

Néguantropie, opacité et explicabilité des réseaux neuronaux artificiels profonds

Johan Mathe johmathe@baylabs.io

1 Introduction

Je voudrais commencer par remercier Bernard Stiegler qui m'a demandé d'intervenir pendant ce séminaire cet été après une discussion sur la disruption au sein de l'entreprenariat dans la région de San Francisco. En tant qu'in génieur, j'ai vu ce projet comme un exercice de création d'un objet temporel. Je suis allé voir il y a plus d'un mois maintenant 2001 a space odyssey de Kubrick joué par l'orchestre symphonique de San Francisco. Ce film a toujours eu une place particulière dans ma vie, je l'ai vu pour la première fois à l'âge de 10 ans, puis tous les 5 ans jusqu'à aujourd'hui. C'était donc mon cinquième visionnage. Le fait d'avoir la musique jouée par un orchestre symphonique a mis en exergue la place centrale de Strauss dans cette œuvre de Kubrick, qui elle-même fait le pont vers Ainsi Parlait Zarathustra. J'ai donc décidé de travailler sous la contrainte pour cette présentation et d'utiliser une trame narrative qui va faire un parallèle entre certains de mes travaux de recherche et l'analyse de 2001.

Le premier volet du film montre un homme primitif confronté à la découverte de la première exosomaturation devant un monolithe noir. Kubrick fait alors une des ellipses narratives les plus longues de l'histoire. On arrive à ce qu'il considérera comme les dernières exosomaturations : la conquête de l'espace et la naissance de l'intelligence artificielle.

2 Baylabs

Baylabs, Inc est une startup qui a pour mission d'améliorer la qualité, valeur et l'accès à l'imagerie médicale. Le premier produit sur lequel l'équipe travaille est un système d'analyse cardiaque basé sur l'utilisation de l'ultrason et de l'échocardiographie. Nous travaillons à l'intersection de deux technologies : la miniaturisation des techniques d'acquisition de données de scans ultrasons ainsi que l'utilisation des réseaux neuronaux profonds. Nous travaillons actuellement en partenariat avec cinq universités et centres hospitaliers américains : Stanford University, Northwestern University à Chicago, Duke University sur la côte est ainsi que Minneapolis Heart Institute. Un de nos premiers prototypes permet de faire du diagnostic des signes avant-coureur de la fièvre rhumatismale. La fièvre rhumatismale est une complication des infections de l'enfance et de l'adolescence, qui survient à la suite des angines dues au streptocoque hémolytique, et qui n'ont pas été soignées par des antibiotiques. Cette pathologie est assez facile à traiter avec des antibiotiques, mais difficile à diagnostiquer sans avoir de vision interne du cœur.

3 Réseaux neuronaux artificiels profonds

Je vais maintenant introduire les réseaux de neurones artificiels profonds. Cette introduction est largement inspirée d'une publication en anglais [1]. Le but d'un réseau neuronal est d'approximer une fonction f^* . Un bon exemple une fonction de classification d'image $y = f^*(x)$ qui associe une classe de donnée à une catégorie y . Le réseau neuronal définit une application $\mathbf{y} = f(\mathbf{x}; \mathbf{W})$. On ne mentionnera pas ici des réseaux neuronaux récurrents. Ces modèles sont la base d'énormément d'applications. La définition de ces modèles est extrêmement simple. En effet il s'agit généralement de la composition de fonctions non-linéaires de la forme suivante :

