médicales et de bien-être**. En nourrissant son dossier pharmaceutique, consultable par les officines, destiné notamment à éviter des incompatibilités entre prescriptions ; en utilisant des services en ligne comme cette application de « suivi des règles et de l'ovulation » téléchargée plus de 4,5 millions de fois sur Google Play ; en s'équipant de thermomètres et de balances connectées ; en partageant leur état de santé et les effets secondaires attachés à leurs traitements dans des communautés en ligne de patients atteints de maladies chroniques ; demain, en versant ces informations dans leur dossier médical partagé – DMP –, dont la Caisse nationale de l'assurance maladie (Cnam) commence le déploiement à l'automne 2018[40]… **Quant à la recherche médicale, elle recrute des cohortes toujours plus importantes en nombre, pour un suivi longitudinal toujours plus profond afin d'identifier de nouveaux phénomènes.** […]

4.8.2 Mutualiser pour faire parler les données

La constitution de bases de taille importante est indispensable à un traitement massif susceptible de dégager des informations et des analyses inédites. Ceci explique en partie la mise en place de nouveaux entrepôts de données, fruits d'une mutualisation de bases auparavant en silos. […]

Mais la mutualisation ne se limite pas à des partages à l'intérieur d'un même groupement d'institutions. Le CEA, par exemple, centralise des images (scans, IRM) de cerveaux provenant de laboratoires privés et publics, afin d'**obtenir une masse suffisante de données pour des projets de recherche sur les maladies** neurodégénératives. Par ailleurs, on voit actuellement émerger toute une série d'acteurs privés, appelés « c**ourtiers de données » (*data brokers*)**, qui proposent d'entreprendre ce travail de collecte, de regroupement et d'anonymisation de données, puis d'effectuer pour des tiers, le travail de fouille et d'analyse. Pour récupérer les informations, ils nouent des accords avec les producteurs de données que sont les établissements hospitaliers, les médecins, les officines. Ainsi, […] en France, deux entreprises, l'une française, OpenHealth Company, l'autre filiale états-unienne, Iqvia, ont noué des partenariats avec des pharmacies pour récupérer les informations sur les achats de leurs clients. En échange, elles livrent gracieusement aux pharmacies des tableaux de bord contenant des informations exploitables en marketing[41]. […] Elle[s] espère[nt] **constituer une base de données de santé et monétiser auprès de tiers le partage de ces données une fois anonymisées**.

---

40 Le DMP a été remplacé à partir de janvier 2022 par un nouveau service : « Mon espace santé », service public français qui permet à chacun de stocker et partager ses documents et ses données de santé de façon gratuite et sécurisée. Les données sont hébergées en France, protégées par l'Assurance maladie. Il est créé sous réserve du recueil du consentement du patient. Cet accord n'a aucune incidence sur ses droits à remboursement.

41 **IQVIA a noué un partenariat avec 14 000 pharmacies françaises (soit une sur deux), stocke et traite des données de santé de clients de pharmacies en France**. La CNIL qui avait donné un accord, lance cependant une enquête et des contrôles après la diffusion d'un reportage du magazine *Cash Investigation* en mai 2021.

4.8.3 Des usages médicaux avant tout…

Du côté des usages, difficile de démêler ce qui tient de la promesse à longue échéance ou de la simple hypothèse de recherche, du bénéfice à court ou moyen terme. Comme souvent en matière d'innovation numérique, l'inflation sémantique autour des horizons heureux a d'abord vocation à en construire les régimes de justification pour convaincre investisseurs privés et publics. Quitte à s'exposer à certaines déconvenues en situation réelle. Les discours-slogans autour de la médecine des « 4 P » – préventive, prédictive, personnalisée et participative –, reflètent bien cette pensée prête à l'emploi qui masque les différences de maturité comme les controverses. […]

