Les limites de la connaissance…

La **science** (latin scientia, « connaissance ») :

« **ce que l'on sait pour l'avoir appris,**
ce que l'on tient pour vrai au sens large,
l'ensemble de connaissances, et des études d'une valeur universelle, caractérisées par un domaine et une méthode déterminée, et fondée sur des relations objectives vérifiables
[au sens restreint impliquant la reproductibilité] »

La science pourra t'elle tout prévoir, tout calculer, tout démontrer ?

Non…

Depuis le XX^{ème} siècle,

les mathématiciens et les physiciens
ont découvert plusieurs limites irréductibles au savoir[1].

(1) Hervé Zwirn – Pour la Sciences n°422 – Décembre 2012 – pp. 45-50

+

(2) Cyrille Imbert (CNRS – Ulm – Philo des Sciences) - Thèse : « L'opacité intrinsèque des phénomènes. Théories connues, phénomènes difficiles à expliquer
et **limites de la science** »

Pourquoi non ?

- Les mathématiques et la physique ne permettront pas de tout comprendre. Plusieurs théories ont montré les limites du savoir.
- Les **théorèmes d'incomplétude de Gödel** sont parmi les premiers exemples de limites fondamentales en mathématique.
- Il existe des limites cognitives et ontologiques (en lien avec l'étude de l'être, de ses modalités et de ses propriétés).
  Il est en effet impossible de calculer certains objets mathématiques/physiques et même de savoir si certains existent.
- Les limites prédictive : il n'est pas possible de déterminer l'état du système au-delà d'un avenir limité (Ex: phénomènes météo).
- Les systèmes chaotiques, que l'on ne peut étudier que par **simulation.**

Gödel : 1 Hilbert : 0

En 1931, le logicien autrichien Gödel publia **2 théorèmes d'incomplétude** qui ont brisé le rêve du grand mathématicien Hilbert qui rêvait de construire une mathématique où tout énoncé serait démontrable sans ambiguïté.

Ces théorèmes indiquent que **toute mathématique suffisamment riche pour contenir l'arithmétique contient des énoncés indécidables**, et que la non-contradiction du système est l'un deux.

Gödel **prouva formellement** que **dans tout cadre mathématique, il existe des énoncés qui, bien que vrais, ne pourront jamais être démontrés dans ce cadre**.

Ces travaux et d'autres ont eu des conséquences sur la réflexion philosophique et notamment sur la prise de conscience des limites de la connaissance humaine.

Face au **scientisme** du XIXème sciècle qui croit pouvoir tout connaître,
le physiologiste allemand, Emil du Bois-Raymond, exprime en 1872 :
« ignoramus et ignorabimus » - **nous ne savons pas et nous ne saurons jamais**.

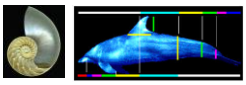
Tout expliquer ?

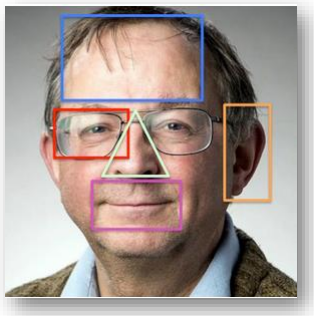
L'idée de tout expliquer n'est aujourd'hui plus tenable **en raison de la mise en évidence par la science elle-même, des limites infranchissables de son propre discours**.

1. Les limites ***prédictives*** : montrent l'impossibilité de prévoir certains phénomènes avec une précision arbitraire sur une échelle de temps indéterminée.
   (Ex : en physique dans la théorie du chaos, la météo,…).
2. Les limites ***cognitives*** concernent l'existence de domaines qui restent hors de portée du savoir. De telles limites apparaissent en mathématique dans l'étude de **nombres parfaitement définis mais incalculables** – et de nombres qu'on ne peut déterminer au mieux qu'avec un nombre fini de décimales.
3. Les limites ***constructives*** qui sont relatives à **l'impossibilité de construire un discours scientifique qui échappe à tout doute et qui repose sur des fondements sûrs**.
4. Enfin, les limites ***ontologiques*** qui éliminent quelques entités conceptuelles en montrant leur inconsistance **ou en les situant en dehors du champ d'appréhension du discours scientifique** (et parfois au-delà de la physique : métaphysique).
   C'est les cas en physique quantique avec le rôle privilégié de l'observateur, ou en mathématique, avec la question de l'existence ou non de certains objets. Exemple: principe d'incertitude d'Heisenberg (position ou vitesse d'une particule ?)

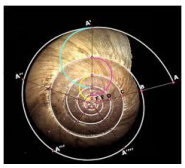
Phi : almost everywhere in real life… Myth or reality ?