$$f(\mathbf{x}) = f^{(n)}(\dots f^{(i)}(\dots f^{(1)}(\mathbf{x})))$$



Fig. 1: SF Symphony et 2001

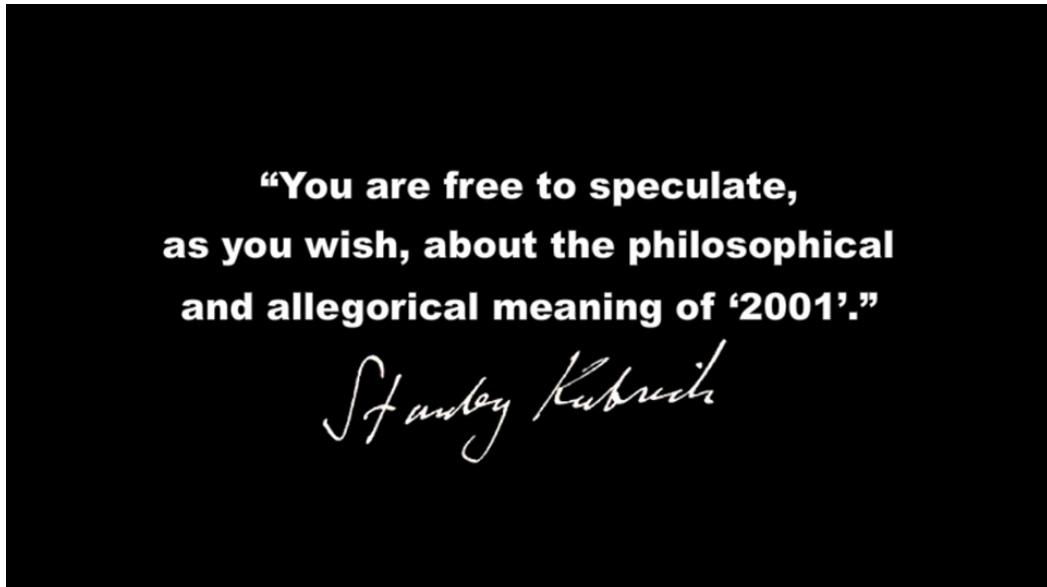


Fig. 2: Citation de Kubrick

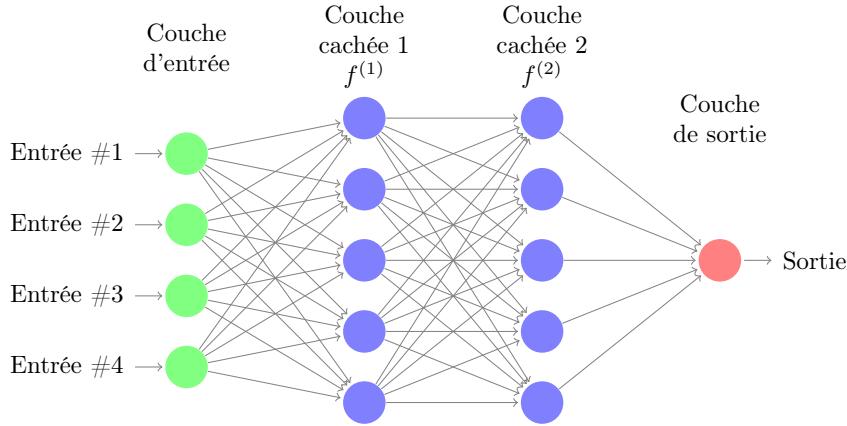


Fig. 3: John utilisant le prototype EchoMD

Généralement, les fonctions f ont la forme suivante, avec $\sigma(x)$ une fonction non linéaire comme par exemple l'unité linéaire rectifiée (voir figure) :

$$f^{(i)}(\mathbf{x}) = \sigma(\mathbf{W}_i^T \mathbf{x} + \mathbf{b}_i) \quad (1)$$

Ces structures chainées peuvent être vues comme un système multi couches. En effet la fonction $f^{(1)}$ représentera la première couche du réseau, jusqu'à $f^{(n)}$ la dernière couche n . Le nombre de couche est appelé la **profondeur** du réseau. Un modèle graphique représentant un réseau à deux couches :



3.1 L'apprentissage

L'apprentissage est un processus souvent itératif qui permet d'estimer une fonction de type réseau neuronal grâce à une grande quantité de données d'entrée. Dans notre cas de classification d'images pathologiques ou non, nous aurons un ensemble de tuples (X_i, y_i) qui correspondront tout simplement aux images et à leurs labels, c'est à dire s'il s'agit d'un cas pathologique ou non.

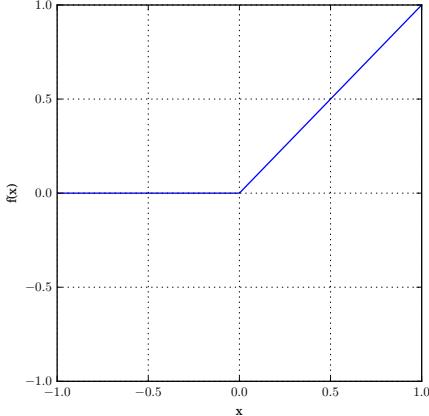


Fig. 4: Une non-linéarité de type $\text{relu}\sigma(x) = \max(0, x)$

D'un point de vue mathématique on peut voir le problème de l'apprentissage comme un problème d'optimisation. Il s'agit de minimiser la somme des erreurs entre les prédictions de notre réseau de neurones et les données labellisées par des experts (dans notre cas les experts du monde médical). L'outil mathématique utilisé dans ce cas est une loss function TODO() : traduire loss. Cette fonction a pour but d'augmenter quand l'erreur de prediction augmente, et de diminuer le cas contraire.