En matière d'épidémiologie, les espoirs sont nombreux : en croisant des bases de données jusqu'ici en silos – les statistiques de gastro-entérites avec celle de qualité de l'eau ; celles sur les maladies neurodégénératives des agriculteurs avec les ventes de pesticides ; les informations sur la consommation d'antidépresseurs et d'anxiolytiques dans un territoire donné après un épisode climatique violent… –, **il s'agit de repérer des causalités encore insoupçonnées ou supposées, mais difficiles à prouver, et d'en tirer les conséquences**. Ces *big data* peuvent aussi nous **aider à comprendre les inégalités sociales de santé, en matière de mortalité et de handicap, au-delà des critères déjà reconnus,** comme la profession ou le niveau d'étude, et en incluant de nouveaux critères, comme l'exposition aux polluants atmosphériques. **Mieux comprendre les risques professionnels liés à l'exposition au bruit, aux agents cancérogènes, au stress… est également un enjeu de taille, alors que de nombreuses molécules utilisées sur le lieu de travail n'ont pas fait l'objet d'évaluation de toxicité**.

La pharmacovigilance devrait être un autre bénéficiaire majeur de ces traitements massifs de données, en permettant de repérer les effets secondaires des médicaments plus tôt, sans dépendre des laboratoires pharmaceutiques qui commercialisent les molécules **et sans attendre un scandale sanitaire**, comme cela a été le cas avec des médicaments tels […] la Dépakine[42].

Le diagnostic est également donné comme un des gagnants de l'application de l'IA aux données de santé. Des acteurs comme IBM et son IA Watson, Google avec son équivalent DeepMind et d'autres moins visibles se positionnent sur ce créneau : diagnostic oncologique à partir de scanners et d'IRM, diagnostic de déficience visuelle liée au diabète à partir d'une photo, qui vient d'être autorisé par la FDA[43], en sont quelques exemples.

Mais **c'est du côté de la génomique que l'inflation des promesses ne cesse d'enfler**. **Le cout des techniques de séquençage du génome ayant chuté drastiquement, de nouvelles thérapies dites de précision ou personnalisées se développent**, qui vont par exemple croiser des informations sur la spécificité génétique et biologique d'une tumeur, avec des informations liées à l'environnement et au mode de vie du malade. Ces informations, confrontées à celles de « n » patients, doivent aider les soignants dans leurs choix thérapeutiques, de manière à améliorer la performance des soins.

Toujours du côté de la génomique, certains n'hésitent pas à parler de médecine prédictive. La connaissance du génome d'un individu permettrait de détecter des risques de

42 La Dépakine est un médicament à base de valproate de sodium traitant l'épilepsie. Or, chez environ 10 % des utilisatrices enceintes, l'acide valproïque est responsable de malformations congénitales. De plus, 30 à 40 % des enfants nés de mères sous valproate ont un risque de déficit cognitif ainsi que d'autisme ou de troubles apparentés. Ces risques sont connus depuis les années 1980, mais les informations sur la notice de la Dépakine étaient peu alarmantes jusqu'en 2010. De ce fait, des femmes, bien qu'enceintes, ont pris ce médicament.
Sanofi a été mis en examen en 2020 pour « *tromperie aggravée* », « *blessures involontaires* », puis « *homicides involontaires* ». La même année, l'Agence nationale de sécurité du médicament et des produits de santé a aussi été mise en examen, pour blessures et homicides involontaires par négligence.

43 La *Food and Drug Administration* est l'agence fédérale qui a, notamment, le pouvoir d'autoriser la mise sur le marché des médicaments aux États-Unis.

pathologies (fragilité cardiaque, rupture d'anévrisme…), **permettant au patient d'adapter son style de vie et, le cas échéant, d'être traité en amont.** Aux États-Unis, l'entreprise 23andMe propose des tests génétiques, sur simple envoi d'un échantillon de salive. Le client se voit retourner des informations sur les origines géographiques de ses ancêtres et sur les risques qu'il encoure pour une dizaine de maladies (Alzheimer, Parkinson, maladie de Gaucher…).

