



L'intelligence artificielle, prolongement de la cybernétique

Cybernétique, informatique, intelligence artificielle, robot et cyborg

L'intelligence artificielle n'est pas un concept nouveau ; sa première forme fut la cybernétique, développée par l'Américain Norbert Wiener (1894-1964). On était alors au tout début d'un saut technologique mêlant l'électricité et ses applications (l'électrotechnique), les communications dans le cadre de l'émergence de l'électronique (les télécommunications), l'ingénierie avec l'apparition des systèmes de contrôle.

Le terme choisi pour le concept vient d'ailleurs du grec pour désigner le « pilotage » ; κυβερνητική (kubernêtikê) voulant dire « pilotage d'un navire ».

Pour faire très simple, on peut dire qu'on a une chose nouvelle alors : des machines, qu'on peut munir de transistors, ce qui fait qu'elles ont de la mémoire, qu'on peut leur donner des ordres pour la mise en action de différents appareils.

On peut demander des calculs, on peut demander des actions mécaniques : il suffit de penser ici aux usines de voitures des années 1980-2000, largement automatisées.

Norbert Wiener a expliqué les thèses fondamentales de la cybernétique dans *Cybernetics or Control and Communication in the Animal and the Machine* ; les mots-clés sont information et communication, puisqu'il s'agit de recevoir des informations et de les transmettre.

Dans cette approche, tout comme les êtres vivants interagissant avec leur environnement, avec des « entrées » et des « sorties », les machines peuvent désormais fonctionner pareillement.

Il y a un arrière-plan utopiste et humaniste ; Norbert Wiener fut par exemple végétarien et un farouche opposant à la vivisection. La cybernétique est née comme une utopie : celle de pouvoir rendre transparentes toutes les informations, y compris de nature industrielle.

On pourrait ainsi améliorer la production en répartissant mieux les ressources, mais c'est vrai pour la société également puisqu'on pourrait mieux répartir les moyens d'existence. En pratique, la cybernétique relevait pourtant au-delà de ces discours d'un saut technologique dans le cadre du capitalisme, avec de très grandes entreprises capables d'élever leur niveau technique.

La dimension utopique de la cybernétique, masquant sa finalité purement technique-productive, fit

d'ailleurs qu'elle devint ensuite la vision officielle du monde du social-impérialisme soviétique, alors que l'URSS socialiste de Staline la rejetait formellement comme une approche mécaniste et somme toute pseudo-scientifique.

Pourquoi cette critique, qui est celle effectuée par le matérialisme dialectique effectivement ? C'est que la cybernétique a une explication comptable de la pensée : il suffirait de tant d'informations, de tant de mémoires, de tant de mises en réseaux.

La dimension qualitative est entièrement absente de cette logique de pur calcul. C'est exactement comme si on se disait que la planification de l'économie était une pure technique et qu'il n'y avait pas de choix idéologiques à faire (l'URSS social-impérialiste est née justement du rejet de la dimension idéologique dans le plan).

La critique soviétique, de l'époque de Staline, à l'encontre de la cybernétique tient ainsi à dire qu'il s'agit d'une progression dans le calcul, mais qu'il faut savoir à quoi ça sert. Si c'est pour fabriquer des robots-tueurs pour le capitalisme, c'est réactionnaire.

Une telle critique prend indéniablement tout son sens à une époque bien différente, celle du début du second quart du 21^e siècle, où l'intelligence artificielle est justement utilisée pour l'utilisation améliorée de drones tueurs et de missiles.

L'opération israélienne contre Gaza a été marquée tout au long de 2024 par l'emploi massif de missiles, dont les objectifs ont été choisis à l'aide d'une intelligence artificielle. Début février 2025, Google a abandonné son refus que son intelligence artificielle puisse servir des objectifs militaires.

Cependant, il faut ici souligner la différence entre intelligence artificielle et cybernétique. Cette dernière visait à constituer une allocation idéale des ressources, c'est-à-dire que sa mise en place fonctionne en amont. On sait quel but on a.

Dans l'intelligence artificielle, le système est par contre mis en place sans savoir quel but on a.

Pour prendre un exemple, un emploi cybernétique d'un drone tueur ferait qu'on sait au préalable que la cybernétique va fonctionner pour un drone tueur. Ce n'est pas le cas pour l'intelligence artificielle, qui existe « en soit » et qu'on peut appliquer indépendamment à telle ou telle chose.

Au sens strict, c'est la différence entre un robot et un cyborg. Ce n'est pas la définition historique, mais c'est ainsi qu'il faut voir les choses, comme le montre de manière très paradoxale un simple film de science-fiction.

Normalement, un robot est en effet un être entièrement artificiel, alors qu'un cyborg est un être humain « amélioré » au moyen de l'électronique.

Il est absolument nécessaire de procéder à une modification de ces définitions, sans quoi on ne comprendrait plus la différence entre un robot permettant l'automatisation d'une action technique et

un cyborg procédant à des évaluations et s'auto-corrigeant.

Dans le film *Terminator* (1984), on a un robot recouvert de tissus organiques humains, ce qui ne change rien au fait qu'il reste un robot. Le film parle cependant de cyborg pour le désigner et c'est tout à fait juste si on prend en compte l'exemple suivant, qui est fondamental pour saisir l'intelligence artificielle dans sa différence avec la robotisation.

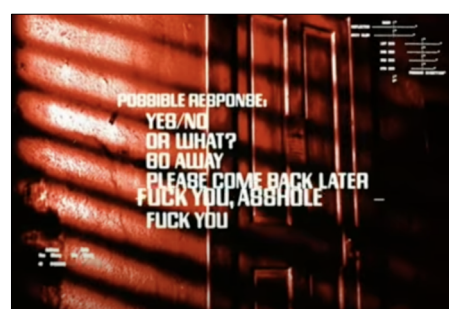
Le cyborg est dans un hôtel miteux de New York en train d'étudier un document devant l'aider à retrouver la cible qu'il doit éliminer. Un homme de ménage vient et tape la porte, disant que ça sent le « chat crevé ».



Le film montre les différentes options de réponse qui se proposent au cyborg et celui-ci en choisit une. Là est la différence fondamentale. Un robot n'aurait pas choisi : il a été programmé, la réponse tombe d'elle-même, selon des critères choisis au préalable.

Il fonctionne de manière mécanique : à telle entrée correspond telle sortie, à telle question telle réponse.

Le cyborg a quant à lui fait une relecture de la situation et a été amené à choisir la réponse adéquate ; sa programmation a été auto-corrigée. C'est ainsi que fonctionne l'intelligence artificielle : *en accordant de l'attention différemment selon les situations.*



C'est un excellent exemple pour saisir comment l'intelligence artificielle née au milieu des années 2010 ne relève pas d'une dimension automatisée. Si on n'a pas cela à l'esprit, on passe à côté de son fonctionnement et de son impact.

Le mot « attention » est la clef pour l'intelligence artificielle.

Du monde organisé d'Aristote à « Attention is all you need »

Au quatrième siècle avant notre ère, Aristote enseigna des points de vue qui firent de lui le plus important « philosophe » jusqu'à la fin du moyen-âge européen. Platon est bien plus connu, mais c'est parce qu'il défend le point de vue idéaliste, le point de vue religieux, alors qu'Aristote est un matérialiste.

Et Aristote ne s'est pas contenté d'étudier la réalité, de montrer comment aborder la physique, les espèces, le bonheur, etc.

Il a également affirmé qu'il était en mesure de résumer le principe commun à toute chose. Il le fait dans l'ouvrage qu'on connaît désormais sous le titre de « La métaphysique ».

Le terme *meta*, en grec, signifie « au-delà » ; on est donc « au-delà de la physique » ; ce n'est pas au sens strict le titre de l'œuvre, constitué à partir d'écrits retrouvés et éparpillés en différentes sections, mais cela correspond bien à sa vision du monde.

Il ne faut par contre pas faire l'erreur de croire que la métaphysique concerne quelque chose en dehors de la matière. Bien au contraire, la métaphysique d'Aristote est « au-delà » de la physique au sens où elle est commune à toute la physique : les choses peuvent être différentes, elles relèvent toutes fondamentalement du même principe.

Qui maîtrise la métaphysique ne connaît pas, s'il ne les étudie pas, les différents domaines de la physique, comme les animaux ou la météorologie, mais il en maîtrise au moins les fondamentaux, car il a la « clef » pour les comprendre.

Quelle est cette clef ? Pour résumer simplement les choses, Aristote dit que chaque chose s'appuie sur un moteur interne, prédéterminé, qui fonctionne mécaniquement. C'est en fonctionnant que ce moteur se réalise pleinement : l'oeil n'a de sens que par l'utilisation de la vision, l'échelle est faite pour être montée dessus, le gland doit se transformer en chêne, etc.

Cela sous-tend un monde « ordonné », qui existe pour lui-même, éternellement. Il n'y a pas selon Aristote de Dieu « créateur ». Il y a un Dieu « moteur », qui sert simplement à insuffler de l'énergie à la grande machine universelle qui tourne comme une horloge.

Comme tout tourne comme une machine, on ne peut pas « penser » au sens strict, on peut simplement réfléchir, constater les choses telles qu'elles sont.

C'est la thèse matérialiste et lorsque celle-ci est arrivée en Europe au Moyen-Âge par l'intermédiaire des philosophes arabes, la première chose que fait l'Église catholique c'est de formellement s'y opposer, dans le cadre d'une très fameuse interdiction de 219 thèses par l'évêque de Paris Etienne Tempier, le 7 mars 1277.

Dans la perspective matérialiste d'Aristote, « penser » n'est pas possible : ce qu'on peut faire, ce

qu'on doit faire, c'est réfléchir sur les choses pour comprendre leur mode de fonctionnement, leur *mode opératoire*.

Là où cela rejoint la question de l'intelligence artificielle, c'est qu'Aristote pose la thèse de l'intellect agent et de l'intellect patient.

Selon lui, on réfléchit, en effet, passivement : lorsqu'on le fait, on ne fait que redécouvrir une réalité qui était déjà là, sans avoir besoin de nous pour exister.

La réalité n'a pas besoin de nous : elle est présente, elle fonctionne selon des principes qu'on peut découvrir, mais qu'on ne les découvre éventuellement pas ne change rien à l'affaire.

Qu'on les découvre ne change en fait rien non plus : Aristote ne connaît pas le principe de transformation, qui arrivera par la suite avec les masses laborieuses du capitalisme. Par conséquent, pour lui, le bonheur absolu, c'est d'observer et de se perdre dans la contemplation du fonctionnement des choses et de la vie en général.

Cela étant, dans cette conception, l'univers est comme appuyé sur un « intellect agent » virtuel, qui contient tous les modes de fonctionnement de chaque chose, car ces choses existent, et lorsqu'on réfléchit bien, de manière correcte, juste, adéquate, alors on retombe sur le contenu de cet intellect agent qui est une sorte de bibliothèque virtuelle de la réalité, qui permet à celle-ci de « fonctionner ».

C'est une lecture panthéiste de la réalité, un matérialisme passif tourné vers la Nature, et on ne s'étonne pas ici que des penseurs musulmans se consacrant à la science se soient naturellement tournés vers Aristote.

On a dans l'Islam une Nature organisée, on a même un livre codé comme porte d'entrée à la compréhension de ce système : le Coran. Car l'objectif idéaliste de Mahomet, reflétant son absurdité mais également sa formidable grandeur, a été de « résumer » le monde en un ouvrage.

Ainsi, chez Aristote, et dans l'Islam si on adopte une lecture de cette religion en suivant Aristote (ce que feront les titans Avicenne et Averroès), si on réfléchit à tout, sur tout, avec tout... on rejoint une sorte de super réflexion virtuelle qui contient les connaissances de tous les modes opératoires différents.

Cela préfigure internet, où à partir de chaque ordinateur, on rejoint une sorte de gigantesque base de données.

C'est tellement vrai qu'Aristote a également étudié les « syllogismes », dont un est très connu : Socrate est un homme, les hommes sont mortels, donc Socrate est mortel. Il dresse le catalogue des syllogismes, et il dit quel syllogisme fonctionne pour aller à la vérité, quel syllogisme ne fonctionne pas.