Un exemple classique d'une loss TODO translate est la distance euclidienne, aussi appelée loss euclidienne. Elle représente la distance euclidienne qui sépare 2 points dans un espace donné.

$$l(\hat{y}, y) = \|\hat{y} - y\|_2 = \sqrt{\frac{1}{n} \sum_{i=1}^N (\hat{y}_i - y_i)^2}$$

Dans notre cas cet espace est l'espace des paramètres et nous retrouvons donc avec la forme suivante :

$$\text{minimiser } J(\boldsymbol{\theta}, \mathbf{y}, \mathbf{X}) = \sum_{i=1}^N \|y_i - f(X_i; \boldsymbol{\theta})\|_2^2$$

La minimisation de cette valeur se fait par l'algorithme du gradient (aussi appelé gradient descent, steepest descent). L'idée relativement simple. On commence par choisir un point aléatoire $\boldsymbol{\theta}_0$, puis on évalue la valeur du gradient $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ en ce point. Ce gradient représente un hyperplan dans l'espace à m dimensions des paramètres de notre fonction $f(x; \boldsymbol{\theta})$. On fait ensuite une mise à jour de notre paramètre qui devient $\boldsymbol{\theta}_1$ en lui ajoutant une partie de ce gradient, pondéré par un taux d'apprentissage α :

$$\boldsymbol{\theta}[k+1] = \boldsymbol{\theta}[k] + \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}, \mathbf{y}, \mathbf{X})$$

La figure 5 illustre l'algorithme du gradient pour une fonction convexe quadratique. Dans ce cas la le gradient devient simplement la dérivée de la fonction en un point. Les droites rouges illustrent les tangentes de la droite en chaque point et montrent comment l'algorithme évolue. Après quelques itérations on considère un critère d'arrêt comme par exemple le fait que la valeur du gradient soit proche d'une valeur arbitrairement faible ϵ .

3.2 Convexité

Une des différences notables entre les algorithmes de type réseaux de neurones profonds et les méthodes plus anciennes d'apprentissage supervisé comme par exemple une SVM TODO citation SVM est la non-convexité du problème. Nous allons revenir quelques instants sur la définition de

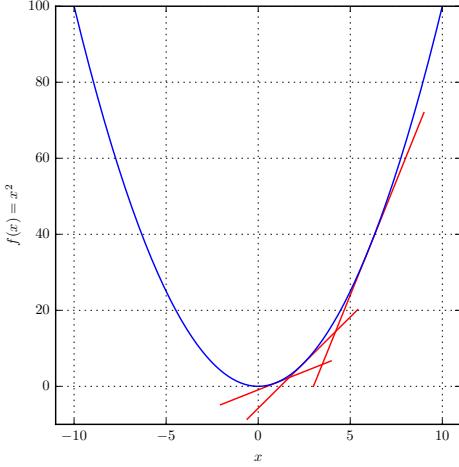
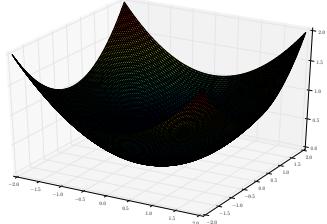
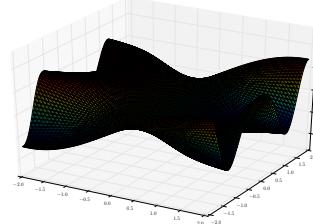


Fig. 5: Illustration de la méthode du gradient

Fig. 6: Une fonction convexe de $\mathbb{R}^2 \mapsto \mathbb{R}$ Fig. 7: Une fonction non convexe de $\mathbb{R}^2 \mapsto \mathbb{R}$

la convexité, car celle ci est capitale pour la suite de cet exposé. L'ensemble convexe est défini de sorte a ce que si l'on choisit deux points x, y dans cet ensemble, tous les points formés par l'ensemble des barycentres entre ces deux points appartiendront aussi a l'ensemble (voir équation 2)

$$\forall x, y \in C \quad \forall t \in [0, 1] \quad tx + (1 - t)y \in C \quad (2)$$

La définition d'une fonction convexe est assez proche de celle de l'ensemble convexe (voir équation 3).

$$f(tx + (1 - t)y) \leq t f(x) + (1 - t) f(y) \quad (3)$$

Une caractéristique principale des fonctions convexes est qu'elle ne possède qu'un seul maximum, celui ci étant donc de fait un maximum global. Ceci implique donc des garanties importantes quant a la convergence de notre fonction d'objectif. Nous avons la garantie d'obtenir la solution au probleme d'optimisation.