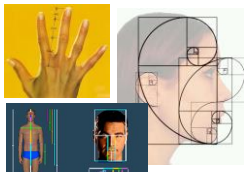
$$\varphi = \frac{1+\sqrt{5}}{2} = 1.6180339887\ldots$$

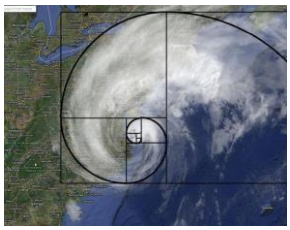


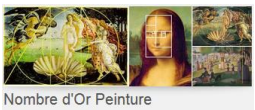



Nombre d'Or Nature

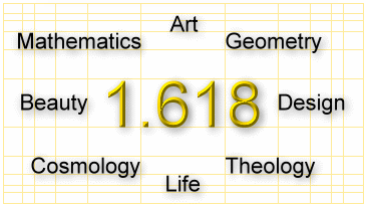







Nombre d'Or Peinture


Nombre d'Or Architecture

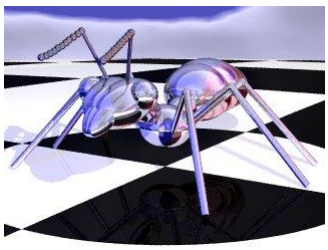


An algorithmic example
Cellular Automaton
Langton's Ant

The Langton ant is a two-dimensional cellular automaton with a very simple set of rules.

It was named after Christopher Langton, his inventor, a pioneer in artificial life.

It is one of the simplest systems for highlighting an example of emerging behavior.



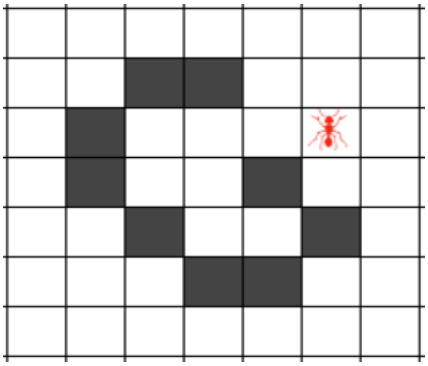
2D black & white grid

Squares on a plane are colored variously either black or white.

We arbitrarily identify one square as the "ant".

The ant can travel in any of the four cardinal directions at each step it takes.

Despite very simple rules, a complex phenomenon is emerging.

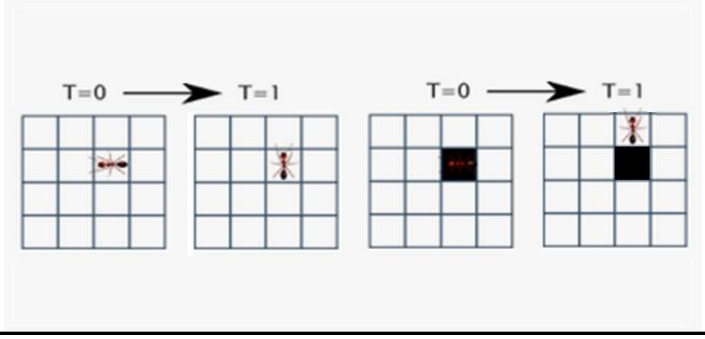


Ant Simple Rules

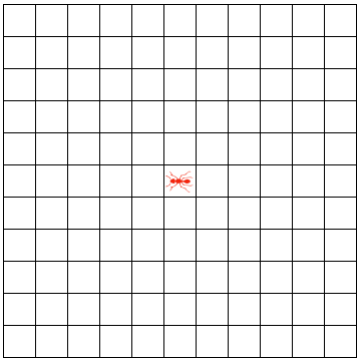
The "ant" moves according to the rules below:

At a white square, turn 90° right, flip the color of the square, move forward one unit

At a black square, turn 90° left, flip the color of the square, move forward one unit

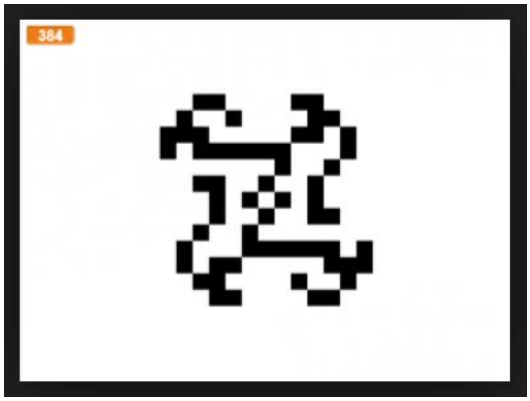


Let's have a look at the first 200 steps



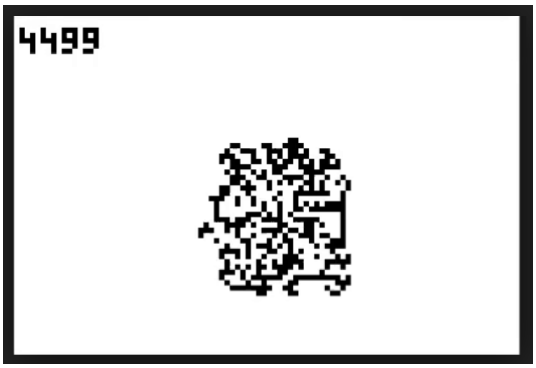
1. Simplicity

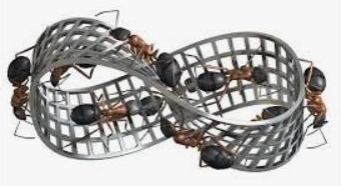
During the first few hundred moves it creates very simple patterns which are often symmetric.