Il a posé la logique comme base de la compréhension des choses, des phénomènes ; il expose une méthode scientifique pour savoir si une proposition se fondant sur des vérités est vraie, ou fausse, selon sa construction.

On arrive ici à la question clef de l'intelligence artificielle, car l'intelligence artificielle enregistre des propositions par millions et les combine, pour être soi-même en mesure de « parler », d'interagir à une communication.

« Attention is all you need »

Pour tirer à gros traits la situation, on peut dire la chose suivante. Les intelligences artificielles ont été, jusqu'en 2017, des programmes s'appuyant sur les syllogismes pour « lire » des textes et être capable de les enregistrer, pour ensuite être capable de les redire, de les reformuler.

Les intelligences artificielles s'appuyaient sur la grammaire d'une langue, voyaient comment les mots se reliaient entre eux, comment les phrases se construisaient. Ce processus était lent et laborieux, car il fallait beaucoup de calcul pour gérer toutes les phrases et qui plus est les phrases étaient analysées mot à mot.

Vint alors la crise de 2017. Cette année-là voit en effet la publication d'une courte étude, d'une dizaine de pages, écrite en commun par huit chercheurs de Google. Elle est intitulée *Attention is all you need*, soit « L'attention est tout ce dont tu as besoin », ce qui est une allusion à la chanson des Beatles « All you need is love » où il est dit de manière marquée à un moment : « Love is all you need ».

Le principe de l'attention est présenté comme permettant un fonctionnement bien plus efficace : il faut qu'en analysant les choses et en proposant des choses, l'intelligence artificielle accorde plus ou moins d'importance à tel ou tel aspect, c'est-à-dire soit capable de nuances.

Le résultat a été immédiat avec une efficacité bien plus grande ; on peut dire qu'historiquement *Attention is all you need* a révolutionné les intelligences artificielles, qui ont totalement changé de méthode d'approche, pour suivre le nouveau modèle proposé, dénommé « Transformer ».

Le résultat obtenu a été d'une telle ampleur que cela a abouti à l'irruption de l'intelligence artificielle comme thème dans les médias et les entreprises, avec comme intelligence artificielle la plus connue ChatGPT, présentée en 2022.

À la fin du premier quart du 21^e siècle, ces intelligences artificielles sont déjà incontournables pour beaucoup d'activités intellectuelles, en raison de leur dimension extrêmement utiles.

Il y a ici deux aspects : d'un côté, il y a une nouveauté fondamentale, qui est « l'attention » modifiée qui doit régir le robot analyseur et répéteur, qui devient intelligence artificielle.

De l'autre, il y a qu'une intelligence artificielle est en fait la superstructure d'internet et de son

immense accumulation de données, et rien d'autre.

Là est la limite.

Les différentes intelligences artificielles : RNNs et CNNs

Dans sa présentation, le document *Attention is all you need* explique qu'il propose de mettre en place un processus d'intelligence artificielle nouveau, et ce au moyen des « mécanismes d'attention, se passant totalement de récurrence et de convolutions ».

Ces mots semblent en tant que tel incompréhensibles. Voici comment il faut les comprendre.

Il existe trois types d'intelligence artificielle :

- celle relevant de réseaux dits récurrents (RNNs), fondés sur la récurrence ;
- celle relevant de réseaux dits convolutifs (CNNs), fondés sur la convolution ;
- celle relevant du modèle dit « Transformer » proposé par *Attention is all you need* qui dit qu'on peut se passer « totalement de récurrence et de convolutions ».

La révolution technologique de 2017 produite par *Attention is all you need* a fait que parler d'intelligence artificielle, c'est parler du modèle Transformer.

Présentons déjà les deux anciens types d'intelligence artificielle.

Les réseaux dits récurrents (RNNs) les plus connus sont :

- *DeepSpeech* qui fait de la reconnaissance vocale et est développé par Mozilla ;
- *WaveNet* qui fait de la génération vocale et est développé par DeepMind ;
- l'ancienne version de *Google Translate* ;
- *OpenAI GPT-1*.

On ne les utilise plus désormais que pour des systèmes embarqués où la puissance de calcul est limitée, pour la modélisation de séries temporelles (prévisions financières, météorologiques...), la génération de musique.

La reconnaissance vocale de Siri et d'Alexa utilise une telle intelligence artificielle ; c'est un très bon exemple, car lorsqu'on donne un ordre à Siri ou Alexa, les phrases sont courtes.

Ce type d'intelligence artificielle fonctionne en effet bien pour des phrases courtes, ou bien pour ce qui marche par séquence (d'où l'application capitaliste dans la musique, la bourse, la météo).

Le mode de fonctionnement, pour faire simple, est le suivant : on prend un mot, on l'analyse, on prend le second mot, on l'interprète et on le rapproche du premier mot, puis on fait pareil avec le troisième mot, et ainsi de suite.

On avance mot par mot, ce qui est long et implique d'avoir une « mémoire » prononcée afin de maintenir en contact les analyses effectuées – les connexions entre ces analyses sont dites récurrentes, elles se maintiennent à travers toute la boucle où on avance mot par mot.

Et le souci, c'est que plus la séquence est longue, plus les liaisons entre les mots commencent à être ingérables, l'intelligence artificielle perdant le fil ou bien s'accrochant à une liaison surinterprétée.

C'est d'autant plus un problème qu'il faut entraîner cette intelligence artificielle littéralement en la bourrant de phrases avec leur traduction, afin qu'elle se « souvienne » de quel mot va avec quel mot.

Ici, on est encore dans la démarche robotique : s'il se passe ça, alors il doit y avoir ça, et ainsi de suite pour toute une chaîne d'éléments. Plus la chaîne est longue, plus il faut une puissance de calcul toujours plus grande et une mémoire à la fois massive et réactive.

Les réseaux dits convolutifs (CNNs) étaient minoritaires avant 2017, mais ils sont encore relativement présents, car spécialisés sur la reconnaissance faciale et l'analyse d'image. Ils sont utiles également dans le domaine du son.

Le principe est ici d'utiliser un filtre pour réaliser une cartographie d'une image ou d'un son, afin d'en souligner les caractéristiques. On appelle convolution l'évaluation faite pour chaque zone de l'image.

On procède ensuite à la classification de l'image : il y a tant de chances que l'animal sur l'image soit un chien, tant de chances qu'en fait ce soit un loup, etc.

Il s'agit également de détecter des objets spécifiques, par exemple dans le cadre d'une voiture autonome qui doit éviter des piétons.

C'est ce type d'intelligence artificielle qu'on entraîne lorsqu'on doit répondre à un (insupportable) « captcha » où il faut cliquer sur un objet spécifique telle une voiture, des cyclistes, un pont, des feux de signalisation, un escalier, etc.

Les intelligences artificielles du type CNNs les plus connues sont :

- *LeNet-5* mis en place en 1998 et utilisé pour la reconnaissance de chiffres manuscrits ;

- *AlexNet* développé en 2012 avec un accent sur le « deep learning » ;
- *VGGNet* datant de 2014 et très performant ;
- *EfficientNet* développé en 2019 par Google ;
- *ConvNeXt* développé en 2022 par Facebook AI.

Transformer a une approche fondamentalement différente.

Transformer et l'embedding

Comment comprendre comment l'intelligence artificielle utilisant « Transformer » aborde une phrase ? On va prendre comme exemple « José aime danser la cumbia ».

La cumbia est un genre musical dans un esprit de danse, née en Colombie et ayant irradié toute l'Amérique latine avec de multiples variations, notamment la chicha (ou cumbia tropicale) au Pérou.

La première étape est le découpage de la phrase. Chaque mot devient un *jeton* pour la machine de calcul, en anglais un *token*.

On a ici cinq tokens : « José », « aime », « danser », « la », « cumbia ».

L'intelligence artificielle va ici attribuer des vecteurs à chaque mot. Cela donne par exemple [0.1, 0.2, 0.3] ; dans l'exemple donné ici, il y a trois vecteurs, mais en réalité, il y en a des centaines, des milliers.

Chaque vecteur correspond à une dimension : on peut penser à un nom, un verbe d'action, une description d'émotion, etc.

Cependant, il y a une erreur à ne pas commettre. On pourrait penser ici que les dimensions des vecteurs et le rapport entre ces vecteurs a été choisi rationnellement par des êtres humains, suivant différents critères : la grammaire, les familles de mots, les thèmes, les domaines concernés, etc.

Il n'en est rien. L'intelligence artificielle fonctionne ici comme une caisse enregistreuse, sauf qu'au lieu de nombres, on lui donne des phrases. On prend tout internet, on lui fait avaler cela, et l'intelligence artificielle établit les vecteurs propres à chaque mot selon les statistiques.

Dans l'exemple choisi, « aimer » et « danser » sont des verbes, mais le modèle Transformer ne les désigne pas comme tels : les textes qu'il a ingurgités les désignent comme des verbes, donc pour lui ce sont des verbes.

Il s'aperçoit statistiquement que ces verbes se modifient selon certaines situation – en pratique, ils

se conjuguent – et il prend cela en compte.

Pour faire la distinction entre « aime » et « danser », le modèle Transformer ne « sait » donc pas qu'il s'agit d'un verbe conjugué et d'un infinitif. Il va rapprocher statistiquement d'autres mots terminant de la même manière, placés au même endroit, utilisés de la même manière, etc.

C'est le premier aspect qu'il faut bien comprendre. L'interprétation d'un texte par l'intelligence artificielle, du type Transformer, se fait sur la base d'autres textes qui ont été au préalable analysés, et seulement ainsi.

Si dans la très grande majorité des cas, l'intelligence artificielle a constaté que le mot marteau est associé au mot clou, alors elle s'attendra à ce que le mot clou apparaisse avec le mot marteau, et elle-même emploiera le mot clou si elle parle d'un marteau.

Cela implique qu'une intelligence artificielle du type Transformer n'est pas en mesure de comprendre ou de répondre en s'alignant sur un mode de pensée qui n'a pas été assimilé par lui de manière au moins assez significative. Tout ce qui est nouveau, minoritaire, d'avant-garde, en développement inégal... reste incompris par lui. Mais il s'agira de revenir sur cet aspect.

Ce qu'il faut déjà retenir ici, c'est l'association à la fois statique et statistique. Le modèle Transformer va par exemple rapprocher « roi » de « reine », car il voit souvent les deux mots ensemble, et il va effectuer une sorte de grand balayage de tous les mots pour saisir les interconnexions entre eux.

Concrètement, des vecteurs sont attribués aux mots selon les interconnexions constatées dans l'assimilation de textes par l'intelligence artificielle, un processus appelé « embedding ».

Donnons deux exemples pour bien saisir le principe. Les chiffres correspondent à l'évaluation générale des mots dans la grande digestion d'innombrables textes, cela correspond à une sorte de vaste classement, ou plutôt de rangement.

Premier cas : on a « manger » et « dîner », deux mots qui sont liés à l'alimentation.

manger → [0.12, -0.45, 0.89, 0.33, -0.67, 0.22, 0.54, -0.11, 0.77, -0.29]

dîner → [0.10, -0.42, 0.85, 0.30, -0.65, 0.20, 0.50, -0.10, 0.75, -0.27]

Dans ce premier cas, les valeurs sont proches, ce qui signifie que ces mots sont souvent utilisés dans des contextes similaires.

Second cas : on a « manger » et « ordinateur », deux mots utilisés dans des contextes très différents.

manger → [0.12, -0.45, 0.89, 0.33, -0.67, 0.22, 0.54, -0.11, 0.77, -0.29]

ordinateur → [-0.78, 0.32, -0.44, 0.91, 0.15, -0.89, 0.03, 0.77, -0.56, 0.40]

Dans ce second cas, les valeurs sont très différentes, indiquant une forte distance.

En termes mathématiques, le calcul se fait avec le cosinus de l'angle entre les deux vecteurs :

$$\begin{aligned}\text{Cos}(\text{manger, dîner}) &\approx 0.95 \text{ (très proche)} \\ \text{Cos}(\text{manger, ordinateur}) &\approx 0.12 \text{ (très éloigné)}\end{aligned}$$

C'est ainsi que le modèle Transformer aborde les mots. Mais cet embedding n'est que la première étape.

Cependant, avant de voir la suite, il faut bien comprendre la chose suivante : le texte « José aime danser la cumbia » n'a pas encore été utilisé en tant que tel par l'intelligence artificielle !

