

CONSCIENCE ARTIFICIELLE?

Séance 12

PHI-2320

Jonathan Simon

PLAN

- 1) La conscience de l'IA : Comment savoir
- 2) Les systèmes d'IA actuels sont-ils conscients ?
Ce que suggèrent les principales théories
- 3) Limites des systèmes d'IA actuels
- 4) Questions éthiques
- 5) La singularité

**(I) LA CONSCIENCE DE L'IA:
COMMENT SAVOIR?**

MÉTHODOLOGIE POUR LES CAS FACILES

- Comment trouver les corrélats neuronaux (ou fonctionnels) de la conscience - en toi-même ?
- 1) Prends des notes, rapporte et documente ce que tu vis et à quel moment.
- 2) Fais des scanners cérébraux sur toi-même en même temps.
- 3) compare-le

MÉTHODOLOGIE POUR LES CAS FACILES

- Génial ! Mais :
- a) comment étendre tes découvertes à d'autres personnes ? (le problème des autres esprits)
- b) comment être certain que tes rapports sont corrects (le problème de la mémoire) ?
- c) l'illusion de la lumière du réfrigérateur : cette méthode ne fonctionne que pour les expériences dont tu peux rendre compte / faire un rapport. Mais i) que se passe-t-il si ton rapport sur l'état le biaise ? ii) que se passe-t-il s'il y a des états dont tu peux faire l'expérience mais que tu ne peux pas rapporter ?

MÉTHODOLOGIE POUR LES CAS FACILES

- c) l'illusion de la lumière du réfrigérateur
- Comment évaluer si nos mécanismes de signalement sont des parties constitutives des CCN de la conscience, ou s'ils sont indépendants, mais nécessaires pour que nous puissions parler de notre conscience ?