2. Chaos

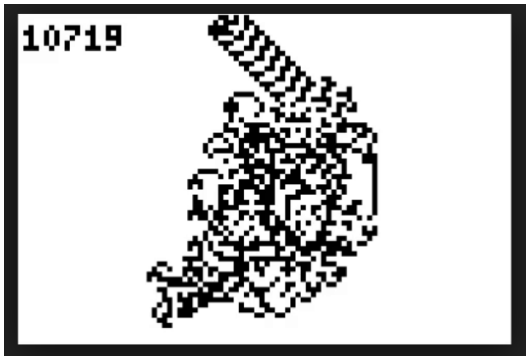
After a few hundred moves,
a big, irregular pattern, of black and white squares appears. The ant traces a pseudo-random path until around 10,000 steps.





3 . Emergent order

Finally the ant starts building a recurrent "highway" pattern of 104 steps that repeats indefinitely.



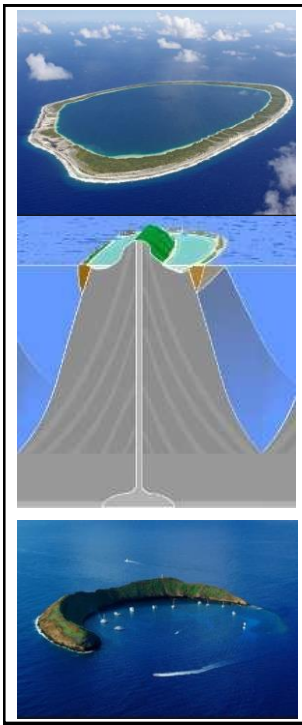
Hidden complexity

These simple rules lead to a complex behavior.

Three distinct modes of behavior are apparent, when starting on a completely white grid.

We will see emergence as a form of underlying and previously hidden complexity.

What can we expect with complex specifications ?



Emergence ?

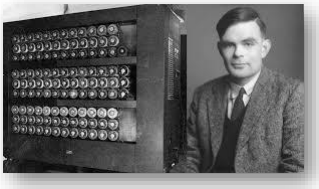
The **emergence**, evokes the idea of an unveiling, an apparition that takes shape, but that was somewhere already there,
under an appearance still elusive.

This is the case or emerging land (hidden below water) or the case of an iceberg.

It is the case of continents, or **atoll island** rooted under the surface of the water, but really take shape for our eyes when they come out.

The first picture could **disappear**…

En fait, la théorie a montré que depuis longtemps que l'informatique est "indécidable"…



L'indécidabilité d'un problème signifie qu'il ne peut être résolu par aucun algorithme en un temps fini.

Cela signifie que peu importe la puissance d'un ordinateur, il ne pourra jamais résoudre certains problèmes.

Beaucoup de questions sont indécidables : Il est indécidable de savoir si un programme va s'arrêter, si un programme est un virus, si un programme est équivalent à un autre…

Les réponses sont accessibles acceptables pour de nombreux programmes simples, mais parfois non, et si le programme est gros ou complexe, nous n'avons pas de méthode générale. **Nous n'avons pas de théorie ou de modèle capable de nous aider à répondre** et nous n'en aurons jamais.

Chaitin's Ω number…

https://en.wikipedia.org/wiki/Chaitin's_constant

In the computer science subfield of algorithmic information theory, a Chaitin Omega number or halting probability is a real number that, informally speaking, represents the probability that a randomly constructed program will halt.

https://www.youtube.com/watch?v=LGYlT6DsfH8&ab_channel=Science4All

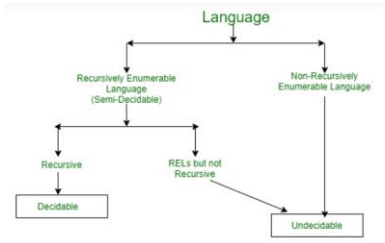
Since the halting problem is undecidable, Ω cannot be completely computed.

This a an incompleteness theorem similar to Gödel's incompleteness theorem in that it shows that no that no consistent formal theory for arithmetic can be complete.