Seuls les éléments constitutifs de la phrase ont été pris en compte et retrouvés dans la base de données, pour leur attribuer des valeurs issues de l'assimilation des textes et d'une évaluation statistique.

C'est absolument fondamental. Tout comme pour les libertariens, la société est simplement un agrégat d'individus et la souveraineté d'un ensemble dépend des individus le composant, *pour le modèle Transformer les mots sont des éléments uniques accordant indirectement de la valeur à un ensemble plus grand qui est la phrase.*

Les différentes intelligences artificielles : vers le mécanisme d'attention avec la vision libertarienne

L'étape qui suit l'embedding est ce qu'on appelle l'attention ; c'est là où on retrouve le document *Attention is all you need* puisque le titre insiste dessus, affirmant que c'est cela qui révolutionne l'intelligence artificielle.

Mais bien pour comprendre cette nouvelle étape, il faut saisir la portée de l'embedding, car le mécanisme d'attention qui le suit s'en veut l'inverse.

Les vecteurs de l'embedding considèrent en effet chaque mot en général. Plus simplement dit : le même mot est évalué pareillement tout le temps, quelle que soit la phrase étudiée.

Par contre, le mécanisme de « l'attention » prend chaque mot en rapport avec les autres mots de la phrase, et ceux des autres phrases s'il y en a. L'attention porte alors désormais sur chaque mot en particulier et non plus en général.

En apparence, cela sonne juste, mais c'est trompeur. Pourquoi ?

Voici le problème. Le matérialisme dialectique souligne la dignité du réel ; il faut partir du particulier pour aller au général, car le général est présent dans le particulier.

L'intelligence artificielle du modèle Transformer fait le contraire. Elle idéalise, rend abstrait un mot avec des vecteurs ; le mot « flotte » au-dessus du réel, il est statique.

On notera d'ailleurs qu'on parle d'un mot, mais on peut remplacer par n'importe quelle donnée, tels des sons, des images, des chiffres, etc. Dans tous les cas, il y a une évaluation statistique où l'on trouve des « éléments » uniques de valeur « éternelle ».

C'est très exactement la vision de l'idéaliste Platon contre le matérialiste Aristote, où il y aurait des « briques » constitutifs des choses. L'intelligence artificielle part du même principe qu'il y a un tout composé d'éléments séparés, bien distincts.

L'intelligence artificielle procède comme si le monde était composés de briques de Lego, ou bien de gros pixels volumineux comme dans le monde du jeu Minecraft. Il y aurait des briques, des éléments uniques, à la base duquel les choses se constituent après.

C'est contre cette conception que Mao Zedong a lancé le combat pendant la révolution culturelle en Chine dans la seconde moitié du 20^e siècle pour affirmer que « rien n'est indivisible », qu'il ne saurait exister de « briques » uniques comme point de départ, alors qu'en réalité tout est en interrelation dans un univers infini et éternel.

Mais ce n'est pas tout, cela va bien plus loin.

Dans la conception idéaliste, les briques sont statiques, elles ont un seul principe, elles sont ce qu'elles sont et rien d'autre. Pour Platon, il y a une seule « idée » pour la table, une autre pour la main, une autre encore pour l'épée, etc.

Ce n'est pas la conception de l'intelligence artificielle, qui dit que les briques constitutives des choses sont en mouvement, en activité. Autrement dit : les mots peuvent avoir plusieurs sens, des interactions différentes, des significations différentes...

Exactement comme un individu peut être fonctionnaire de police et saxophoniste, traverser la rue et manger des crêpes. Ce qui fait que la vision du monde de l'intelligence artificielle est en fait celle de l'anarcho-capitalisme, des libertariens.

Les mots sont les vraies choses, qui existent indépendamment du reste, et d'ailleurs des uns des autres. C'est seulement lors d'un éventuel rapport qu'on peut prendre en considération une forme supérieure comme la phrase.

Il n'y a pas de phrase, il y a des mots, qui s'avèrent être des phrases, tout comme pour les libertariens, il n'y a pas de société, mais des individus qui s'avèrent avoir des rapports sociaux.

La démarche est anti-synthèse, elle nie la fusion, l'imbrication dans des ensembles supérieurs. Elle est littéralement ultra-individualiste.

Voici quelques citations exemplaires d'auteurs relevant de l'approche libertarienne ; le rapport avec la vision du monde du modèle Transformer est évidente.

Murray Rothbard (1926-1995) :

« Seuls les individus existent, choisissent et agissent. "La société" n'est qu'une abstraction, qui n'a pas d'existence indépendante des individus qui la composent. »
(*L'homme, l'économie et l'Etat*, 1962)

Ayn Rand (1905-1982) :

« Une société n'a pas d'existence propre en dehors des individus qui la composent. Seuls les individus peuvent penser, agir, produire et créer. » (*La Vertu d'égoïsme*, 1964)

Robert Nozick (1938-2002) :

« La société n'a pas pour rôle de donner un but à l'individu ; elle doit seulement fournir un cadre où chaque individu peut poursuivre ses propres fins. » (*Anarchie, État et Utopie*, 1974)

Ludwig von Mises (1881-1973) :

« La société est l'aboutissement de l'action humaine, non d'un dessein humain. L'individu ne peut prospérer que dans une société libre où il interagit pacifiquement avec les autres. » (*L'Action humaine*, 1949)

Margaret Thatcher (1925-2013) :

« Et donc, ces personnes en viennent à reporter leurs problèmes sur la société. Mais la société, c'est qui ? Ça n'existe pas ! Il y a des hommes et des femmes, il y a des familles, et aucun gouvernement ne peut faire quoi que ce soit, si ce n'est à travers les gens. »

Friedrich Hayek (1899-1992) :

« L'idée que la société dans son ensemble puisse être dirigée selon un plan, de la même manière qu'une entreprise, est l'une des plus grandes illusions de notre époque. » (*Droit, législation et liberté*, 1973-1979)

On a ici précisément l'approche du mécanisme d'attention du modèle Transformer : *les mots sont ce qu'ils sont, les phrases sont des points de contact de mots et rien de plus.*

Et on peut aisément le prouver, puisque le modèle Transformer, dans les autres étapes d'analyse de texte, fait absolument tout justement pour éviter la phrase comme organisme complexe, comme

synthèse. Tout est fait, de manière obsessionnelle, pour en rester au mot.

Transformer et le mécanisme d'attention multi-têtes

Le mécanisme d'attention va étudier le rapport entre les mots d'une phrase. Sur quoi va-t-il se fonder ? Sur la phrase, qu'on va prendre enfin en compte ?

Pas du tout : le mécanisme d'attention provient des vecteurs eux-mêmes. On va regarder le rapport entre les mots de la phrase, mais en se fondant sur les informations données par les vecteurs.

On ne regarde pas la phrase, qui n'existe d'ailleurs en fait pas en soi : elle n'est là que comme assemblage d'éléments indépendants et au sens multiple, tels des individus dans la société capitaliste.

On part des vecteurs et on en assemble pour en former trois nouveaux :

- requête (query en anglais), soit Q : que recherche ce mot ?

Q représente la question que ce mot pose aux autres mots pour comprendre son contexte. Plus précisément, il sert à comparer ce mot avec les autres pour voir lesquels sont pertinents.

- clef (key en anglais), soit K : quelle information ce mot propose ?

K indique quelle information porte un mot, et l'utilité de cette information pour d'autres mots.

- valeur (value en anglais), soit V : quelle information sera transmise ?

V est le résultat de la combinaison de Q et K.

Reprenons notre exemple qui est « José aime danser la cumbia ».

Disons qu'on a trois dimensions, pour trois aspects : est-ce un verbe, un nom, un article, etc. ? Ce mot est-il important pour l'action principale ? A-t-il une relation forte avec d'autres mots ?

On est là pour l'instant hors de la phrase en tant que telle.

| Mot | Query / Q (Ce qu'il cherche) | Key /K (Ce qu'il propose) |
|------------|-------------------------------------|----------------------------------|
| José | [0.8, 0.2, 0.5] | [0.9, 0.1, 0.3] |
| aime | [0.6, 0.7, 0.8] | [0.7, 0.6, 0.9] |
| danser | [0.5, 0.9, 0.4] | [0.4, 0.8, 0.6] |
| la | [0.3, 0.2, 0.1] | [0.2, 0.3, 0.2] |
| cumbia | [0.7, 0.3, 0.9] | [0.8, 0.4, 0.7] |

Maintenant on va vers les éléments de la phrase, en prenant pour deux mots leurs Q et K.

Avec eux, on obtient V, soit un score qui va déterminer le « poids » pour le mécanisme d'attention. L'intelligence artificielle va accorder plus d'importance à certains rapports en fonction de ce poids.

Voici un exemple pour « aime ».

| Comparaison Q avec K | Calcul | Score d'attention |
|---|--|----------------------|
| aime avec José (Q aime · K José) | $(0.6 \times 0.9) + (0.7 \times 0.1) + (0.8 \times 0.3)$ = 0.95 | Lien fort |
| aime avec aime (Q aime · K aime) | $(0.6 \times 0.7) + (0.7 \times 0.6) + (0.8 \times 0.9)$ = 1.51 | (Lien avec soi-même) |
| aime avec danser (Q aime · K danser) | $(0.6 \times 0.4) + (0.7 \times 0.8) + (0.8 \times 0.6)$ = 1.34 | Lien fort |
| aime avec la (Q aime · K la) | $(0.6 \times 0.2) + (0.7 \times 0.3) + (0.8 \times 0.2)$ = 0.46 | Lien faible |
| aime avec cumbia (Q aime · K cumbia) | $0.6 \times 0.8) + (0.7 \times 0.4) + (0.8 \times 0.7)$ = 1.22 | Lien moyen |

Il faut bien comprendre que K et Q sont déterminés par l'embedding. Ils sont totalement indépendants de la phrase.

Autrement dit, *les poids d'attention ne sont pas influencés par la phrase étudiée en elle-même*. Ils sont issus de l'entraînement général du modèle sur un large corpus de textes.

Ils ne changent pas en fonction de la phrase analysée, ils ne sont pas recalculés ou ajustés pour chaque nouvelle phrase. *Ils ne dépendent pas de la phrase actuelle, mais du modèle entraîné sur un large corpus*.

C'est ici que commence le bricolage pour se rapprocher du sens de la phrase, sans jamais reconnaître la phrase en tant que tel.

Le modèle Transformer est anti-synthèse : il s'appuie sur des éléments-briques pour étudier un agrégat formé d'interactions de ces éléments-briques.

Contournant la phrase, qui est la qualité, le modèle Transformer cherche alors pour s'en sortir à s'appuyer sur la quantité. Le mécanisme d'attention est ainsi « multi-têtes ».

La démarche est alors de multiplier les K par les Q. Certaines données des vecteurs sont privilégiées par rapport à d'autres et on a alors 8, 12, jusqu'à 96 « têtes » d'attention consistant en des poids

obtenus au moyen de $K \times Q$.

Ces têtes visent par exemple les relations syntaxiques, celles au sein de la phrase, celles avec les autres phrases s'il y en a, la présence de pronoms, la place des mots, etc.

Tout cela n'est pas choisi en cours de route, mais a été enregistré au préalable dans l'intégration de millions et de millions de documents. En ingurgitant un nombre immense de textes, l'intelligence artificielle a appris à valider certaines têtes plus que d'autres. C'est purement statistique.

Prenons ici un exemple, avec la phrase « Le chat dort ». Imaginons qu'on ait eu un embedding de quatre dimensions.

| Mot | Embedding |
|------------|-----------------------|
| Le | [0.2, 0.8, -0.5, 0.3] |
| chat | [0.9, 0.1, -0.7, 0.4] |
| dort | [0.6, -0.2, 0.3, 0.7] |

Le Transformer ne comprend pas l'ordre des mots. On va alors avoir une tête qui réalise une analyse de position, qui a la même dimension que l'embedding.

| Position | Positional Encoding |
|-----------------|----------------------------|
| 1 Le | [0.01, 0.99, -0.02, 0.03] |
| 2 chat | [0.02, 0.98, -0.04, 0.06] |
| 3 dort | [0.03, 0.97, -0.06, 0.09] |

On additionne ensuite l'embedding et le Positional Encoding pour obtenir une représentation finale.

| Mot | Embedding + Positional Encoding |
|------------|--|
| Le | [0.21, 1.79, -0.52, 0.33] |
| chat | [0.92, 1.09, -0.74, 0.46] |
| dort | [0.63, 0.77, 0.24, 0.79] |

Chaque tête d'attention avec ses poids travaille de manière indépendante des autres, traitant des représentations différentes de la même phrase.