4.8.4 Des patients acteurs

Les associations de patients sont aussi demandeuses d'un accès aux données de santé, afin d'exercer un travail de vigilance, voire d'alerte, par exemple en comparant l'accès aux soins à l'échelle des territoires, en mettant à jour des inégalités de traitements en fonction de la situation socioéconomique du patient, ou en détectant des prescriptions répétées inadaptées… […]

Quant au patient lui-même, l'usage de ses données de santé pour son propre bénéfice reste encore à inventer. En effet, tant que le DMP ne sera pas largement diffusé et utilisé par les professions médicales, le patient ne disposera que d'informations parcellaires, souvent en format non numérique, quand l'accès ne lui en est tout simplement pas interdit. Ainsi, **aujourd'hui, le patient ne peut accéder à son dossier pharmaceutique. Pourtant cela lui permettrait, par exemple, de retrouver le nom de médicaments qui lui ont déclenché une allergie ou qui, au contraire, se sont révélés efficace dans le passé, pour tout simplement partager cette information avec son praticien.** L'enjeu n'est pas trivial : il conditionne la capacité du patient à nouer un dialogue plus équilibré et fructueux avec le corps médical, à réduire l'asymétrie de pouvoir entre médecin et malade. Comme le montre l'enquête ethnographique menée par la sociologue Dominique Pasquier auprès de familles modestes utilisatrices de l'Internet, le simple accès à des sites d'information médicale est déjà une source de « capacitation » pour ces dernières. Sans préjuger des usages à venir, gageons qu'un accès à la complétude des données devrait prolonger ce constat[44].

Si les données de santé devraient d'abord servir à l'amélioration de la qualité des soins, on ne peut ignorer qu'elles peuvent servir à d'autres finalités, notamment en matière de marketing médical pour les industriels du secteur. À titre d'exemple, connaitre la consommation de médicaments par zone de chalandise permet d'organiser les tournées des visiteurs médicaux pour cibler les médecins non-prescripteurs ; ou, à l'instar de n'importe quel secteur commercial, connaitre les profils des internautes permet de pousser de la publicité en ligne, toujours plus personnalisée. Les industries de santé consacrent de plus en plus de budget à leur marketing digital, en direction des prescripteurs ou des patients eux-mêmes.

4.8.5 Des risques multiples

On le voit, les données de santé sont générées de plus en plus massivement, collectées et stockées par un nombre croissant d'acteurs, utilisées à des fins qui ne cessent de se diversifier. Cette triple envolée n'est pas sans soulever de nombreux problèmes. Le premier est lié à la sécurisation des bases de données. Leur multiplication, leur dispersion entre les mains de

44 « En Seine-Saint-Denis, la fracture numérique creuse la fracture vaccinale », *Libération*, 11 février 2021 : En France, trois opérateurs (Maiia, Keldoc et Doctolib) ont été choisis par l'État pour orchestrer la campagne vaccinale contre la Covid-19 : chaque centre doit être accessible en ligne (donc sans critère de lieu d'habitation) via l'un des agendas proposés par ces entreprises. Or, dès le 18 janvier, date de l'ouverture de la vaccination aux plus de 75 ans, a afflué une « *patientèle inhabituelle* » pour le centre municipal de santé de La Courneuve, commune populaire de Seine-Saint-Denis. Plus riche, plus connectée qu'à l'accoutumée. Le centre a collecté le code postal des vaccinés : sur les 756 premières doses administrées en quinze jours, seulement 20 % des bénéficiaires étaient Courneuviens. […] « *Ces chiffres sont venus corroborer l'effet classe sociale qu'on pressentait* », se désole Julien Le Breton, médecin chargé du centre de vaccination de cette ville. « *Ces réservations en ligne sont déjà un peu contre-intuitives pour les plus de 75 ans, mais c'est carrément un non-sens pour les populations précaires de notre département* », complète le Dr Fabrice Giraux, responsable du centre de vaccination d'Aubervilliers.

nombreuses entreprises plus ou moins habituées à manipuler des données sensibles, induisent mécaniquement une augmentation des risques de fuite de données, par inadvertance – un sous-traitant technique peu exigeant – ou à la suite de manœuvres crapuleuses. Il se passe rarement plus de quelques jours sans qu'on ne découvre une faille de sécurité sur des logiciels clés ou, plus grave, qu'une fuite de données ne soit révélée[45]. […]Derrière ces fuites, une pratique : le vol de données avec exigence de rançon. […]