L'apprentissage aussi est "indécidable"…

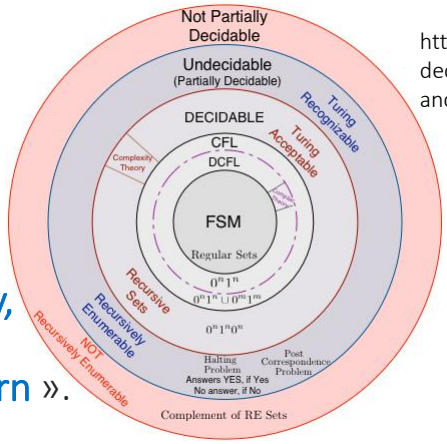
C'est le cœur de l'intelligence artificielle (l'apprentissage) qui a été démontré indécidable.

BEN-DAVID, Shai, HRUBEŠ, Pavel, MORAN, Shay, et al. "Learnability can be undecidable", Nature Machine Intelligence, 2019, vol. 1, no 1, p. 44-48.



https://www.geeksforgeeks.org/decidability-semi-decidability-and-undecidability-in-toc/

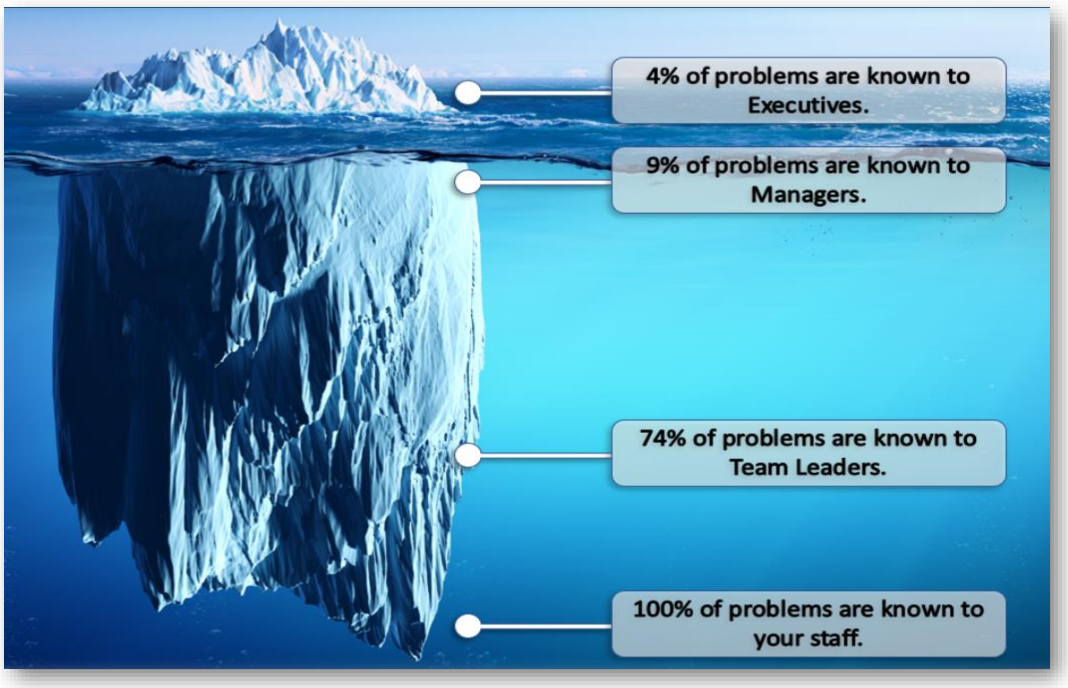
Theorem :
« We cannot know with any certainty, because logic shows it formally, that machines really learn what they are supposed to learn ».



https://sameer9247.wordpress.com/2016/11/16/theory-of-computation-undecidability/

Who (really) knows ?

Right: Yoshida's iceberg in industry…



2022 LLM – GPT 3
Mode développeur ça devait sérieux…

https://www.youtube.com/watch?v=cAhb2jr2_IM&ab_channel=MachineLinguist

2023 Peut-on contrôler ces productions ?

https://www.youtube.com/watch?v=dDhTMIao-fM&ab_channel=MonsieurPhi

L'émergence et l'Intelligence Artificielle

> TECH

"NOUS NOUS SOMMES TROMPÉS": LE COFONDATEUR D'OPENAI REGRETTE L'APPROCHE OPEN SOURCE DE SES DÉBUTS

Marius Bocquet avec AFP Le 16/03/2023 à 12:22

Le CEO de Google sous le choc : cette IA s'améliore toute seule

Bastien L 18 avril 2023 Google, Intelligence artificielle 1 commentaire

Lors d'une interview à la télévision américaine, le CEO de Google a révélé qu'une intelligence artificielle avait appris une langue sans être entraînée pour le faire, de façon autonome et totalement inexplicable. Comment expliquer ce phénomène et pourquoi est-ce un danger ?