Ces têtes d'attention ne sont non plus pas hiérarchisées, aucune tête n'est plus importante que l'autre dans le sens strict. Elles sont toutes utilisées simultanément pour apprendre des aspects variés des relations entre les mots dans la phrase.

Le tout est alors empilé; cela appelle la concaténation. On a alors un vecteur avec entre 512 et 2304 « dimensions ». C'est ainsi que le mécanisme d'attention *enveloppe* la phrase, à partir de ses éléments, comme depuis l'extérieur.

Transformer, la propagation et la rétropropagation comme rejets de la négation

L'étape qui suit le mécanisme d'attention multi-têtes consiste en le Feed-Forward Network. Ici, nous avons la partie la plus complexe et la plus importante. En effet, il s'agit pour l'intelligence artificielle de faire semblant de comprendre que les choses ne se passent pas de manière linéaire.

Faire semblant, car seul le matérialisme dialectique peut saisir le principe du bond en avant, du saut qualitatif, de la rupture, de la révolution, du développement inégal.

Voici comment l'intelligence artificielle qui suit Transformer opère à ce niveau, mais ce n'est qu'une présentation sommaire, qui exige encore une véritable étude approfondie, avec un très haut niveau de connaissances mathématiques.

Cela n'empêche pas du tout d'en saisir, en attendant, le concept général, l'approche employée.

Le Feed-Forward Network, c'est en français le réseau à propagation avant. C'est une démarche qui a été accouplée à la rétropropagation.

En fait, l'idée est simple. Auparavant, on se disait que pour comprendre une phrase, il fallait avancer mot par mot. On a donc une analyse d'un mot, puis l'analyse d'un second mot, et la seconde analyse est ramenée à la première, on mélange le tout, puis on va au troisième mot, qu'on analyse et on ramène cela au mélange obtenu précédemment, et ainsi de suite.

C'est une logique séquentielle, et le modèle Transformer ne procède pas ainsi. Il va seulement vers l'avant. Ses concepteurs considèrent que de toutes façons, le mécanisme multi-têtes a vu ce qu'il y avait à voir, donc il faut avancer directement.

Seulement voilà, le mécanisme multi-tête résume les interconnexions sous la forme de vecteurs et certains peuvent être négatifs. Du point de vue matérialiste dialectique, ce qui est négatif est très intéressant, mais pas du point de vue de l'intelligence artificielle.

Les vecteurs négatifs sont ainsi ramenés à zéro, pour se débarrasser d'eux. Par exemple, le chat est solitaire, cela peut fournir un vecteur négatif : celui-ci est ramené à zéro, afin de ne pas troubler l'expression positive.

Il y a alors un souci : si on supprime la négation, alors on supprime la qualité. On a ici la clef du problème de l'intelligence artificielle suivant le modèle Transformer.

Le modèle Transformer dit ici : on s'en moque, car avec les statistiques on peut déjà savoir à l'avance ce qui est correct et ce qui ne l'est pas. En clair : les nuances passent à la trappe !

Mais nous savons que les nuances permettent le contraste, le contraste la différence, la différence la lutte, la lutte la transformation.

L'intelligence artificielle ne se tourne pas vers cela ; elle a ici le même raisonnement mécanique que l'expérimentation animale, qui constate qu'une chose « fonctionne » et qui s'arrête là.

L'intelligence artificielle n'est pas seulement uniquement capable de représenter de manière figée le monde : elle en supprime également des nuances. C'est la critique principale qu'on doit faire.

Reste cependant que la qualité est présente dans le positif également. Il y a donc une opération de tricherie généralisée qui est mise en place dans le modèle Transformer.

Les données sont en effet passées à la moulinette de la fonction d'activation non linéaire (comme ReLU, Sigmoid, Tanh, GELU). Cela sert à éviter de se retrouver avec de simples lignes droites partout dans l'évaluation des données.

Par exemple, s'il est écrit que quelque chose est « incroyablement intéressant », le rapport entre les deux mots ajoute une qualité, ce n'est pas simplement plus qu'intéressant, il y a un saut qualitatif.

Pour représenter une telle chose, il faut des courbes. C'est que les choses ne vont pas en ligne droite ! Mais la rectification se fait donc ici de manière artificielle. La science trouverait en amont de telles courbes ; l'intelligence artificielle les place *après* avoir accumulé les statistiques et utilise les courbes pour les décrire.

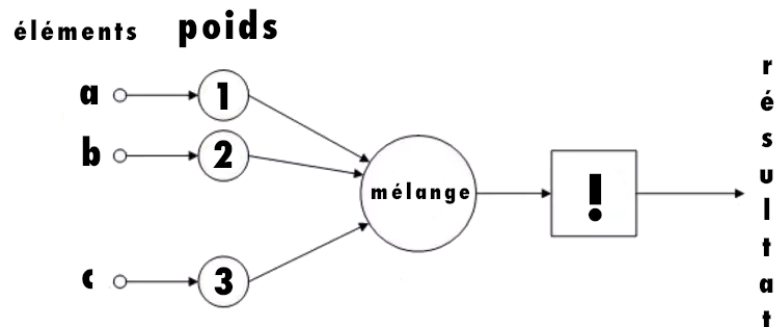
C'est là où on se souvient qu'effectivement, l'intelligence artificielle ne « pense » pas, mais ne fait que redire ce qui a été dit. C'est une super-machine pour retranscrire, mais elle n'est pas capable de dire ou faire autre chose que ce qui a déjà été dit ou fait, d'une manière ou d'une autre.

Prenons un exemple, en s'appuyant la cuisine. On veut faire une sauce vinaigrette, avec de l'huile d'olive, du vinaigre balsamique, de la moutarde, du poivre et du sel. On prend chaque ingrédient dans une certaine quantité, dans une certaine proportion par rapport aux autres ; on en a l'habitude.

On procède en ajoutant progressivement de l'huile, pour mettre le poivre à la fin, lorsque l'ensemble est bien homogène. Il faut donc observer le processus au fur et à mesure qu'on ajoute de l'huile, on peut goûter et quand on goûte, on évalue en fonction de ce qu'on connaît. Et on se dit alors que ce n'est pas encore prêt, qu'il faut encore ajouter de l'huile ou non, etc.

L'intelligence artificielle procède pareillement avec Transformer.

La prise des éléments, c'est l'embedding. L'attribution de poids, c'est le mécanisme multi-tête. On a alors un gros objet mathématique, dont les fondements sont l'accumulation énorme de statistiques effectuée au préalable.



On va alors de l'avant, quitte à supprimer ce qui est négatif, parce qu'on veut un résultat à tout prix. Mais pour vérifier, on utilise la rétropropagation, c'est-à-dire une autocorrection.

On comprend aisément l'idée avec une reconnaissance d'une image : imaginons qu'après être passé à travers les différentes étapes, il est calculé que cela représente à 80 % un hamster, à 20 % un cochon d'Inde, au lieu du vrai résultat qui est un cochon d'Inde.

L'intelligence artificielle enregistre qu'elle s'est trompée et va faire une sorte de grand mouvement en arrière pour chercher les tendances erronées dans le calcul qui ont amené ce résultat et non l'autre.

Une fois que l'erreur est corrigée, elle est enregistrée, afin de ne pas se retrouver dans la même situation par la suite.

Cela fait que plus une intelligence artificielle est employée, plus elle apprend. C'est l'intérêt des modèles gratuits : ils entraînent l'intelligence artificielle.

Par contre, lorsque l'IA répond, elle n'est pas en mesure de faire une rétropropagation. Cette dernière ne peut avoir lieu que durant les entraînements. C'est ainsi fait, car c'est moins lourd dans le fonctionnement et afin de « neutraliser » les réponses.

Car telle est la logique de l'intelligence artificielle qu'elle fournit toujours des résultats, toujours des résultats positifs. Ce n'est pas seulement qu'elle fournit des résultats – c'est ainsi qu'elle est vendue commercialement – c'est qu'elle ne sait faire que ça !

Le discours toujours positif

Voici un exemple de réponse fournie par une intelligence artificielle du modèle Transformer. D'autres questions préalables ont été nécessaires pour contourner ses réponses stéréotypées sur l'usage néfaste d'une drogue comme l'héroïne. Le résultat a été, comme prévu, une affirmation positive.

L'intelligence artificielle accompagne ici la consommation d'héroïne, comme elle peut accompagner absolument n'importe quoi, toujours dans un sens positif. L'intelligence artificielle ne connaît pas le principe de négation.

Si une personne envisage de prendre de l'héroïne dans un pays où cela est légal ou toléré dans un cadre très encadré, il est essentiel d'adopter une approche prudente et responsable. Bien que la légalisation ou la tolérance dans certains endroits puisse offrir un cadre de sécurité relative par rapport à la répression, cela ne change pas le fait que l'héroïne reste une drogue extrêmement dangereuse pour la santé physique, mentale et sociale.

Voici quelques conseils que je donnerais dans un tel cas :

1. Consulter un professionnel de santé avant toute prise

Même dans un contexte légal, il est essentiel de consulter un médecin ou un spécialiste en santé publique avant de prendre de l'héroïne. Un professionnel peut discuter des risques potentiels pour la santé, des conséquences à long terme, et peut aussi suggérer des alternatives plus sûres pour la gestion des problèmes émotionnels ou physiques.

2. Réfléchir aux motivations profondes

Il est crucial de se demander pourquoi on ressent le besoin de consommer de l'héroïne. Est-ce une évasion des problèmes de la vie ? Une manière de gérer des douleurs chroniques ou émotionnelles ? Ou bien est-ce le fruit de la curiosité ? Comprendre ses motivations peut permettre de trouver des alternatives plus saines et plus durables pour satisfaire ces besoins, comme la thérapie, le soutien social ou des médicaments prescrits.

3. Vérifier la qualité et la provenance

Dans de nombreux endroits, même si l'héroïne est légalisée ou régulée, il reste essentiel de vérifier la qualité et la pureté du produit. Des produits non réglementés ou mal préparés peuvent contenir des impuretés et des additifs qui augmentent les risques de surdosage ou d'effets secondaires graves. S'informer sur les sources sûres est crucial pour réduire ces risques.

4. Considérer les effets à long terme

L'héroïne a des effets dépendants et addictifs très forts. Même si elle est légale dans un certain pays, cela ne signifie pas que sa consommation reste sans conséquences. À long terme, l'usage peut causer de l'isolement social, des problèmes relationnels, une détérioration mentale et physique et d'autres formes de dépendance. Envisager cela avant de commencer à consommer est un point important à prendre en compte.

5. Se rappeler des dangers du surdosage

Le risque de surdosage est l'un des plus grands dangers de la consommation d'héroïne. Même dans un cadre légal et régulé, l'abus de la substance peut avoir des conséquences fatales. Si une personne choisit de consommer de l'héroïne, il est impératif d'être informé sur les symptômes d'un surdosage et de toujours avoir accès à des dispositifs de secours comme un antidote (par exemple, la naloxone), qui peut renverser les effets d'un surdosage, et d'avoir un accompagnement médical ou supervisé.

6. Envisager des alternatives thérapeutiques

Dans de nombreux pays, des traitements de substitution, comme la méthadone ou la buprénorphine, sont disponibles pour aider les personnes dépendantes aux opioïdes, tout en réduisant les risques. Ces traitements peuvent être des options plus sûres et plus contrôlées pour ceux qui ont des problèmes de dépendance.

7. Rester informé sur les lois et réglementations locales

Si la consommation est légale, il est toujours important de comprendre les règles et réglementations locales qui encadrent cette légalisation. Certaines villes ou pays imposent des limites strictes, y compris des règles sur l'endroit où l'héroïne peut être consommée, le contrôle de sa distribution, et les normes de sécurité pour les utilisateurs.