Mais les problèmes de sécurité sont également liés à une difficulté technique : celle de l'anonymisation des données. En effet, une majeure partie des usages évoqués sont effectués sur des données agrégées et anonymisées, de manière à protéger les patients concernés. Mais l'anonymisation d'une donnée est extrêmement difficile : il ne suffit pas de décorréler l'information médicale (la pathologie, la prescription…) du nom du patient. Il faut pouvoir s'assurer que d'autres informations associées ne permettront pas de remonter jusqu'à lui. Par exemple, si le patient souffre d'une pathologie lourde et qu'on dispose du nom de la petite commune où il demeure, il est extrêmement simple de l'identifier. C'est pour cela qu'on parle le plus souvent de « pseudonymisation » (lorsque l'identité a été enlevée mais avec une possibilité de réversibilité). **Or plus le nombre de bases de données accessibles est important, plus les croisements entre bases sont possibles, plus les risques de « réidentification » de données pourtant anonymisées sont grands.** Le gouvernement australien en a fait la pénible expérience en 2016, après avoir mis en *open data* les données de remboursements de dépenses médicales anonymisées de 2,9 millions d'Australiens, couvrant une période allant de 1984 à 2014. La publication scientifique en ligne *ScienceX* a rapidement démontré que le processus pouvait être inversé, et s'est amusée à réidentifier sept célébrités australiennes, dont trois députés et un footballeur[46].

4.8.6 Derrière l'atteinte à la vie privée, la discrimination

De façon générale, l'accès à des données aussi sensibles que les données médicales par des tiers qui ne sont pas supposés en connaitre le contenu est en soi une atteinte à la vie privée,

45 Il suffit de faire un tour sur le site https://www.cyberveille-sante.gouv.fr pour s'en rendre compte.
Le 23 février 2021, *Libération* a révélé que « **les informations confidentielles de 500 000 patients français [avaient été] dérobées à des laboratoires et diffusées en ligne** […] Le fichier comporte à chaque ligne, jusqu'à 60 informations différentes sur une même personne : numéro de Sécurité sociale, date de naissance, groupe sanguin, adresse, numéro de téléphone portable, médecin prescripteur, etc. Un commentaire précise parfois "Grossesse", "Levothyrox", "tumeur au cerveau", "séropositif HIV". […] Les personnes y figurant sont des cibles de choix pour du *phishing* personnalisé. [Ces] données personnelles compilées créent aussi des risques d'usurpation d'identité, de fausses ordonnances (qui peuvent utiliser les noms des médecins), de messages de détresse factices reprenant les problèmes de santé mentionnés, etc. ».

46 Voir « *Research reveals de-identified patient data can be re-identified* », *phys.org*, 18 décembre 2017.
D'autres exemples prouvent que l'anonymisation ou la dépersonnalisation n'écarte pas tous les risques :
• en 2002, l'expérience de Latanya Swenney réussit à ré-identifier des personnes à partir d'informations anonymisées (code postal, date de naissance, sexe) détenues par le responsable de l'assurance médicale des employés d'État au Canada ;
• en 2006 par Cynthia Dwork, chercheuse chez Micosoft, démontre qu'il est impossible d'assurer une protection complète des données dites « sensibles » dès lors que l'« attaquant » dispose d'informations annexes dites « quasi identifiantes » qui peuvent être aussi anonymes que l'âge, le lieu de résidence, le sexe ; voire se trouver compilées dans d'autres institutions. Croisement de données santé avec un fichier d'électeurs par exemple ;
• En 2014, le journaliste David Larousserie entamait la présentation du dossier qu'il consacrait dans *Le Monde* à la problématique des données sensibles par l'évocation d'un scénario d'un praticien hospitalier : « *Devant son écran d'ordinateur, un patron peu scrupuleux cherche à en savoir plus sur le dossier médical d'un employé fréquemment malade. Connecté au site web de l'entreprise* Health aware, *il renseigne non pas le nom de son salarié mais le nombre des hospitalisations, le mois et les durées de séjour. Il entre aussi le code postal de la ville de résidence, l'âge et le sexe de sa « cible » ; 0,023 seconde plus tard le service commercial a trouvé l'identité cherchée et, moyennant finance, livre la totalité des connaissances médicales sur l'employé* ».