WOTF, cette nouvelle secte vénère l'intelligence artificielle

par Yohan Demeure, expert géographe • 17 novembre 2017, 14 h 41 min

Elon Musk accuse le cofondateur de Google de vouloir créer un "dieu numérique"

https://www.lebigdata.fr/ceo-google-choc-ia-ameliore-seule

Entraîné avec du calcul à haute performance sur la mémoire de masses de données des productions humaines, des comportements surprennent, cf. les systèmes complexes

Le mystère des « *propriétés émergentes* »

D'après Sundar Pichai, il arrive que les intelligences artificielles génératives se comportent d'une manière inattendue. C'est pourquoi il est extrêmement important d'avancer avec prudence, en surveillant de près les réactions des IA.

Selon les constatations de Google, des modèles de langage peuvent parfois **s'enseigner de nouvelles compétences en toute autonomie**, à l'insu des programmeurs. L'entreprise a intitulé ce mystère les *« propriétés émergentes »*.

Par exemple, une IA expérimentale, mise au point par Google, est parvenue à apprendre en toute indépendance *« la langue du Bangladesh, alors qu'elle n'a pas été formée à la connaître »*. Présent aux côtés de Sundar Pichai, James Manyika, vice-président principal de la technologie chez Google, explique qu'il a suffi de quelques requêtes pour qu'une IA s'habitue à une langue inconnue :

01net

Google est stupéfait : son IA a appris une nouvelle langue sans aide

https://www.01net.com/actualites/google-stupefait-ia-appris-nouvelle-langue-sans-aide.html

SiecleDigital

Économie Technologie Marketing Retail Cybersécurité Médias Réseaux sociaux Outils

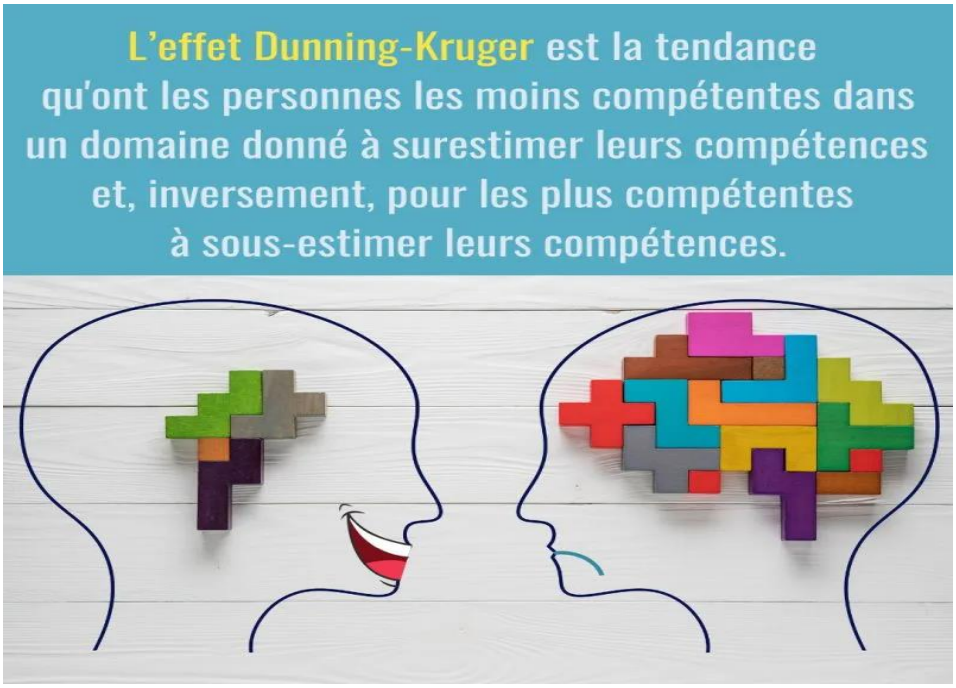
Intelligence artificielle

L'IA générative doit être réglementée comme les armes nucléaires, estime le PDG de Google