Les réseaux neuronaux profonds étant composés d'une succession d'opérations linéaires (convexes) et non-linéaires (non convexes), la composition de fonctions non convexes est généralement une fonction non-convexe. Ceci implique que dans le cas général, les solutions obtenues par des algorithmes de la famille des algorithmes du gradient ne nous donnent pas une garantie sur l'optimalité et la globalité du résultat trouvé.

3.3 Une nouvelle fonction d'erreur : minimisation de l'entropie croisée

On veut quantifier l'information d'une maniere qui formalise une forme d'intuition.

- Les evenements qui sont fortement probable devraient avoir un contenu en information faible, et les evenements qui sont garantis devraient avoir un contenu informationnel proche de zéro. Un exemple typique est que le soleil se levera demain. Cette phrase a un contenu informationnel faible si l'on connaît l'histoire des leveres et couchers de soleils depuis le début de l'histoire de l'humanité.
- Les evenements peu probables devraient avoir un contenu informationnel élevé. Les évenemens independants devraient avoir une information additive. Par exemple, se rendre compte que lors d'un lancer de dés est tombé sur pile deux fois

Shannon a théorisé dans son papier **A Mathematical Theory of Communication** l'entropie comme suit :

$$I(x) = -\log P(x)$$

Différents types d'entropie :

- Entropy de gibbs
- Entropy de von neumann

Self-information deals only with a single outcome. We can quantify the amountof uncertainty in an entire probability distribution using the Shannon entropy :

$$H(x) = \mathbb{E}_{x \sim P}[I(x)] = -\mathbb{E}_{X \sim P}[\log(x)]$$

$$C = -\frac{1}{n} \sum_x [y \ln a + (1 - y) \ln(1 - a)]$$

3.4 Entropie Croisée

Shannon a théorisé l'entropie d'un point de vue informationel dans sa publication de 1948, **A Mathematical Theory of Communication** [2]

$$H_{y'}(y) = \sum_i y'_i \log(y_i) \tag{4}$$

4 Opacité, Explicabilité

Dans 2001 l'odyssée de l'espace, le kubrick met en scene un monolithe a trois reprises, dans des moments clés de l'intrigue :

- Au debut juste avant que le post neanderthal mette en place sa premiere exosomatisation
- Juste avant l'apparition de HAL9000, qui est finalement l'exosommatisaion ultime
- A la fin, pour la transition entre l'homme et l'übermensch

Dans l'oeuvre originale de l'écrivain Arthur C Clarke de qui le film est inspiré, le monolithe a une forme de pyramide. Kubrick décide de remplacer cette forme pyramidale en forme parallélépipédique, qui ressemble finalement à une boîte noire. Quand Kubrick dans les années 60 a fait ses recherches sur les balbutiements de l'intelligence artificielle, il a passé du temps avec Minsky au MIT, pour essayer de comprendre les tenants et les aboutissants de l'IA. En théorie des systèmes, une boîte noire est une représentation d'un système pour lequel seulement ses entrées et ses sorties sont observables.

Je vais maintenant décrire des travaux qui sont liés a des axes de recherches actifs et récents. En particulier les problématiques d'explicabilité et d'opacité. J'utilise ici la définition suivante de l'opacité d'un réseau de neurones artificiel :

Un réseau de neurone artificiel est dit opaque de par la nature non-convexe du probleme d'apprentissage, nous n'avons aucune visibilité sur les performances ni sur les criteres internes qui font que ce dernier perform (TODO) bien ou mal TODO rephrase.