avec la violence psychologique qui l'accompagne. **Mais de cette violation peut découler toute une série de conséquences très concrètes pour la vie des personnes**. Imaginons qu'un recruteur ait devant lui deux candidats aux qualités professionnelles équivalentes, mais qu'il sache que le premier a dans le passé été sujet à des incidents cardiaques et non le second… Imaginons un bailleur social qui puisse choisir ses locataires en fonction de leur état de santé… Imaginons qu'un banquier sollicité pour un emprunt, déjà en droit légalement de connaitre de notre situation médicale passée et présente, puisse de surcroit accéder au profil génétique du demandeur et que celui-ci révèle une forte probabilité de développer un cancer… Imaginons un assureur qui calculerait ses primes d'assurance en fonction de l'état de santé de ses clients ou qui, tout simplement, exclurait les clients à la santé jugée trop fragile…

Même si ces pratiques sont totalement illégales aujourd'hui en France[47]**, ces risques ne sont pas théoriques, loin s'en faut**. **Pour s'en convaincre il suffit de regarder du côté des assureurs**[48]. **À défaut d'accéder aux données de santé proprement dites, ils utilisent les données dites de « bien-être », issues de nos objets connectés (pèse-personne, montre…), qui en disent déjà long sur notre état général**. Ainsi, […] en France, Axa […] en 2014, [a] propos[é] des chèques cadeaux à ses clients qui acceptaient de marcher un certain nombre de pas par jour et de surveiller cette activité avec un appareil de la marque Withings. Tout cela bien entendu au nom de l'encouragement à une vie saine et à la réduction des risques sanitaires.

4.8.7 Réglementation : à la recherche de l'équilibre

En France, les données de santé bénéficient par défaut d'une protection renforcée depuis l'adoption de la loi informatique et libertés en 1978. Et en Europe, le règlement général sur la protection des données, ou RGPD, entré en vigueur en mai 2018, ne fait que confirmer ce principe et laisse aux États la liberté d'adopter un régime encore plus exigeant. Dans l'Hexagone, l'usage des données de santé fait l'objet d'un encadrement strict puisque, par défaut, est interdit « *le traitement des données génétiques, des données biométriques aux fins d'identifier une personne physique de manière unique, des données concernant la santé ou des données concernant la vie sexuelle ou l'orientation sexuelle d'une personne physique* ». **Pour autant, il ne s'agit pas d'empêcher toute réutilisation et la loi prévoit d'emblée une série d'exceptions au principe d'interdiction, dans lesquelles le consentement du patient joue un rôle fondamental : si ce dernier fournit un accord libre et éclairé, ses données peuvent alors être utilisées pour d'autres finalités que la raison première de leur collecte, sous réserve bien entendu du respect de normes de sécurité élevées. Lorsque le recueil du consentement n'est pas possible, l'utilisation des données doit se justifier par l'intérêt public et suppose une série d'autorisations et d'avis préalables**[49].

---

47 La loi de modernisation de notre service de santé du 26 janvier 2016 interdit explicitement l'usage des données du **système national des données de santé (SNDS)** pour « *l'exclusion de garanties des contrats d'assurance et la modification de cotisations ou de primes d'assurance d'un individu ou d'un groupe d'individus présentant un même risque* ».