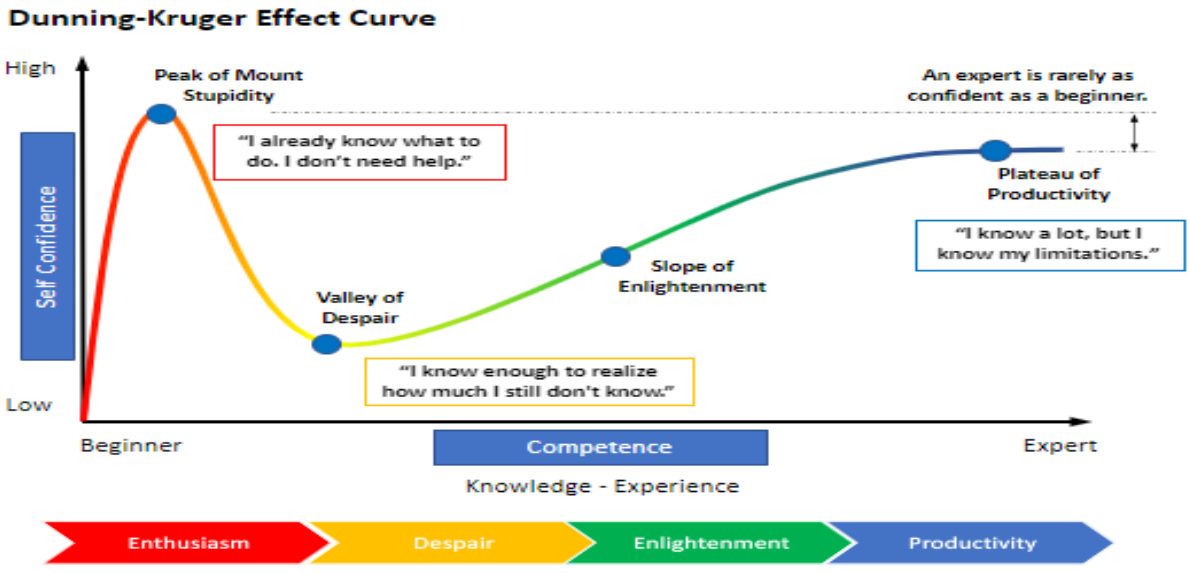
Lors d'une interview, Sundar Pichai a fait part de ses nombreuses inquiétudes dans le cas où l'IA serait « mal déployée »

https://siecledigital.fr/2023/04/18/sundar-pichai-inquietudes-ia-generative/

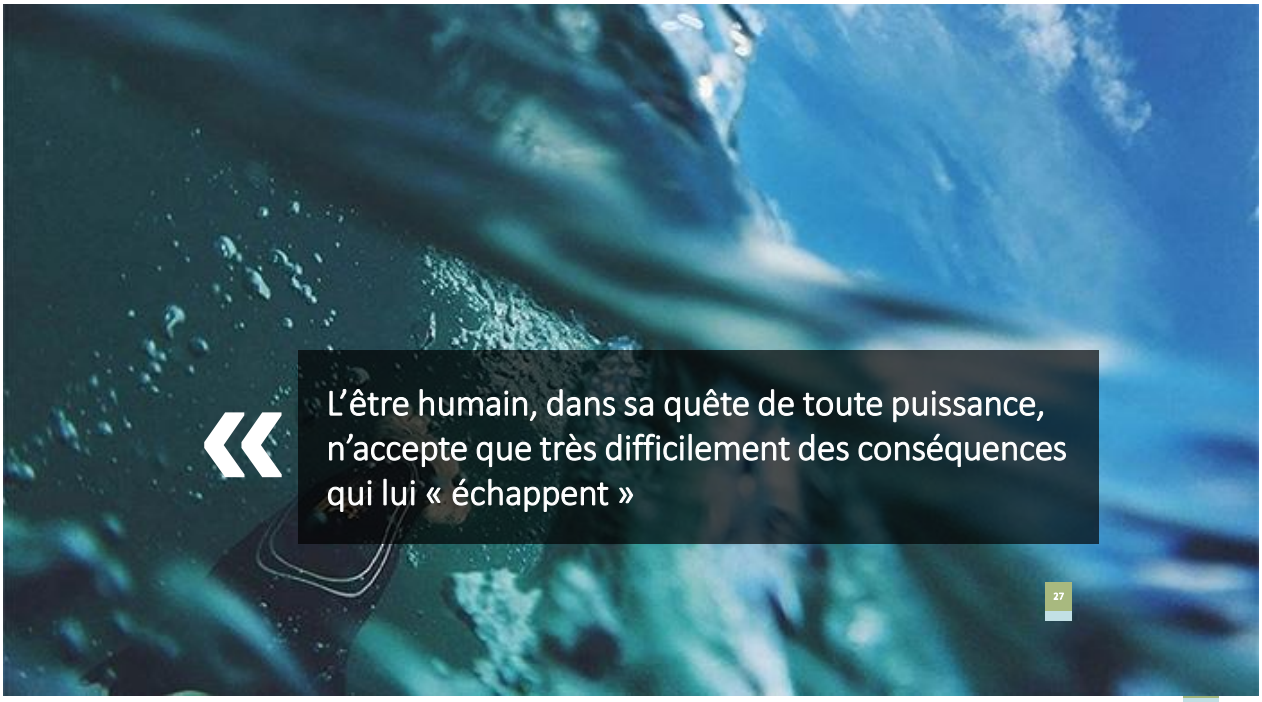
L'Effet Dunning Krugger



Curve of the Denning-Kruger effect



https://www.linkedin.com/pulse/i-wore-juice-dunning-kruger-effect-false-illusion-rafael



Explainable Artificial Intelligence

Faced to the Gödel's wall > "Machine learning undecidability"…

An **Explainable AI** (XAI) or **Transparent AI**
is an artificial intelligence (AI) whose actions can be understood by humans.

It contrasts with the concept of the "black box" in machine learning, meaning the "interpretability" of the workings of complex algorithms, where even their designers cannot explain why the AI arrived at a specific decision.