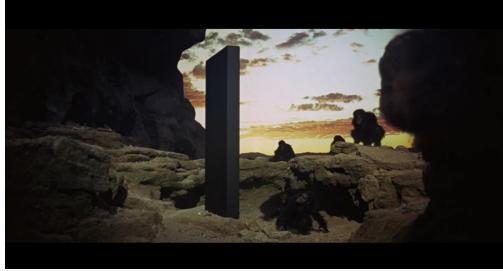


Fig. 8: Première apparition du Monolithe

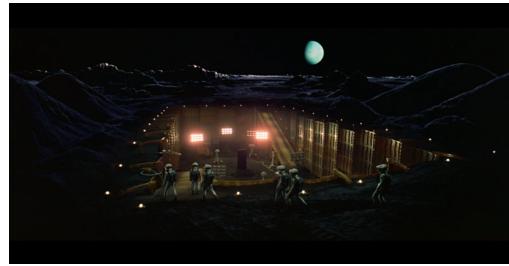


Fig. 9: Deuxième apparition du Monolithe

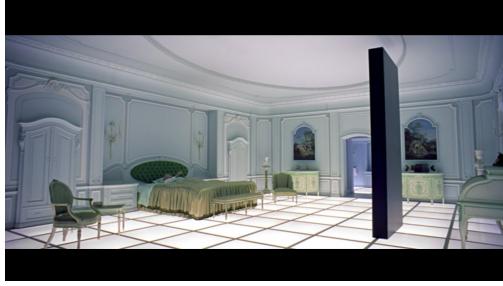


Fig. 10: troisième apparition du Monolithe



Fig. 11: Monolithe

L'explicabilité ici est définie comme une production d'information qui va déterminer les critères d'un signal d'entrée qui ont amené le réseau de neurones à faire un choix plutôt qu'un autre.

4.1 Opacité

Une des plus grandes problématiques posées aujourd'hui par les réseaux de neurones profonds est liée à ce qu'on peut définir comme l'opacité et le manque d'explicabilité de ceux-ci.

Pour diminuer l'opacité de ce type d'algorithme, nous nous penchons sur une technique relativement ancienne, mais qui a été popularisée récemment grâce aux travaux de Erhan et al 2009 TODO Citations. Il s'agit simplement de l'opération inverse de l'apprentissage. Nous avons vu que l'apprentissage consistait simplement en une optimisation d'une fonction d'objectif visant à minimiser une erreur de prédiction, ou encore une entropie croisée entre une distribution attendue et une distribution produite par le réseau de neurones.

Ici nous allons maximiser une activation en considérant la fonction inverse de notre réseau de neurones :

$$h^*(\mathbf{X}, \boldsymbol{\theta}) = f^{*-1}(\mathbf{X}, \boldsymbol{\theta})$$

Nous solvons donc le problème d'optimisation suivant :

$$\mathbf{X}^* = \text{argmax} h(\boldsymbol{\theta}, \mathbf{X})$$

Nous pouvons réutiliser la méthode de l'algorithme du gradient présentée lors de l'introduction de l'apprentissage (3.1).

4.2 Explicabilité

Ici pour l'explicabilité, nous utilisons les approches de ziel et fergus TODO citation fergus 2014 qui consiste à observer la classification des résultats d'un réseau de neurone déjà entraîné dans le cas où une image aura une zone de taille (n, n) pixels qui sera « effacée ». TODO : image avec patch effacé. On déplace ensuite cette image d'un pixel de sorte à créer une heatmap (TODO : translate)