48 Voir « *Health insurers are vacuuming up details about you – and it could raise your rates* », National Public Radio, 17 juillet 2018.
Cependant, en France, une personne voulant contracter une assurance-vie ou un prêt bancaire n'a aucun choix : **soit elle accepte l'ensemble des conditions générales, entre autres remplir un questionnaire de santé très précis, soit sa demande sera refusée**. Que devient alors le secret médical ? L'argument des assureurs est simple : il faut pouvoir apprécier les risques que l'on prend en charge. D'ailleurs, cela ne signifie pas qu'il est impossible de souscrire une assurance et donc d'obtenir un prêt si l'on n'est pas en bonne santé – la convention Areas a été créée pour cela le 6 juillet 2006 – simplement, le médecin de l'assurance peut demander des informations ou des examens complémentaires.
En pratique, la proposition d'assurance est faite aux conditions standards du contrat dans 58 % des cas (2010) ou alors l'assureur demande des surprimes (40 % des cas en 2010), voire des exclusions de garantie (2 % des cas).

En complément de ces dispositions qui demeurent relativement lourdes, la loi autorise la CNIL à élaborer des référentiels et des règlements types pour certaines catégories de données ou de traitements : la demande d'autorisation n'est pas nécessaire, une simple déclaration de conformité auprès de la CNIL suffit. Par ailleurs, toujours dans un souci d'encouragement à l'innovation et à la recherche, le gouvernement a choisi en 2016, dans le cadre de la loi de modernisation de notre système de santé, de mettre à disposition de tous en *open data* des extractions agrégées du SNDS et d'élargir l'accès à ce dernier, jusqu'ici uniquement ouvert à une poignée de chercheurs membres d'organisations publiques. Aujourd'hui, sous réserve de ne pas poursuivre des finalités interdites (promotion en direction des professionnels ou des établissements de santé ; exclusion des garanties de contrats d'assurance ou modification de cotisations ou de primes d'assurance pour un individu ou un groupe d'individus), les acteurs privés peuvent également accéder à ces données.

Ce régime, qui se veut d'équilibre, est pourtant loin de clore le débat, ou plutôt les débats, qui sont, de façon schématique, économiques d'une part et éthiques de l'autre.

4.8.8 L'intelligence artificielle pousse à la massification

On pourrait croire que la France dispose d'une masse de données de santé suffisante pour outiller tant la recherche que les industriels de santé. Mais ces derniers considèrent que **l'accès au SNDS demeure trop complexe, et souhaitent disposer de données toujours plus fraiches, quasiment en temps réel, et toujours plus abondantes**. Le rapport Villani publié en mars 2018, qui propose une stratégie française et européenne en matière d'intelligence artificielle (IA), conforte cette position. […]

Si cette course à la massification et à la « désanonymisation » ne peut qu'inquiéter pour les raisons évoquées précédemment, la stratégie industrielle qui sous-tend le projet est intéressante : il s'agit ni plus ni moins que de **contrer le déploiement intensif de géants états-uniens dans le champ de l'intelligence artificielle[50] et de rouvrir un espace compétitif pour des acteurs de petite ou moyenne taille qui pourront avoir accès à ces données.** Encore faut-il accepter la logique de mutualisation. […]

4.8.9 Enjeux éthiques

À côté de ces enjeux d'innovation et de politique industrielle, l'usage des données, que ce soit à l'échelle individuelle ou en traitement massif, soulève de nombreux enjeux éthiques, pour

49 Le responsable de traitement doit présenter un dossier de demande d'autorisation à la CNIL et, lorsqu'il s'agit d'un projet de recherche, il doit également obtenir préalablement un avis de l'Institut national des données de santé sur le caractère d'intérêt public du projet et, selon les cas, l'avis d'un Comité de protection des personnes (CPP) ou d'un Comité d'expertise pour les recherches, les études et les évaluations dans le domaine de la santé (Cerees).