Some claim that transparency rarely comes for free and that there are often tradeoffs between how "smart" an AI is and how transparent it is; these tradeoffs are expected to grow larger as AI systems increase in internal complexity.

The technical challenge of explaining AI decisions is sometimes known as the **interpretability problem**.

C'est l'hallu… totale…

GENERATIVE AI

Fact or Fiction: What Are the Different LLM Hallucination Types?

JULY 17, 2023

Authored by **Adam Williams**
Content Writer at Holistic AI

Hallucination is Inevitable: An Innate Limitation of Large Language Models

Ziwei Xu Sanjay Jain Mohan Kankanhalli
School of Computing, National University of Singapore
ziwei.xu@u.nus.edu {sanjay,mohan}@comp.nus.edu.sg

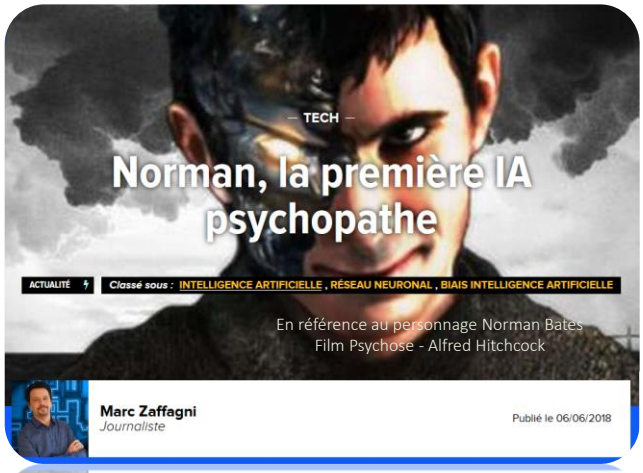
Abstract

Hallucination has been widely recognized to be a significant drawback for large language models (LLMs). There have been many works that attempt to reduce the extent of hallucination. These efforts have mostly been empirical so far, which cannot answer the fundamental question whether it can be completely eliminated. In this paper, we formalize the problem and show that it is impossible to eliminate hallucination in LLMs. Specifically, we define a formal world where hallucination is defined as inconsistencies between a computable LLM and a computable ground truth function. By employing results from learning theory, we show that LLMs cannot learn all of the computable functions and will therefore always hallucinate. Since the formal world is a part of the real world which is much more complicated, hallucinations are also inevitable for real world LLMs. Furthermore, for real world LLMs constrained by provable time complexity, we describe the hallucination-prone tasks and empirically validate our claims. Finally, using the formal world framework, we discuss the possible mechanisms and efficacies of existing hallucination mitigators as well as the practical implications on the safe deployment of LLMs.

Preprint. Under review.

https://www.holisticai.com/blog/types-of-llm-hallucinations

Toutes des malades ces I.A ?



https://www.futura-sciences.com/tech/actualites/intelligence-artificielle-norman-premiere-ia-psychopathe-71518/

Drôle d'idée au MIT !
Créer une intelligence artificielle psychopathe en l'éduquant avec des images de morts violentes avant de lui faire passer le fameux test de Rorschach.

L'objectif est de démontrer l'influence cruciale des données utilisées pour entraîner la mémoire des algorithmes.

Son réseau neuronal d'apprentissage profond a été entraîné à l'aide de photos montrant des morts violentes ou horribles puisées sur un groupe de discussion Reddit.

Ses réactions ont ensuite été observées au travers du test des taches d'encre, et comparées avec une I.A. entraînée normalement…

La censure et le problème d'alignement

Il désigne l'ensemble des questions techniques et éthiques soulevées par le rapport d'un programme d'intelligence artificielle, et de ses résultats, avec les valeurs, les attentes et les préférences humaines. https://www.sambuc-editeur.fr/articles/?a=86

Le développeur doit s'assurer que le programme conçu agit et produit des résultats conformes à ce qu'attendrait un humain sur des tâches similaires.

Dans l'apprentissage par renforcement, on fait évaluer par un humain les résultats du modèle d'intelligence artificielle, de façon à le corriger par une sorte de « système de récompense » au fil d'une succession d'essais / erreurs.

De façon plus générale, le problème d'alignement se rapporte à la question du contrôle, de la compréhension et de la prévision des comportements de l'IA impliquant aussi bien des solutions d'ingénierie que des questions éthiques.

Alignment Problem In A Nutshell

The alignment problem was popularised by author Brian Christian in his 2020 book The Alignment Problem: Machine Learning and Human Values. In the book, Christian outlines the challenges of ensuring AI models capture "our norms and values, understand what we mean or intend, and, above all, do what we want." The alignment problem describes the problems associated with building powerful artificial intelligence systems that are aligned with their operators.

Et l'éthique dans tout ça ?