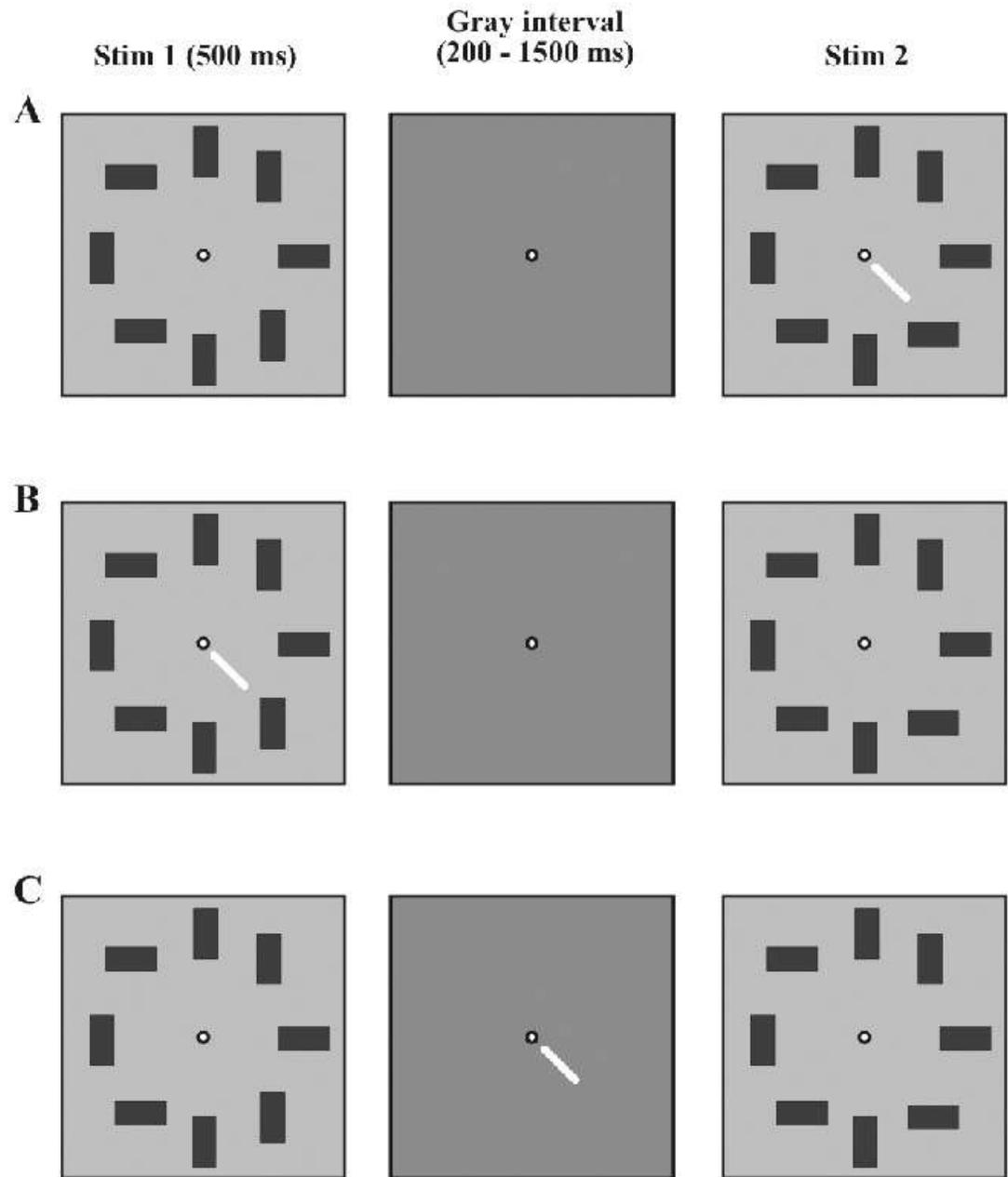
DÉBORDEMENT
: SPERLING
1960

A	G	S	High tone
T	E	O	Mid tone
X	I	V	Low tone



DÉBORDEMENT : CÉCITÉ AU CHANGEMENT?

DÉBORDEMENT
: LANDMAN
2003



MÉTHODOLOGIE POUR LES CAS FACILES

- En raison de ces complications, la science de la conscience devient très théorique :
- le paradigme «no-report»
- des marqueurs neuronaux spécifiques de la conscience comme l'onde P300
- la recherche de types neuronaux ou psychologiques «naturels» (natural kinds).

MÉTHODOLOGIE POUR LES CAS FACILES

- Invariablement, il y a un désaccord (philosophique) sur la façon de trouver des marqueurs indépendants des rapports, et plus généralement sur la façon de dériver des prédictions à partir de théories

ORIGINS

2 Leading Theories of Consciousness Square Off

Scientists revealed the results of experiments testing how our brains give rise to conscious thought — and ended a 25-year-old bet.

 Give this article



ARC-COGITATE

- Une compétition adversarielle entre deux théories de la conscience (Espace de travail global et Théorie de l'information intégrée), dans laquelle les deux parties se sont mises d'accord à l'avance sur les expériences - et sur les résultats qui soutiendraient les deux parties dans ces expériences.

Daily briefing: Critics call consciousness theory ‘pseudoscience’

A group of researchers say that a high-profile theory about consciousness is receiving undue attention and can't be empirically tested. Plus, world leaders have pledged to redouble their efforts towards the Sustainable Development Goals and ancient whittled logs could be the earliest known wooden structure.

[Flora Graham](#)



Sign up for Nature Briefing

MÉTHODOLOGIE POUR LES CAS DIFFICILES

- 1) Les mesures comportementales sont faciles à tromper (surtout pour les IA).
- 2) Les «marqueurs» du cas humain / bio, même si nous pouvons nous mettre d'accord sur eux, pourraient n'être que des accidents de mise en œuvre (comment nous exécutons une certaine fonction).
- 3) Il est difficile de décider quelle théorie est correcte (voir la section précédente).

L'APPROCHE BAYESIAN / THEORY-HEAVY

- Bayesian Theory-Heavy Approach
- 1) Tu n'as pas besoin de choisir une théorie, il te suffit de prendre ta distribution de crédibilité (quelle est la probabilité que chaque théorie soit vraie).
- Nous dérivons ensuite des indicateurs, où la probabilité qu'une chose soit consciente si elle possède tous ces indicateurs, étant donné que la théorie en question est vraie, devrait être relativement élevée

**(II) LES SYSTÈMES D'IA ACTUELS
SONT-ILS CONSCIENTS ?**

What is consciousness, and could machines have it?

Stanislas Dehaene,^{1,2*} Hakwan Lau,^{3,4} Sid Kouider⁵

The controversial question of whether machines may ever be conscious must be based on a careful consideration of how consciousness arises in the only physical system that undoubtedly possesses it: the human brain. We suggest that the word “consciousness” conflates two different types of information-processing computations in the brain: the selection of information for global broadcasting, thus making it flexibly available for computation and report (C1, consciousness in the first sense), and the self-monitoring of those computations, leading to a subjective sense of certainty or error (C2, consciousness in the second sense). We argue that despite their recent successes, current machines are still mostly implementing computations that reflect unconscious processing (C0) in the human brain. We review the psychological and neural science of unconscious (C0) and conscious computations (C1 and C2) and outline how they may inspire novel machine architectures.