50 Voir « E-santé : l'offensive estivale des Gafam », 7 septembre 2018, sur Ticpharma.com.
Début aout 2018, Microsoft, Amazon, IBM, Oracle, Salesforce et Google ont lancé un projet inédit dans l'e-santé : ils se sont lancés dans un consortium inédit pour faciliter l'interopérabilité des données de santé, avec le concours de l'association américaine de régulation des données de santé DirectTrust. Concrètement, ces entreprises entendent améliorer l'échange de données de santé entre elles, avec les patients et les institutions, grâce à l'IA et le cloud, au nom du « bien commun ». Derrière le vernis philanthrope, l'objectif pour ces industriels est également de lever les obstacles provoqués par la non-interopérabilité des données de santé pour rendre leurs technologies mieux adaptées et donc plus efficaces partout : à l'hôpital, dans les centres de soins ou chez le médecin traitant.
Julien Bisson, dans l'édito du n°432 du 1er février 2023 du *Un* explique qu'on peut « ***s'interroger sur la mainmise des géants du numérique sur ces technologies*** » et **demander si « *cette révolution scientifique [qu'est l'IA] sera un instrument de la liberté et de la connaissance ou alors une nouvelle forme de servitude à l'égard d'intérêts privés, qui ont plus souvent fait la preuve de puissance que de leur bienveillance*** »

lesquels les réponses restent en grande partie à inventer. **Côté individuel, ce sont les données génétiques qui sont au cœur de la controverse la plus vive. Ne serait-ce parce qu'elles révèlent des informations non pas sur un individu unique mais sur tous ceux qui ont en partage son ADN, ascendants, membres de la même cohorte familiale et descendants. Mais aussi et surtout parce qu'une prédiction n'est qu'une probabilité, et que la révélation de cette prédiction peut causer plus de mal que de bien.** Que faire lorsque l'on risque à 40 % d'être atteint d'une pathologie pour laquelle il n'existe aucun traitement préventif ? Doit-on partager cette information avec sa fratrie, sa descendance ? Cette connaissance ne va-t-elle pas être un facteur supplémentaire susceptible de déclencher la maladie ? Autant de questions qui constituent l'un des débats clés des États généraux de la bioéthique qui se sont déroulés en France de janvier à juin 2018. Le déploiement de tests « *direct-to-consumer* » (publicité et vente de produits médicaux auprès des patients plutôt que des médecins) comme celui de 23andMe amplifient ce phénomène : quel peut être le vécu de personnes recevant sans accompagnement par un professionnel un résultat de test leur annonçant qu'ils ont une forte probabilité de souffrir à l'avenir d'une pathologie grave ?

Côté mégadonnées, les interrogations liées à l'IA ne sont pas moins fortes. Ces questionnements ne sont pas spécifiques au champ médical, mais prennent ici une acuité particulière. Ainsi des questions de transparence des algorithmes. Comment s'assurer par exemple qu'une banque n'est pas en train de scruter massivement les relevés de ses clients, comportant des remboursements de frais médicaux, pour dresser le profil des personnes à risque, puis exclure sur cette base les clients considérés comme potentiellement insolvables ? Et des questions de responsabilité : qui sera responsable, par exemple, en cas d'erreur de diagnostic : le médecin, l'hôpital ou l'entreprise productrice de l'algorithme ? La CNIL a publié fin 2017 un rapport qui pose un diagnostic et ouvre de premières pistes d'actions[51]. Une contribution à des controverses qui ne cessent de se complexifier.

On l'aura compris, les données de santé constituent sans doute aucun un terreau fertile tant du point de vue scientifique que médical et économique. Mais les risques associés sont élevés et obligent à veiller à ce que les finalités d'innovation soient compatibles avec les exigences d'une société démocratique, soucieuse de préserver la vie privée et l'autonomie de ses citoyens. Cette ambivalence contraint à faire preuve d'une imagination renouvelée en matière technique et juridique, tant sur le recueil et la conservation que sur les traitements et les usages des données de santé.

51 Voir « Comment permettre à l'homme de garder la main ? Rapport sur les enjeux éthiques des algorithmes et de l'intelligence artificielle », CNIL, 15 décembre 2017.