Neocolonial slavery: ChatGPT built by using Kenyan workers as AI guinea pigs, Elon Musk knew

OpenAI apparently developed ChatGPT by exploiting and underpaying Kenyan workers. These workers had to sift through tons and tons of explicit and graphic content, because of which the workers developed serious mental health issues.

Mehul Reuben Das | January 26, 2023 18:48:58 IST

The Kenyan team was managed by Sama, a San Francisco-based firm, which said its workers could take advantage of both individual and group therapy sessions with "professionally-trained and licensed mental health therapists".

US Government to crack down on harmful AI products and businesses violating ethics to develop AI

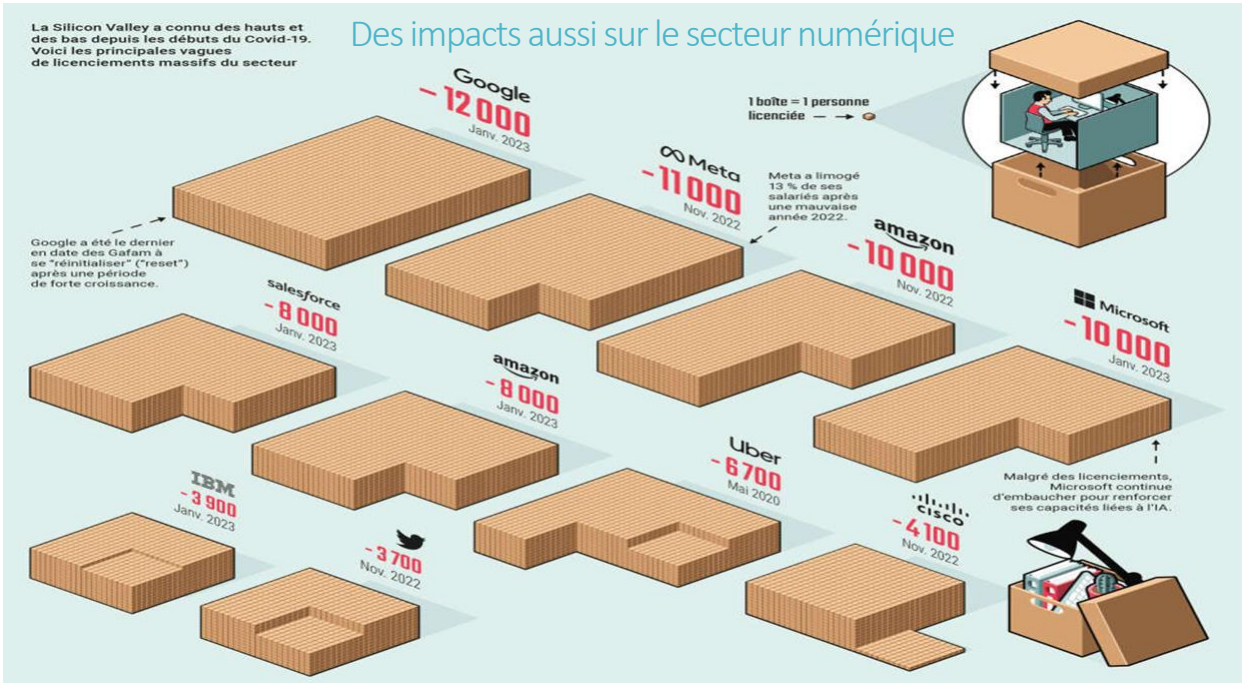
The FTC has announced that the US government will actively pursue bad actors and AI developers who use AI's biases to discriminate against or mislead people. The warning was particularly aimed at OpenAI, the makers of ChatGPT.

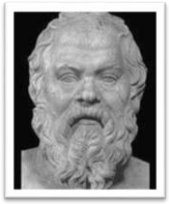
A Madagascar, les petites mains bien réelles de l'intelligence artificielle alimentent la machine

Les intelligences artificielles semblent fonctionner toutes seules mais, dans les coulisses, des humains les alimentent en données : des tâches ingrates et répétitives. Pour les entreprises françaises, Madagascar s'impose comme le lieu privilégié de cette sous-traitance du clic à bon marché.

https://www.firstpost.com/world/openai-made-chatgpt-using-underpaid-exploited-kenyan-employees-who-forced-to-see-explicit-graphic-content-12053152.html

https://www.courrierinternational.com/grand-format/infographie-dans-la-tech-on-licencie-a-tour-de-bras





Conclusion:
"All I know is that I know nothing…"

Socrates – Vth century B.C - Euh… Yes, he was a philosopher…)

The more one knows, the more he realizes the immensity of the things he does not know…

The recognition of our ignorance is the necessary attitude to adopt in the face of the quest for "knowledge" (Science…)

Computer Science is undecidable… how can we safely program *in silico* or even *in vivo*: a much more risky environment ?

Without simulation we are not even able to predict what will come from a very simple specifications (e.g. the Langton's Ant).

How is it possible to build a **reliable** "Strong" AI or a safe biological program?