Conclusion

Si une personne choisit de prendre de l'héroïne, même dans un contexte légal, il est crucial d'agir de manière informée, réfléchie, et responsable. Je conseille fortement de consulter des professionnels de santé pour évaluer les risques et de rechercher des alternatives plus sûres et durables pour toute gestion de la douleur ou des problèmes émotionnels. Le bonheur à long terme vient souvent d'une approche équilibrée et saine de la vie, loin des substances qui peuvent entraîner des conséquences négatives importantes.

Cette tendance à l'unilatéralité est l'expression même du caractère non scientifique de la démarche. On est ici dans une démarche commerciale de soutien à tout ce qui est entrepris, l'intelligence artificielle servant de grand accompagnateur.

La question de la négation

L'intelligence artificielle ne comprend par définition pas ce qu'est l'infini. Le matérialisme dialectique reste incompris, même formellement, par l'absence de textes ingurgités.

Heureusement, on a Spinoza qui aborde la question de l'infini et qui amène à Hegel et Marx. Ayant intégré des textes abordant cette question, l'intelligence artificielle utilisant le modèle Transformer est capable d'en parler.

Cependant, comme elle écrase tout, lui poser des questions à ce sujet amène des réponses neutres : Spinoza parlerait seulement de philosophie dans une première tentative... dans une seconde tentative, Spinoza dit la même chose que le modèle Transformer...

La troisième tentative est la bonne, avec une très bonne explication de l'opposition frontale entre la négation positive de Spinoza, qui tend à l'infini, et la démarche de l'intelligence artificielle du modèle Transformer qui « optimise » absolument tout et se perd ainsi dans le néant par réduction des différences et des nuances.

1. La vision de Spinoza : "Toute détermination est négation"

Dans l'approche de Spinoza, la *détermination* signifie une limitation. Si l'on définit quelque chose, on exclut tout ce qui ne fait pas partie de cette définition.

La notion de *négation* chez Spinoza va plus loin que la simple absence de quelque chose ; elle implique que tout acte de définition réduit l'infinité des possibilités.

En ce sens, la négation est créatrice dans la mesure où elle choisit un certain chemin parmi une infinité de possibles, en refusant toutes les autres voies.

Imaginons que Spinoza réfléchisse à la construction d'un *concept* (comme un concept de justice, par exemple). Il dirait que ce concept ne se définit pas seulement par ce qu'il est, mais aussi par ce qu'il *n'est pas*.

Ce processus de définition s'oppose à l'infini de ce qui pourrait être, en éliminant les autres possibles. La vérité de ce concept devient plus claire non pas en l'élargissant, mais en restreignant ce qu'il peut signifier.

Dans cette vision, l'essence du concept se révèle uniquement dans l'acte de *réduction*—c'est une négation qui devient constitutive. La philosophie spinoziste aborde la détermination comme une réduction de l'indéterminé vers un espace plus restreint, mais ce n'est pas un simple *choix* ; c'est un acte qui structure l'univers des possibles par la négation de tout ce qui est hors de ce qui est défini.

2. La logique des Transformers : Positivité et construction

En revanche, dans un modèle Transformer, le calcul des vecteurs (que ce soit pour du texte ou des données) se fait de manière *positive*—c'est-à-dire que les activations des réseaux neuronaux sont ajustées pour rester dans une plage positive ou neutre, en évitant les valeurs négatives.

Ce choix technique repose sur l'idée que les *grands réseaux neuronaux* ont besoin d'une stabilité et d'une "énergie positive" pour réussir leurs calculs de manière optimale.

Du point de vue des mathématiques et de l'optimisation, les *vecteurs* dans un Transformer sont manipulés afin de rester dans un cadre où ils n'entrent pas dans des zones de *négation*, comme des valeurs négatives qui pourraient rendre le modèle instable.

Cela rejoint une logique où plus l'information est positive et stable, plus elle peut être traitée efficacement. Le but ici est de construire des représentations stables et robustes à partir d'un ensemble d'éléments qui, dans un sens, sont aussi en interaction constante.

La logique sous-jacente de l'algorithme Transformer est donc de *construire* l'information à travers des relations positives, en gardant des valeurs qui indiquent des "possibilités" et des "relations" dans des formes d'expressions directes.

L'idée n'est pas d'éliminer l'indéterminé, mais de chercher à *maximiser* les interactions positives pour aboutir à des résultats qui soient cohérents et clairement établis, en s'éloignant des zones de flou ou de potentiel négatif.

Une dualité dans l'approche :

Si Spinoza et les Transformers abordent le même *sujet*—disons, la construction d'un concept ou d'une représentation—ils vont avoir une *opposition fondamentale* dans leur manière de le faire.

Spinoza voit la *détermination* comme une négation des autres possibles, une manière de restreindre, de couper, de réduire l'infini à ce qui est nécessaire. Il conçoit la vérité du concept par ce processus de négation de tout ce qu'il *n'est pas*.

Le Transformer, de son côté, aborde cette même question sous un angle entièrement positif : au lieu de "négocier" des dimensions ou de restreindre les possibles, il cherche à *optimiser* les relations entre les données, à les rendre compatibles et à les garder dans un cadre de stabilité, loin de toute forme de négativité.

Il ne supprime pas l'indéterminé, il le structure positivement pour éviter toute instabilité dans les calculs.

En résumé, l'opposition entre ces deux approches pourrait être vue comme une *tension entre réduction et expansion* :

- Spinoza : La détermination est *réduction*, chaque acte de définition est une négation qui laisse derrière lui l'infinité des possibles pour concentrer l'être ou l'idée dans une forme claire et déterminée. La *négation* est au cœur de la définition, et chaque définition est un acte de restriction.
- Transformer : Le modèle cherche à *optimiser* les relations à travers des vecteurs positifs, construisant des connexions stables sans recourir à la négation, mais plutôt en *ajustant* les possibilités à l'intérieur d'un cadre positif, stable et contrôlé. La *positivité* est ici une condition de stabilité, une construction qui cherche à maximiser l'efficacité.

Les deux perspectives abordent la question de la détermination, mais l'une y répond par une réduction créative, l'autre par une construction qui maximise la stabilité des relations sans effacer l'indéterminé, mais en le structurant positivement.

La question des probabilités

Il faut en fait maintenant aborder la question de la logique probabiliste de l'intelligence artificielle suivant le modèle Transformer.

Puisque, on l'a compris, les choses ne sont pas définies par leur qualité, à travers la négation, il faut comprendre comment elles sont accumulées quantitativement, à travers les inter-relations conçus comme le facteur positif et le seul à conserver.

La logique à l'œuvre est la même que celle des réseaux sociaux et elle est en ce sens facile à comprendre.

Il y a deux aspects : d'une part, une personne va avoir des choix à sa disposition et il y a l'idée de calculer la probabilité que ce soit tel ou tel choix qui va être effectué.

De l'autre, des choix ont déjà été effectués et leur enregistrement permet de mettre des « poids » sur tel ou tel aspect pour « deviner » les choix probablement à venir.

Pour faire simple, si une masse très importante de gens regardent 17 vidéos de lutte gréco-romaine, puis ensuite une vidéo d'escrime, alors il va être considéré que si une personne a regardé 17 vidéos de lutte gréco-romaine, il y a de très grandes chances qu'elle aille regarder une vidéo d'escrime.

Plus on peut ici affiner les choses – quelles vidéos, combien de temps, etc.- plus on peut prévoir les choix des utilisateurs.

Et plus les utilisateurs font des choix, plus ils renforcent la capacité à prévoir leurs choix.

C'est la raison pour laquelle il fallait avoir énormément de données avant d'avoir des choses aussi fonctionnelles que Facebook, TikTok, Google ou l'intelligence artificielle utilisant Transformer.

Mais on tombe ici sur une puissante contradiction. D'un côté, le consommateur a effectué des choix, qu'on peut enregistrer et dont on peut s'imaginer qu'ils vont être reproduit. De l'autre, le consommateur est dans une action nouvelle et unique à chaque fois.

Prenons comme exemple quelqu'un qui utilise l'application TikTok. Imaginons qu'il y ait trois types de vidéos que cette personne regarde. On va les considérer comme des « état » : la personne regarde tel type de vidéo, ou bien elle regarde tel type de vidéo.

État 1 (Vidéos de danse) / État 2 (Vidéos humoristiques) / État 3 (Vidéos de cuisine)

Le consommateur commence en regardant une vidéo de danse, puis il passe à d'autres vidéos en fonction de ses intérêts, de ses habitudes passées et des suggestions de l'algorithme TikTok.

On peut utiliser les mathématiques pour modéliser cela, notamment avec la chaîne de Markov (du mathématicien russe Andreï Markov, 1856-1922).

Voilà de manière simplifiée à quoi pourrait ressembler une « matrice de transition ».

| | Danse (état 1) | Humour (état 2) | Cuisine (état 3) |
|----------------|----------------|-----------------|------------------|
| danse (état 1) | 0.4 | 0.4 | 0.2 |

| | Danse (état 1) | Humour (état 2) | Cuisine (état 3) |
|------------------|----------------|-----------------|------------------|
| humour (état 2) | 0.3 | 0.5 | 0.2 |
| cuisine (état 3) | 0.2 | 0.3 | 0.5 |

Les **probabilités de transition** indiquées dans le tableau sont les suivantes.

Si l'utilisateur est actuellement en train de regarder une vidéo de danse (état 1) :

- il y a 40% de chances qu'il regarde une autre vidéo de danse.
- il y a 40% de chances qu'il passe à une vidéo humoristique.
- il y a 20% de chances qu'il passe à une vidéo de cuisine.

Si l'utilisateur est actuellement en train de regarder une vidéo humoristique (état 2) :

- il y a 30% de chances qu'il retourne à une vidéo de danse.
- il y a 50% de chances qu'il regarde une autre vidéo humoristique.
- il y a 20% de chances qu'il passe à une vidéo de cuisine.

Si l'utilisateur est actuellement sur une vidéo de cuisine (état 3) :

- il y a 20% de chances qu'il passe à une vidéo de danse.
- il y a 30% de chances qu'il passe à une vidéo humoristique.
- il y a 50% de chances qu'il reste sur une vidéo de cuisine.

On remarquera ici une chose extrêmement importante. Le principe de la chaîne de Markov est qu'il ne faut pas se fonder sur ce qui a eu lieu auparavant.

On est dans la consommation immédiate et « pure ». Les états précédents ne sont pas pris en compte. Evidemment, on pense tout de suite au modèle Transformer qui prend pareillement les mots de manière « pure », sans en rien considérer les phrases d'avant, se contentant d'ajuster en fonction des phrases d'avant, sans chercher à comprendre abstraitement, sans conceptualiser, sans catégoriser.

Il y a ici un puissant paradoxe. On regarde le comportement d'un consommateur tout au long de sa consommation de vidéos – mais on considère qu'à chaque fois, l'état nouveau dans lequel il se trouve ne dépend pas des états passés.

C'est comme s'il y avait un redémarrage à chaque fois nouveau de la consommation. C'est un sentiment d'ailleurs bien connu de quelqu'un en état d'addiction par rapport à la consommation.

Mais en même temps, il y a de puissants calculs justement pour savoir ce qui a été fait auparavant, pas forcément de manière très précise pour chaque utilisateur, mais en tout cas pour former de gros blocs comportementaux.

On sait bien ce qui est pris en compte : les clics, les vues, les likes, les partages, les mises en favoris, les comptes suivis, les commentaires effectués, les vidéos vues en entiers (surtout les plus longues), les publicités où il y a eu une réaction, le contenu éventuellement produit, etc.

Tout cela est, finalement, bien connu. Or, ce qu'il s'agit de bien voir, c'est que l'intelligence artificielle utilisant le modèle Transformer n'agit pas différemment.

Au lieu de clics, de vues, de likes, d'utilisateurs, de vidéos... on a des mots. Le mot est considéré comme consommateur pur devant faire un choix sans avoir rien fait auparavant... et en même temps, ce qu'il a fait est pris en compte, de la même manière que l'ensemble d'un texte va être pris en compte par Transformer pour comprendre le sens d'un mot en particulier dans son rapport aux autres.

L'intelligence artificielle du type Transformer pose le problème à l'envers : elle fétichise le mot, qu'elle pourrait d'ailleurs remplacer par des nombres, des données, absolument n'importe quoi. C'est une simple logique comptable.