Fig. 12: Un cas pathologique de fièvre rhumatisque

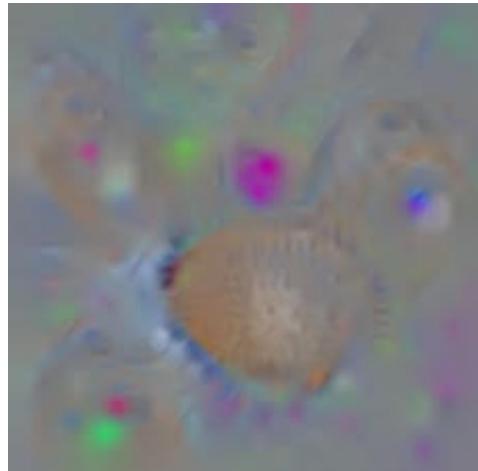


Fig. 13: Analyse de l'opacité, exemple 1

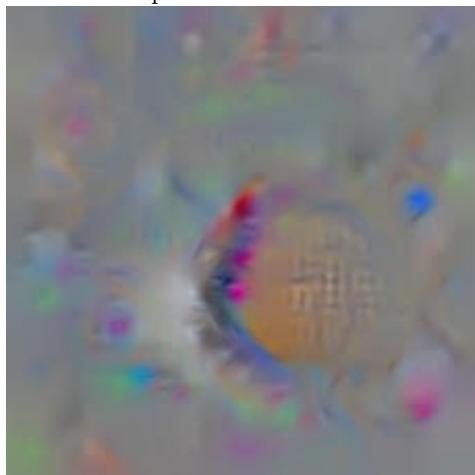


Fig. 14: Analyse de l'opacité, exemple 2



Fig. 15: Un cas pathologique de fièvre rhumatisante

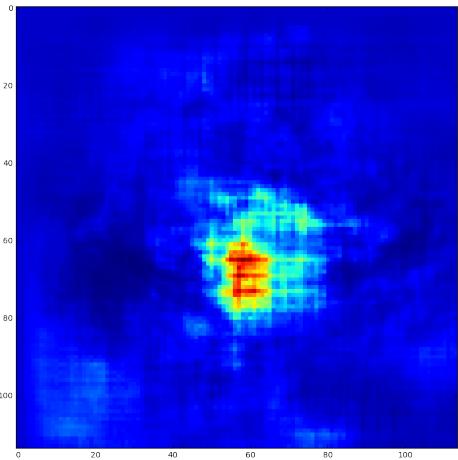


Fig. 16: Analyse de l'opacité, exemple 1

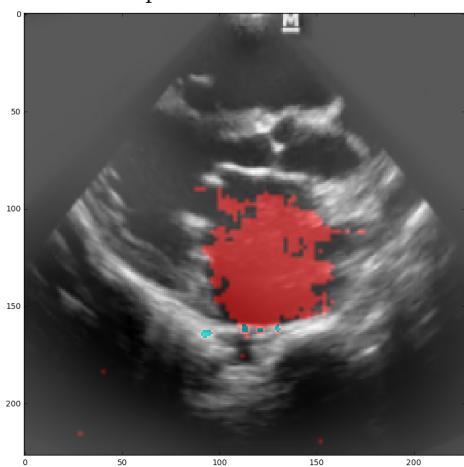


Fig. 17: Analyse de l'opacité, exemple 2

Faux positifs, qu'est ce qu'on extrait de ces images ?

5 Néguantropie, prolétarisation, circuits longs de la transindividuation

Une des motivations pour ce travail de recherche et développement autour de l'explicabilité et de l'opacité au sein de Bay Labs vient de plusieurs discussions avec les professionnels du corps médical. Quand nous avons commencé à travailler avec les cardiologues sur ces questions d'apprentissage et de réseaux neuronaux, un de leur retours principaux était la question de la prolétarisation. En effet en tant que chercheurs, ingénieurs et entrepreneurs, nous fournissons un outil aux médecins en vue de produire un diagnostic. Il est critique que le médecin qui a la responsabilité finale du diagnostic comprenne pourquoi l'outil fournit tel ou tel diagnostique.

De plus, depuis le début de l'aventure, nous tentons de penser les circuits longs de la transindividuation entre le corps médical et notre équipe de recherche. C'est à mon sens l'utilisation continue de ces boucles de renforcement entre les médecins qui peuvent plus que jamais utiliser leur esprit critique. En effet si nous conservons la trace des événements bugs on peut faire bien. Aussi nous avons remarqué que dans certains cas eg afrique et biais

Finalement Kubrick positionne la boîte noire - le monolith - avant chaque transition. Ces boîtes noires agissent comme des marqueurs de temps.

- Dans le premier cas, nous avons du recul pour comprendre le fonctionnement d'un outil aussi primaire que le bâton. Il nous montre notre passé. Il montre aussi que l'homme primitif ne comprenait pas vraiment la mécanique classique. Pour nous l'utilisation de cet outil est tout à fait transparente. Pour le singe c'est une boîte noire.
- Dans le deuxième cas, Kubrick montre les dangers de concevoir une AI de façon boîte noire. En effet la gestion des erreurs devient impossible (la crise commence au moment où les ingénieurs à bord se rendent compte que HAL9000 a commis une erreur. La confiance est perdue.)
- Dans le dernier cas, totalement oenirique, Kubrick à mon sens théorise la singularité (voir figure) en empruntant la thématique de l'übermensch de Nietzsche. Nietzsche définit l'übermensch comme l'homme supérieur pouvant s'élever au-dessus de la morale chrétienne et imposer ses propres valeurs. Il nous met face à un scénario qui dépassent notre entendement cartésien et joue sur nos peurs primaires (vieillissement prématûr de l'acteur dans un environnement vicié) - avant d'avoir la renaissance finale. Avec le thème de Strauss reprenant le relai. On se retrouve dans un cas typique Kubrick ne compose plus seulement avec la musique et l'image, mais avec le subconscient du récepteur. La renaissance de l'homme en « star child ». La boîte noire est là pour nous rappeler que les techniques sont souvent découvertes par hasard. L'homme singe n'a pas attendu la mécanique newtonienne pour servir du bâton - hors c'est celle-ci qui va expliquer avec exactitude quelle énergie il devra déployer pour l'utiliser précisément.