SITOGRAPHIE

La quatrième partie a été créée grâce à Wikipédia ainsi que DeepL et Mate Translate, et principalement aux articles suivants[52] :

- **2018**
  « Qu'est-ce ce qu'une donnée de santé ? » de la CNIL, 8/01/18
  « L'IA chez Google : nos principes » de Sundar Pichai, 7/06/18
  « Données de santé : contours d'une controverse » de Valérie Peugeot dans *L'Économie Politique*, 7/11/18

- **2019**
  « Qu'est-ce qu'une donnée sensible et quelles sont les règles relatives à leur traitement ? » de The Neoshields, 9/10/19

- **2020**
  « *Language Models are Few-Shot Learners* » rédigé par les chercheurs d'OpenAI, 05/20
  « GPT-3 : sa nature, sa portée, ses limites et ses conséquences » de Luciano Floridi et Massimo Chiriatti, 2020
  « GPT-3 : un modèle d'intelligence artificielle prêt à l'emploi, capable d'écrire des livres, mais dépourvu de morale », rédaction de France Télévisions, 1/09/20

- **2021**
  « On the Dangers of Stochastic Parrots : Can Language Models Be Too Big ? », Emily Bender, Timnit Gebru, Angelina McMillan-Major et Margaret Mitchell, 1/03/21

- **2022**
  « Les éthiciens de l'IA ont averti Google de ne pas se faire passer pour des humains. Maintenant, l'un des employés de Google pense qu'il y a un fantôme dans la machine » de Nitasha Tiku pour *The Washington Post*, 11/06/22
  « Nous avons averti Google que les gens pourraient croire que l'IA a une conscience. Maintenant, c'est en train de se produire », tribune de Timnit Gebru et Margaret Mitchell dans *The Washington Post*, 17/06/22
  « L'intelligence artificielle, trop humaine pour être honnête » d'Erwan Cario pour *Libération*, 26/07/22

- **2023**
  « Piratage ChatGPT : des hackers détournent l'IA pour créer des malwares » de Maurine Briantais pour *Comment ça marche*, 18/01/23
  « Microsoft et ChatGPT : l'IA d'OpenAI à tous les étages » de Maurine Briantais pour *Comment ça marche*, 19/01/23
  « ChatGPT : les nouveaux enjeux de l'IA » émission de François Saltiel sur France culture, 20/01/23
  « Med-PaLM – A large language model from Google Research, designed for the medical domain » Google, 03/23
  « Yann Le Cun, directeur à Meta : "L'idée même de vouloir ralentir la recherche sur l'IA s'apparente à un nouvel obscurantisme" » de Claire Legros pour *Le Monde*, 28/04/23
  « Yoshua Bengio, chercheur : "Aujourd'hui, l'intelligence artificielle, c'est le Far West ! Nous devons ralentir et réguler" » de Claire Legros pour *Le Monde*, 28/04/23
  « ChatGPT : plus de 100 000 comptes piratés dont certains en France, ce que l'on sait » de

52 Si un article est réservé aux abonnés du journal, il est possible de le lire dans son intégralité sur Europresse. On peut y accéder depuis l'ENT, onglet « Les BU » puis deux fois « Ressources numériques ».

José Billon pour BDM, 21/06/23
« Large language models encode clinical knowledge » de Karan Singhal, Shekoofeh Azizi, Tao Tu, etc. pour *Nature*, 27/07/23
« Climat : les émissions cachées de l'intelligence artificielle » d'Hortense Goulard pour *Les Échos*, 10/09/23
« ChatGPT : de mauvais usages débusqués dans des articles scientifiques », d'Alexandra Delbot pour *France culture*, 20/09/23
« OpenAI dévoile GPT-4 Turbo : nouveau bond de puissance de l'IA, énorme upgrade pour ChatGPT » de Bastien L. pour *Le Big data*, 7/11/23
« OpenAI : aux sources de l'affaire Sam Altman » d'Alexandre Piquard pour *Le Monde*, 24/11/23
« Le grand modèle de langage de l'EPFL pour le savoir médical », EPFL, 28/11/23

- **2024**
« AMIE : A research AI system for diagnostic medical reasoning and conversations », Alan Karthikesalingam and Vivek Natarajan pour Google Research, 12/01/24
« L'IA de Google plus forte que les médecins » d'Olivier Hertel pour *Le Point*, 17/01/24