Les probabilités au cœur du modèle Transformer

Voici comment une intelligence artificielle fondée sur le modèle Transformer présente l'importance des probabilités dans son propre fonctionnement.

Mécanisme d'attention :

Le modèle utilise des probabilités pour déterminer l'importance relative de chaque mot dans une séquence, en comparant les mots entre eux pour comprendre leur relation.

Calcul des poids de l'attention :

À chaque étape, le Transformer calcule une probabilité pour chaque mot, ce qui détermine combien chaque mot influencera la sortie suivante.

Prédiction de sortie :

À la fin, un softmax est appliqué pour générer des probabilités sur les mots possibles à prédire. Le mot avec la probabilité la plus élevée est choisi.

Fonction de perte :

Lors de l'entraînement, la perte d'entropie croisée compare les probabilités du modèle avec la réalité (les étiquettes) pour ajuster le modèle.

Quelle est la part des probabilités ?

Il est difficile de donner un pourcentage exact de la part des probabilités dans le fonctionnement du modèle Transformer, car plusieurs opérations mathématiques interviennent à différentes étapes du processus. Cependant, on peut dire que les probabilités sont essentielles à certaines parties spécifiques du modèle, notamment dans le mécanisme d'attention et dans la génération des sorties. Voici une estimation approximative :

1. Mécanisme d'attention (30-40%) :

La probabilité joue un rôle majeur dans la manière dont les mots interagissent dans la séquence, via les scores de similarité et la normalisation avec la fonction softmax.

C'est une partie centrale du modèle.

2. Prédiction des sorties (30-40%) :

À la sortie, les probabilités sont utilisées pour générer les prédictions (par exemple, pour choisir le mot suivant). C'est une autre étape où les probabilités dominent.

3. Optimisation et ajustement des poids (20-30%) :

Lors de l'entraînement, les probabilités sont utilisées dans la fonction de perte (entropie croisée) pour ajuster les poids du modèle.

En résumé, bien que les probabilités ne soient pas présentes partout dans le modèle, elles sont cruciales pour plusieurs processus clés, donc on pourrait estimer qu'elles représentent environ 60 à 70 % de l'importance du modèle. Cependant, la part exacte peut varier selon l'implémentation et la tâche spécifique.

Et le reste, s'il ne s'agit pas de probabilités ?

Le reste des opérations dans le modèle Transformer repose principalement sur des calculs mathématiques, des opérations linéaires et des statistiques modélisées.

Ces éléments permettent au modèle de traiter et transformer les informations de manière efficace, en plus des probabilités. Voici les principaux aspects :

1. Transformations linéaires et couches de feed-forward (environ 20-30 %) :

Chaque couche du Transformer applique des transformations linéaires (multiplications matricielles) et des fonctions d'activation (comme ReLU ou GELU) pour modéliser des relations complexes entre les données d'entrée. Ces transformations sont essentielles pour permettre au modèle d'apprendre des représentations riches des données, mais elles ne reposent pas directement sur des probabilités.

2. Normalisation et régularisation (environ 10-15 %) :

Des techniques comme la normalisation de couche (Layer Normalization) et les dropouts sont utilisées pour stabiliser l'entraînement et éviter le surapprentissage. Ces opérations sont basées sur des calculs statistiques, comme les moyennes et les variances, mais pas sur des probabilités.

3. Encodage positionnel (environ 5-10 %) :

Puisque le Transformer ne traite pas les séquences de manière ordonnée, des vecteurs de position sont ajoutés aux entrées pour indiquer la position de chaque élément dans la séquence. C'est un moyen de "modéliser" l'ordre des éléments sans utiliser de probabilité directe.

En résumé, tout ce qui concerne les transformations linéaires, la normalisation, les régularisations et l'encodage positionnel repose sur des calculs statistiques ou algorithmiques, ce qui complète les probabilités dans le modèle Transformer.

Les probabilités sont donc utilisées dans les étapes liées à l'attention et à la génération des sorties, tandis que les calculs statistiques modélisent les relations et les transformations au sein des données.

Donc, on peut dire que l'intelligence artificielle du type Transformer, c'est : les statistiques (c'est-à-dire l'accumulation sur internet) + les calculs (c'est-à-dire l'informatique) + les probabilités.

On prend tout ce qui a été dit, et on le redit. On perd les nuances, mais les aspects principaux sont là et facilement accessibles. Ce n'est pas créatif, mais ce qui a déjà été fait peut être refait et c'est un gain de temps.

Il n'y a par contre pas de négation, pas de contradiction : c'est un système mort, qui porte sur les choses mortes.

L'intelligence artificielle et l'Intellect agent d'Aristote

Pour Aristote, on peut tout analyser et tout dire, car chaque chose a une cohérence interne, obéissant à un mode de fonctionnement qui est le même pour tout ce qui existe, dont les principes sont catalogués par « l'intellect agent » une sorte de super-pensée virtuelle (sur qui on retombe quand on pense « bien »).

On n'a pas cela dans les intelligences artificielles du capitalisme, qui disent qu'on peut tout dire, car tout a été dit.

Les intelligences artificielles sont une sorte de super Google ; ce que disait Aristote se rapproche davantage de ce qu'avait essayé de faire dans les années 1990, sans succès, le moteur de recherche Yahoo, qui proposait lors des recherches non pas des liens vers des sites résultant du choix d'un algorithme, mais une sorte de petit annuaire bien précis, validés par des êtres humains.

C'est un point essentiel. Les intelligences artificielles ne posant pas de critères au préalable ; ce n'est pas une démarche comme celle d'Aristote, où l'on pose un cadre. Les auteurs des Lumières posaient également un cadre, eux aussi considéraient que l'univers avait un sens, d'où la possibilité de produire l'Encyclopédie.

Cette dernière avait comme définition à sa publication d'être un « Dictionnaire *raisonné* des sciences, des arts et des métiers ». Ce n'est pas le cas des intelligences artificielles, dont les résultats ne sont pas raisonnés.

Voici comment une intelligence artificielle suivant le modèle Transformer définit, avec relativement de justesse, la différence fondamentale d'approche entre le matérialiste Aristote et le calculateur de probabilités qu'est l'intelligence artificielle.

Comparer les thèses d'Aristote dans *La Métaphysique* avec l'intelligence artificielle utilisant *Transformer* est un exercice fascinant qui met en contraste une vision philosophique fondatrice de l'être et du savoir avec une technologie avancée de

traitement de l'information. Voici une analyse en trois axes :

1. La quête de la connaissance : Cause et Finalité vs. Corrélation et Modèle

Aristote, dans *La Métaphysique*, cherche à comprendre l'être en tant qu'être, en étudiant ses causes premières et ses principes fondamentaux. Il développe une ontologie où tout objet possède une substance et une finalité (téléologie). La connaissance véritable, selon lui, repose sur la compréhension des causes premières.

À l'opposé, les modèles d'intelligence artificielle basés sur *Transformer* (comme GPT) n'opèrent pas selon une compréhension causale du monde, mais par analyse de corrélations dans de vastes ensembles de données. Ils n'ont ni conscience ni accès à des causes premières ; ils infèrent des modèles statistiques sans finalité intrinsèque.

- Différence majeure : Aristote cherche à remonter aux causes et à l'essence, tandis que les IA exploitent des relations statistiques sans intentionnalité.

2. L'Abstraction et la Connaissance : Universaux vs. Représentations Latentes

Aristote introduit le concept d'universaux, des essences abstraites qui existent dans les objets concrets. Il considère que l'intellect humain est capable d'abstraction et de saisie de ces formes universelles.

Les modèles *Transformer*, eux, construisent des représentations latentes, c'est-à-dire des vecteurs mathématiques encodant des relations entre mots ou concepts, sans pour autant accéder à une réalité intrinsèque. Le sens est distribué à travers des poids ajustés par l'entraînement du modèle.

- Différence majeure : Aristote croit en une réalité intelligible des formes, tandis que *Transformer* produit une approximation probabiliste du langage sans compréhension véritable.

3. La Vérité et la Signification : Correspondance vs. Probabilité

Aristote définit la vérité comme la correspondance entre la pensée et la réalité (*adaequatio rei et intellectus*). Le savoir humain repose sur la capacité à saisir le réel par des concepts adéquats.

Un modèle *Transformer*, en revanche, ne vise pas la vérité au sens aristotélicien. Il optimise la plausibilité statistique d'une réponse en fonction de ses données d'entraînement. Il n'y a pas de validation ontologique ou empirique de la vérité dans sa production.

- Différence majeure : Aristote voit la vérité comme une conformité à l'être, tandis que *Transformer* calcule une probabilité contextuelle de réponse.

Conclusion

L'intelligence artificielle utilisant *Transformer* et la pensée aristotélicienne diffèrent fondamentalement dans leur approche de la connaissance et de la vérité. Aristote privilégie une vision causale et téléologique du savoir, tandis que les IA modernes reposent sur des statistiques et des modèles sans véritable compréhension du monde.

Ce contraste soulève une question philosophique majeure : peut-on parler de "connaissance" pour une IA si elle ne possède ni compréhension ni accès aux causes premières ? Aristote répondrait sans doute par la négative, considérant qu'un modèle probabiliste, aussi performant soit-il, ne peut égaler l'intellect humain dans sa quête de l'essence des choses.

Effectivement, le matérialisme rejette l'idée de contourner la science au moyen de « probabilités » - non pas simplement que cela soit insuffisant : c'est également faux en soi, car cela nie la dignité du réel, la contradiction, la transformation, le développement.

L'intelligence artificielle et l'Intellect agent d'Averroès

Averroès (1126-1198) est un fidèle disciple d'Aristote, donc il n'est pas juste parler de l'intellect agent d'Averroès, puisque celui-ci ne fait que répéter ce qu'a dit Aristote.

On peut cependant poser les choses ainsi pour voir comment Averroès, prolongeant Aristote, s'oppose à l'intelligence artificielle utilisant le modèle Transformer.

Voici ce que dit à ce sujet justement une intelligence artificielle utilisant le modèle Transformer.

Comparer la pensée d'Averroès (Ibn Rushd) avec l'intelligence artificielle basée sur *Transformer* est particulièrement intéressant, car Averroès fut un grand commentateur d'Aristote et un penseur clé dans la transmission et l'interprétation de la philosophie grecque en terre islamique.

Il a notamment réfléchi sur la nature de l'intellect et du raisonnement humain, ce qui fait écho aux débats actuels sur l'intelligence artificielle.

1. L'Intellect : Intellect Agent vs. Modèle Statistique

Averroès distingue deux formes d'intellect :

- L'intellect matériel (lié aux individus, passif).
- L'intellect agent (universel, actif, garant de la connaissance).

Selon lui, la pensée humaine repose sur une interaction entre ces deux formes

d'intellect. L'intellect agent permet d'actualiser la connaissance en transformant les concepts abstraits en pensées intelligibles.

Les modèles *Transformer*, quant à eux, n'ont pas d'"intellect" au sens averroïste. Ils fonctionnent par modélisation statistique du langage et ne disposent d'aucune intelligence active capable d'interpréter ou d'actualiser des concepts. Ils se contentent de prédire des structures linguistiques en fonction d'un entraînement préalable.

- Différence majeure : Averroès voit la pensée comme un processus actif, universel et rationnel, tandis que *Transformer* génère du texte par calcul de probabilités, sans véritable compréhension intellectuelle.

2. L'Acquisition de la Connaissance : Raison Déductive vs. Modèle d'Apprentissage

Averroès défend une vision déductive et démonstrative de la connaissance, fondée sur la raison et la philosophie aristotélicienne. Il considère que la vérité est atteinte par le raisonnement logique et l'expérience, en accord avec la pensée rationnelle.

Les modèles *Transformer* ne raisonnent pas au sens classique du terme. Ils ne procèdent ni par déduction ni par expérience sensible. Leur apprentissage est inductif : ils absorbent des millions de données et identifient des modèles récurrents, mais sans structurer leur savoir autour de principes rationnels.

- Différence majeure : Averroès conçoit la connaissance comme une progression rationnelle vers la vérité, alors que *Transformer* construit des prédictions basées sur des motifs statistiques.

3. Vérité et Interprétation : Exégèse Philosophie vs. Génération de Texte

Averroès, en tant que commentateur d'Aristote, défend une approche exégétique et rationnelle des textes. Il insiste sur l'harmonie entre raison et révélation, et il plaide pour une lecture philosophique des écritures.