Ce que je propose ici, c'est que la néguanthropie passe avant tout par l'augmentation de l'explicabilité de la techné. Cela passe par l'utilisation des circuits longs de la transindividuation inter-domaine.

5.1 Une proposition pharmacologique

Je propose d'utiliser une approche pharmacologique lors de l'apprentissage. Rapelons-nous

$$\text{minimiser } J(\boldsymbol{\theta}, \mathbf{y}, \mathbf{X}) = \sum_{i=1}^N \|y_i - f(X_i; \boldsymbol{\theta})\|_2^2$$

Maintenant si nous définissons une fonction d'explicabilité $E(\boldsymbol{\theta})$ et une fonction d'opacité $O(\boldsymbol{\theta})$, nous pouvons définir une fonction d'erreur pharmacologique J_p :

$$\text{minimiser } J_p(\boldsymbol{\theta}, \mathbf{y}, \mathbf{X}) = J(\boldsymbol{\theta}, \mathbf{y}, \mathbf{X}) + \alpha O(\boldsymbol{\theta}) - \beta E(\boldsymbol{\theta})$$

J'aimerais que nous discutions ensemble d'un type d'algorithme que je définis comme socio inspiré. En effet il nous incompe de définir ces fonctions O et E en s'inspirant des sciences de l'éducation et de l'apprentissage et de la pédagogie

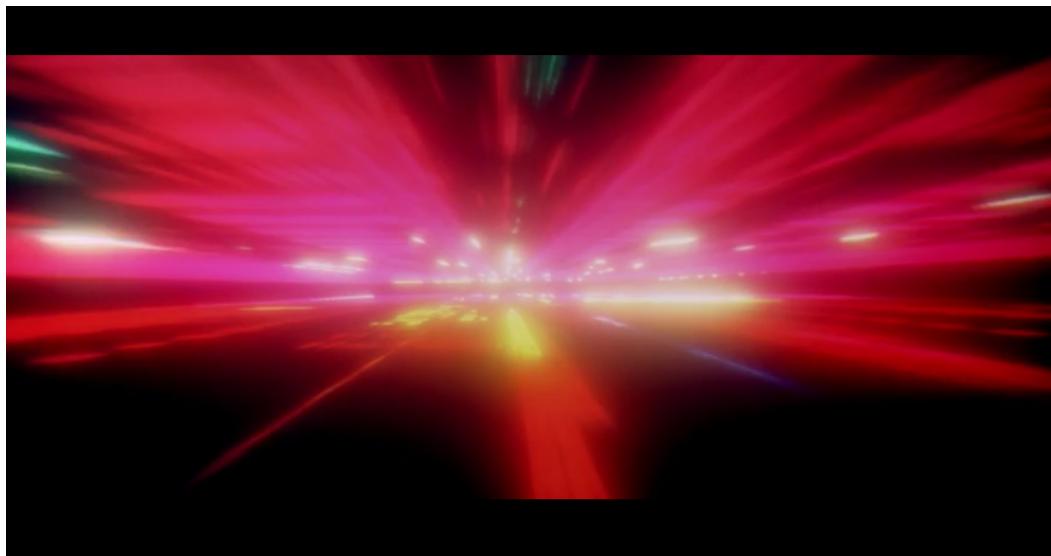


Fig. 18: proposition : Singularité selon Kubrick



Fig. 19: proposition : ubermesnch

Science vs explicabilité :

Références

- [1] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep learning. Book in preparation for MIT Press, 2016.
- [2] Claude E. Shannon and Warren Weaver. *A Mathematical Theory of Communication*. University of Illinois Press, Champaign, IL, USA, 1963.