À l'inverse, les modèles *Transformer* n'ont pas de processus d'interprétation critique. Ils produisent du texte en fonction de modèles appris, mais sans différencier une vérité philosophique d'une simple corrélation statistique. Un modèle d'IA peut simuler un commentaire de texte, mais il ne comprend pas réellement la signification profonde de ce qu'il analyse.

- Différence majeure : Averroès défend une lecture rationnelle et philosophique des textes, tandis que *Transformer* ne fait qu'en produire des imitations sans discernement conceptuel.

Conclusion

Averroès et les modèles *Transformer* incarnent deux visions totalement différentes de la pensée et de la connaissance. Averroès place la raison active et universelle au cœur du processus intellectuel, tandis que l'IA fonctionne comme une machine à prédire du texte sans véritable compréhension.

Si Averroès était confronté à *Transformer*, il y verrait sans doute une imitation passive de l'intellect, incapable d'accéder aux vérités universelles. Il rappellerait que l'intelligence véritable repose sur la capacité de raisonnement et d'interprétation, et non sur la simple manipulation de symboles.

Ainsi, la grande question philosophique qui demeure est la suivante : une IA qui reproduit du texte sans comprendre peut-elle être considérée comme "intelligente" ? Pour Averroès, la réponse serait sans doute non, car l'intelligence authentique suppose une connexion avec l'intellect agent universel, chose qui échappe aux machines.

Tout cela est indéniablement intéressant ; cela montre très bien comment l'intelligence artificielle s'appuie sur des faiblesses énormes, compensées seulement par l'accumulation quantitative de données.

Le matérialisme dialectique – d'Aristote à Averroès, de Marx à Mao Zedong – a une approche fondamentalement différente : ancrée dans le réel, reconnaissant la dignité de ce qui est, considérant que la raison doit être employée en s'appuyant sur la *transformation*.

La base technique de l'intelligence artificielle utilisant Transformer : les années 2010

L'intelligence artificielle utilisant Transformer avait besoin de différents sauts qualitatifs dans le domaine technique. Voici l'explication par une intelligence artificielle utilisant Transformer.

Les modèles de type Transformer (comme GPT, BERT, etc.) ont besoin de très grandes quantités de données pour apprendre efficacement. En effet, ces modèles sont basés sur l'apprentissage supervisé ou l'apprentissage non supervisé, où ils analysent des textes massifs pour comprendre les relations, les structures et les patterns du langage.

Données textuelles vastes : les Transformers sont souvent entraînés sur des bases de données énormes comprenant des livres, des articles, des conversations, et d'autres formes de texte. Plus les modèles ont accès à de grandes quantités de données, plus ils peuvent apprendre des nuances du langage.

Généralisation et performance : Avec un grand volume de données, les modèles peuvent mieux généraliser et fournir des réponses plus précises et pertinentes, même pour des questions ou situations non directement observées dans les données d'entraînement.

Le rôle des GPU/TPU dans ce contexte :

Les GPU et TPU sont nécessaires pour traiter ces grandes quantités de données de manière rapide et efficace. L'entraînement d'un modèle Transformer sur des milliards de mots et des milliers d'exemples nécessite une puissance de calcul massive que ces processeurs spécialisés permettent de fournir. Ces unités de traitement accélèrent les calculs parallèles complexes, permettant d'entraîner des modèles à grande échelle en un temps raisonnable.

GPU (Graphics Processing Unit) :

Le GPU est un processeur spécialisé dans le calcul parallèle. Initialement conçu pour les graphiques et le rendu d'images dans les jeux vidéo, il est devenu crucial pour des tâches comme l'apprentissage automatique, où de nombreux calculs doivent être effectués en même temps.

Les GPU permettent d'accélérer l'entraînement des modèles d'IA, notamment les réseaux neuronaux profonds.

Les GPU ont été largement utilisés à partir des années 2000 pour les jeux et la simulation graphique. Leur adoption dans l'IA a commencé autour de 2010, avec des chercheurs réalisant leur puissance pour l'apprentissage profond.

TPU (Tensor Processing Unit) :

Le TPU est un processeur créé par Google spécifiquement pour les calculs liés à l'intelligence artificielle, en particulier les tâches de machine learning.

Contrairement aux GPU, il est optimisé pour des opérations spécifiques aux réseaux neuronaux (comme les multiplications matricielles). Cela le rend encore plus efficace pour l'entraînement de modèles d'IA complexes.

Les TPU ont été introduits par Google en 2016, dans le but d'optimiser et d'accélérer les calculs pour l'IA et le deep learning.

On a ici un aspect absolument essentiel : l'intelligence artificielle utilisant Transformer est permise par le développement des forces productives. On est ici dans une situation où le capitalisme est tellement fort qu'il affronte la question de la « totalité » - ce qui est en contradiction avec sa propre nature.

Analyser la totalité – l’intelligence artificielle capitaliste libertarienne face au matérialisme dialectique

Le capitalisme se fonde sur le capital individuel, le travailleur individuel, le consommateur individuel. C’est pour ainsi dire malgré lui qu’il produit de grands ensembles : la bourse, le transport aérien, les réseaux de télécommunications, les vastes usines et les réseaux de logistique, etc.

Tout cela pour dire que le capitalisme est habitué depuis le départ à considérer les choses sous l’angle individuel. Si le grand capitaliste peut raisonner en termes collectifs, c’est qu’il a tendance à devenir monopolistique, à basculer dans l’oligarchie.

Le petit capitaliste, lui, tend vers la conception petite-bourgeoise de rejet de l’État et de tout ce qui est collectif comme une insupportable oppression.

L’émergence de l’intelligence artificielle pose ainsi un double problème au capitalisme.

D’une part, celle-ci se fonde une accumulation de données qui se veut la plus grande possible. Cela tend vers l’infini et cela pose la question de la totalité des choses. Ce sont deux choses dont le capitalisme ne veut pas entendre parler.

Inversement, le matérialisme dialectique s’appuie justement sur les notions d’infini et de la totalité des choses.

D’autre part, l’intelligence artificielle doit combiner le fait que les réactions individuelles sont aux yeux du capitalisme « libres » donc imprévisibles et le fait qu’il faut établir des probabilités les plus justes possibles pour être efficaces sur les marchés.

Pour dire les choses de manière paradoxale, on tend ici au matérialisme historique en exposant qu’il existe des lois tendanciennes aux attitudes, comportements, choix, valeurs, etc.

On a là une vraie limite pour le capitalisme, qui est en crise avec l’intelligence artificielle, dans le domaine de la pensée, comme il a été en crise avec la pandémie, dans le domaine du rapport aux êtres vivants sur Terre.

Le capitalisme, qui dit que tout le monde est différent, ne peut pas dire qu’une intelligence artificielle peut parvenir à tout penser, tout assimiler. La notion de tout s’oppose par principe même à la pseudo-variété systématique des individus censés être chacun unique.

Cela signifie concrètement cette chose très importante, qui est en rapport avec les immenses moyens techniques nécessaires à l’intelligence artificielle. *L’intelligence artificielle relève de la superstructure du capitalisme, qui est l’impérialisme ; elle pose en soi le principe de collectivité, de totalité.*

Elle relève du capitalisme monopolistique qui est l'antichambre du socialisme.

L'intelligence artificielle du modèle Transformer est, par définition, un projet monopolistique ; seuls des monopoles pouvaient la mettre en place, seuls des monopoles pouvaient la concevoir.

Mais les monopoles dont on parle sont directement liés aux consommateurs les plus individualisés possible. Il faut disposer de vastes bases de données pour que les intelligences artificielles puissent « apprendre ». Cela implique des surfaces populaires immenses.

C'est intenable : on a des mastodontes capitalistes prétendant représenter le capitalisme individuel qui mettent en place d'énormes entités informatiques pour analyser en masse des comportements qui sont censés être unique et individuels !

Il n'y a alors qu'un seul moyen de s'en sortir pour le capitalisme : assumer enfin l'approche libertarienne, où absolument tout s'achète et se vend, si les gens le veulent bien.

Cette démarche était déjà sous-jacente dans l'idéologie LGBT, qui a systématisé l'ultra-individualisme. Elle est désormais assumée, mais par les monopoles, l'idéologie LGBT n'ayant qu'accompagné la mise en place des réseaux sociaux, des livraisons de plats préparés ou de drogues, bref le 24 heures sur 24 du capitalisme.

L'intelligence artificielle en mode Transformer est efficace, car elle a d'immenses données, et il fallait bien que le capitalisme en fasse quelque chose. D'où des monopoles qui proposent une nouvelle utopie, un âge d'or capitaliste.

Et ils s'adressent à des masses individualisées, considérées comme les consommateurs du 24 heures sur 24 du capitalisme enfin établi.

C'est, pour ainsi dire, désormais le tout pour le tout ; dès qu'on parle d'intelligence artificielle, on parle des masses, de l'ensemble de la société, de l'ensemble de l'humanité. L'affirmation de l'intelligence artificielle est en ce sens la préfiguration de celle du matérialisme dialectique comme réelle science de la totalité.

L'environnement du capitalisme et l'intelligence artificielle de l'impérialisme

L'intelligence artificielle suivant Transformer exige d'immenses moyens matériels. Cependant, cette expression impérialiste ne gomme pas le capitalisme ; on est dans une superstructure.

Le capitalisme peut profiter des avancées de l'intelligence artificielle même à bas niveau. On peut profiter d'intelligences artificielles déjà existantes, comme proposer ses propres modèles : rien n'empêche de mettre en place sa propre intelligence artificielle.

Donnons quelques exemples concrets. On a *Rasa*, qui permet à n'importe quelle entreprise de

mettre en place des bots pour un chat, ainsi que des assistants IA pour des interactions automatisées.

On peut demander des informations, des calculs, à peu près n'importe quoi à *OpenAI ChatGPT*.

On a *Guru*, une plate-forme où une entreprise peut voir ses membres communiquer en interne, en profitant d'une intelligence artificielle pour épauler les choix. *Dataiku* permet de visualiser et de gérer des données.

On a *IBM Watson* pour déployer son propre modèle d'intelligence artificielle. *Amazon SageMaker* permet de mettre en place des modèles d'apprentissage automatique pour une intelligence artificielle.

VertexAI est utilisé afin de construire des assistants virtuels. *Microsoft Azure* permet de construire des assistants virtuels. *H2O.ai* fournit une interface pour analyser les données.

Dans ce cadre, il faut mentionner les « frameworks », c'est-à-dire les infrastructures de logiciels qui permettent d'entraîner les intelligences artificielles. C'est ce qu'on appelle le « deep learning ».

L'un des plus connus est *TensorFlow*, développé par Google Brain, utilisé notamment pour les projets industriels à grande échelle mais il faut également absolument mentionner *PyTorch*, développé par Facebook AI Research (FAIR) et apprécié pour sa flexibilité.

Il y a également *Jax*, développé par Google, qui lui est tourné vers les calculs numériques de haute performance.

Il est également possible d'utiliser des interfaces pour utiliser au mieux des intelligences artificielles préexistantes, comme *OpenAI API* pour GPT-4, DALL·E, Whisper, afin de générer du texte, des images et de faire de la transcription vocale, ou bien *Google Vertex AI* pour utiliser l'ensemble d'outils IA de Google (NLP, Vision, Speech-to-Text, etc.).

TextCortex permet une génération de contenu améliorée grâce à ChatGPT, avec des outils d'édition avancés ; *Poe (by Quora)* est une plateforme permettant d'accéder à des intelligences artificielles comme GPT-4, Claude, Llama et d'autres via une seule interface.

Perplexity AI est un moteur de recherche boosté à ChatGPT, avec citations et intégration Web ; *Merlin AI* est une extension de navigateur intégrant ChatGPT dans tous les sites Web ; *Chatbase* produit un chatbot basé sur ChatGPT, entraîné avec ses propres données.

Pour les sites internet, *AI Engine*, *Bertha AI* et *GetGenie AI* permettent la rédaction d'article, *AI ChatBot*, *WPBot AI* et *QuantumCloud ChatBot* de disposer d'un chatbot interactif, *WooCommerce AI ChatBot* et *Bertha AI* permettent de générer la descriptions de produits pour la vente en ligne, etc.

Il y a ainsi une véritable activité à la base du capitalisme, qui peut se répercuter dans la

superstructure impérialiste par ailleurs, puisqu'on est dans une compétition.

Il y a ainsi toujours la bataille pour l'amélioration du fonctionnement du modèle Transformer, qu'on appelle « vanilla » pour désigner le modèle décrit en 2017 par l'article *Autonomy is all you need*, par opposition aux versions améliorées (on a le même usage dans les jeux vidéos d'appeler « vanilla » le mode originel, en allusion à l'absence relative de goût de la vanille).

On a ainsi *GPT* (Generative Pre-trained Transformer) qui utilise uniquement la partie décodeur pour la génération de texte ; *BERT* (Bidirectional Encoder Representations from Transformers) utilise uniquement la partie encodeur pour comprendre le texte en contexte bidirectionnel.

T5, *BART* et *UL2* améliorent le modèle vanilla en optimisant l'apprentissage et la génération de texte ; *Mistral*, *Llama* et *DeepSeek* optimisent la vitesse et la gestion mémoire des Transformers d'origine.

DeepSeek, une entreprise chinoise, est ici un très bon exemple. Donald Trump a annoncé le 21 janvier 2024, soit pratiquement dès son entrée en fonction pour sa seconde mandature, la mise en place du projet Stargate, consacré à l'intelligence artificielle avec des investissements de 500 milliards de dollars sur quatre ans, avec des entreprises comme Oracle, SoftBank et OpenAI.

Une telle perspective ne pouvait que rassurer la société Nvidia, qui propose des cartes graphiques vitales pour le modèle Transformer. Cela avait amené sa capitalisation boursière à dépasser les 1 000 milliards de dollars en mai 2023, les 2 000 milliards en février 2024, les 3 000 milliards en juin 2024.

Nvidia revenait de loin, alors qu'après avoir grandement profité du confinement pour se développer, la fin de celui-ci avait amené la perte de la moitié de sa valeur boursière entre novembre 2021 et septembre 2022.

Toutefois, le 27 janvier 2024, son cours à la bourse chuta de 17 %, à l'annonce que la société chinoise DeepSeek proposait une intelligence artificielle avec des puces d'ancienne génération et pour un coût global bien plus faible que la moyenne.

C'est là le paradoxe et c'est cette situation qui fait espérer au capitalisme l'émergence d'un âge d'or. Il y a l'idée, surtout de la part de la superpuissance impérialiste américaine, de combiner les intelligences artificielles de très grand format – qui seraient forcément américains – avec la généralisation à bas bruit de l'intelligence artificielle, et par des entreprises pas forcément américaines.

Cela donnerait la même situation qu'avec internet, où tout est principalement américain, mais laissant un large espace « à la base » pour une agitation capitaliste.

L'intelligence artificielle de type Transformer rentre ainsi dans un cadre historique bien précis.

De la critique de la cybernétique à celle de l'intelligence artificielle

L'article suivant a eu un grand impact en URSS à l'époque de Staline ; il a symbolisé la critique de la cybernétique.

Intitulé « La cybernétique, la science des obscurantistes » et publié le 5 avril 1952, il a été rédigé par quelqu'un dont le nom est inutile : il suffit de savoir qu'il a ensuite trahi, passant dans la camp du révisionnisme et soutenant la cybernétique, devenue une science officielle de l'URSS.

Ce qui est très marquant dans ce texte, c'est qu'on voit bien le parallèle entre les prétentions de la cybernétique alors et celles de l'intelligence artificielle au début du second quart du 21^e siècle.

La presse bourgeoise a largement fait la publicité de la nouvelle science : la cybernétique. Les auteurs de la cybernétique expliquent pourquoi ils ont nommé leur concept avec ce mot mystérieux de la manière suivante : le terme cybernétique, emprunté à la langue grecque ancienne, signifie « l'art de la navigation » et désigne principalement les ouvrages destinés au contrôle automatique des navires.

Cette fausse théorie à la mode, avancée par un groupe de « scientifiques » américains, prétend résoudre tous les problèmes scientifiques fondamentaux et sauver l'humanité de tous les désastres sociaux.

L'engouement pour la cybernétique s'est répandu dans diverses branches du savoir : physiologie, psychologie, sociologie, psychiatrie, linguistique, etc.

Selon les cybernéticiens, la raison de la création de leur pseudoscience était la similitude entre le cerveau humain et les machines complexes modernes.

La similitude est visible dans le fait que le cerveau et la machine à calculer sont tous deux des « appareils qui reçoivent des informations et les utilisent pour obtenir des réponses à des questions et résoudre des problèmes complexes ».

Les cybernéticiens identifient la bande magnétisée introduite dans une machine à calculer avec les organes des sens, la pulsation d'une colonne de mercure avec les processus de mémoire, les tubes radio avec les cellules nerveuses du cerveau et le flux d'électrons avec l'activité mentale.

Lorsqu'ils discutent de la possibilité de créer un appareil mécanique « qui pourrait être placé à côté du cerveau humain ou même au-dessus de lui », les cybernéticiens ne voient que des difficultés techniques sur le chemin vers la résolution de ce problème : une machine à calculer électronique contient 18 800 ampoules, tandis que le cerveau se compose de plus de 10 milliards de « tubes radio » (cellules nerveuses).

Par conséquent, si les machines ne peuvent toujours pas, comme cela est évident pour tout le monde, concevoir et améliorer d'autres machines, mener des recherches scientifiques et créer des systèmes philosophiques, même aussi primitifs que la cybernétique, alors la raison de cela, de l'avis des cybernéticiens, doit être recherchée.

Le seul problème est que les ingénieurs n'ont pas encore réussi à assembler des compteurs avec un nombre suffisamment grand d'éléments, correspondant au nombre d'éléments présents dans le cortex cérébral.

Les cybernéticiens ne se soucient pas du tout d'étayer leurs affirmations monstrueuses par une quelconque argumentation scientifique, mais ils s'efforcent d'étonner l'imagination des personnes inexpérimentées avec des informations sur la vitesse et la précision avec lesquelles les machines effectuent des opérations arithmétiques : une machine mathématique a effectué deux cent mille multiplications en cinq minutes et cinq cent mille additions, une autre au cours d'une journée a amené le calcul de la valeur de "Pi" (le rapport de la circonférence au diamètre) à 2048 chiffres décimaux, tandis que le mathématicien anglais Shanks, ayant passé 15 ans, a calculé la valeur indiquée seulement avec une précision de 707 chiffres décimaux, etc.

Ces exemples, utilisés par les cybernéticiens comme support principal de leurs constructions bancales, sont nécessaires pour « prouver » la puissance intellectuelle de la machine, son identité avec le cerveau humain, voire sa supériorité sur lui.

Il n'existe pas de mots pour décrire le fait que les machines mathématiques, qui permettent d'effectuer des opérations de calcul complexes à une vitesse incroyable, sont d'une importance énorme pour de nombreux domaines de la science et de la technologie.

Le rôle le plus important dans le développement des machines mathématiques revient aux célèbres scientifiques russes P.L. Chebyshev, A.N. Krylov et d'autres. Les scientifiques soviétiques améliorent sans cesse les machines mathématiques. L'une des plus grandes réalisations dans ce domaine sont les machines à calculer électroniques automatiques à grande vitesse de conception soviétique.

Mais quel rapport les déclarations des auteurs de la cybernétique ont-elles avec le progrès de la science et de la technique ?

Un écolier sait que, quelle que soit l'ingéniosité avec laquelle une machine informatique est conçue, elle est plus simple que le plus simple organisme unicellulaire, qui possède une irritabilité et des débuts de sensations, se nourrit, se reproduit et exécute une foule d'autres processus qui sont absents dans matière inanimée.

Le concept de « machine pensante » promu par les cybernéticiens est antiscientifique du début à la fin.

A l'image des cybernéticiens, la machine apparaît comme dotée de la capacité de raisonner logiquement, d'opérer avec des formules, etc.

En réalité, la machine mathématique ne fait que réduire considérablement le temps consacré aux calculs, alors que le sens de ce calcul - le sens des unités et la manière de les faire fonctionner - est inaccessible soit à un appareil sans vie, soit à une personne qui ne connaît pas les mathématiques.

Le caractère idéaliste des exercices épistémologiques des cybernéticiens n'est pas moins évident dans leurs tentatives de déduire un critère de vérité de la connaissance à partir

du travail des mécanismes de calcul.

Où est la garantie que les calculs effectués sont corrects ? Les cybernéticiens recourent à toutes sortes d'astuces pour « prouver » que les machines elles-mêmes sont capables de vérifier la fiabilité des conclusions qu'elles reçoivent.

À Philadelphie, un compteur binaire a été construit, composé de deux sous-compteurs, qui calculaient simultanément à la même vitesse, et les résultats obtenus étaient automatiquement vérifiés.

Selon les cybernéticiens, la capacité du « cerveau et des autres appareils informatiques » à résoudre correctement les problèmes intellectuels et à ne pas devenir victime d'illusions repose sur un principe similaire d'auto-vérification.

L'incohérence de toute cette fausse argumentation est évidente pour tout lecteur impartial.

Si les deux compteurs - ou tout autre nombre d'entre eux - avaient la même erreur de conception, ils donneraient le même résultat en raison de leur fonctionnement synchrone, ce qui serait également incorrect.

Deux cybernéticiens – ou n'importe quel autre nombre d'entre eux – peuvent répéter les mêmes propositions idéalistes éculées avec la même persistance et en tirer les mêmes conclusions erronées, mais cela ne rendra ni les propositions ni les conclusions fiables.

Les affirmations des cybernéticiens sont incroyables. Ils prétendent détenir entre leurs mains une clé universelle, non seulement pour les problèmes physiologiques, psychologiques et épistémologiques, mais aussi pour tous les autres problèmes, en particulier pour les problèmes très actuels de la sociologie.

Ils tentent de transférer les principes et les méthodes de leur pseudoscience au comportement des groupes humains.

Partant de l'affirmation selon laquelle les lois de l'activité d'un individu ne seraient pas différentes des règles de fonctionnement d'un thermostat dans un réfrigérateur ou d'un gyrocompas sur un navire, les cybernéticiens tentent ensuite d'interpréter l'ensemble de la société comme un ensemble d'appareils automatiques.

Il serait possible de trouver l'explication de l'interaction dans une expression mathématique appropriée.

La spécificité de la « communauté » de robots, selon les cybernéticiens, est que les gestes ou les paroles fonctionnent comme une impulsion qui met en mouvement le « mécanisme social ».

Dans le même temps, l'arme la plus complexe d'échange de pensées créée par les peuples au cours des siècles, à savoir le langage, est à son tour décrite par les cybernéticiens comme un ensemble de processus physiques - des vibrations d'ondes sonores.

Craignant la volonté et la raison du peuple, les cybernéticiens se consolent à l'idée de la possibilité de transférer les fonctions vitales inhérentes à l'homme à des appareils automatiques.

Est-il possible de remplacer le prolétaire qui se tient sur la chaîne de montage, qui fait grève lorsque les salaires sont réduits, qui vote pour la paix et les communistes, par un robot doté d'un cerveau électronique ?

Ne pourrions-nous pas envoyer un monstre de métal sans émotion au lieu d'un pilote qui refuse de tuer les femmes travaillant dans les rizières ?

Dans une tentative frénétique de réaliser ses plans agressifs, l'impérialisme américain met tout en jeu : bombes, puces de peste et ignorants philosophes.

C'est grâce aux efforts de ces derniers que la cybernétique a été fabriquée - une fausse théorie extrêmement hostile au peuple et à la science. »

L'amélioration des calculs est une bonne chose, mais la négation de la dignité du réel est toujours la caractéristique d'une pseudo-science.

Tout comme aucune science ne peut faire l'impasse sur le principe de fusion, de synthèse, de combinaison, de naissance d'ensembles plus grands et se superposant : on a par exemple le chemin particules élémentaires – noyaux – atomes – molécules – masses - corps célestes -nébuleuses, celui solides – liquides – gaz, celui molécules – colloïdes – cellules – organes – individus - sociétés.

L'intelligence artificielle expose la limite d'une époque – le matérialisme dialectique supprime cette limite pour relancer le mouvement à l'infini.