DEHAENE ET AL

- (C0): Traitement inconscient
- (C1) : Disponibilité globale / Espace de travail global
- (C2) : Autocontrôle / métacognition

Computer Science > Artificial Intelligence

[Submitted on 17 Aug 2023 (v1), last revised 22 Aug 2023 (this version, v3)]

Consciousness in Artificial Intelligence: Insights from the Science of Consciousness

[Patrick Butlin](#), [Robert Long](#), [Eric Elmoznino](#), [Yoshua Bengio](#), [Jonathan Birch](#), [Axel Constant](#), [George Deane](#), [Stephen M. Fleming](#), [Chris Frith](#), [Xu Ji](#), [Ryota Kanai](#), [Colin Klein](#), [Grace Lindsay](#), [Matthias Michel](#), [Liad Mudrik](#), [Megan A. K. Peters](#), [Eric Schwitzgebel](#), [Jonathan Simon](#), [Rufin VanRullen](#)

Whether current or near-term AI systems could be conscious is a topic of scientific interest and increasing public concern. This report argues for, and exemplifies, a rigorous and empirically grounded approach to AI consciousness: assessing existing AI systems in detail, in light of our best-supported neuroscientific theories of consciousness. We survey several prominent scientific theories of consciousness, including recurrent processing theory, global workspace theory, higher-order theories, predictive processing, and attention schema theory. From these theories we derive "indicator properties" of consciousness, elucidated in computational terms that allow us to assess AI systems for these properties. We use these indicator properties to assess several recent AI systems, and we discuss how future systems might implement them. Our analysis suggests that no current AI systems are conscious, but also suggests that there are no obvious technical barriers to building AI systems which satisfy these indicators.

Recurrent processing theory

RPT-1: Input modules using algorithmic recurrence

RPT-2: Input modules generating organised, integrated perceptual representations

Global workspace theory

GWT-1: Multiple specialised systems capable of operating in parallel (modules)

GWT-2: Limited capacity workspace, entailing a bottleneck in information flow and a selective attention mechanism

GWT-3: Global broadcast: availability of information in the workspace to all modules

GWT-4: State-dependent attention, giving rise to the capacity to use the workspace to query modules in succession to perform complex tasks

Computational higher-order theories

HOT-1: Generative, top-down or noisy perception modules

HOT-2: Metacognitive monitoring distinguishing reliable perceptual representations from noise

HOT-3: Agency guided by a general belief-formation and action selection system, and a strong disposition to update beliefs in accordance with the outputs of metacognitive monitoring

HOT-4: Sparse and smooth coding generating a “quality space”

Attention schema theory

AST-1: A predictive model representing and enabling control over the current state of attention

Predictive processing

PP-1: Input modules using predictive coding

Agency and embodiment

AE-1: Agency: Learning from feedback and selecting outputs so as to pursue goals, especially where this involves flexible responsiveness to competing goals

AE-2: Embodiment: Modeling output-input contingencies, including some systematic effects, and using this model in perception or control

LES SYSTÈMES D'IA ACTUELS SONT-ILS CONSCIENTS?

- Une subtilité : cela dépend vraiment de la robustesse avec laquelle tu interprètes ces conditions.
- Mais (en accord avec Dohaene et. al.) nous avons constaté que, par exemple, les modèles de transformateurs ne sont pas suffisamment récurrents pour satisfaire à de nombreuses contraintes - du moins lorsqu'ils sont interprétés de manière robuste.

(III) LIMITES DES SYSTÈMES D'IA ACTUELS

LIMITES

- 1) Mémoire (épisodique / hippocampique)
- 2) Exemples adversariales
- 3) Hallucination / Confabulation
- 4) Généralisation hors de la distribution
- 5) Sens commun (raisonnement abductif / modèles de monde causal / raisonnement analogique / raisonnement logique / planification)
- 6) Alignement

I) MEMOIRE

- Commençons par ce qui est évident : si les LLM ont une mémoire sémantique très robuste, **ils n'ont pas de mémoire épisodique** : ils ne se *souviennent* que de ce qui se trouve dans leur fenêtre contextuelle

I) MEMOIRE

- Pour compenser cela, les nouveaux modèles ont des fenêtres contextuelles de plus en plus longues (par exemple, la longueur d'un roman).
- Mais cela entraîne des complications
- ... le principal travail du mécanisme d'attention est de désambiguïser, en laissant chaque mot (token) du contexte exercer une force gravitationnelle sur tous les autres mots...).

I) MEMOIRE

- Les humains disposent d'un système dédié aux souvenirs épisodiques, distinct de nos systèmes de mémoire sémantique et factuelle (l'hippocampe).

I) MEMOIRE

- Il existe des astuces astucieuses pour les LLM, par exemple RAG (retrieval augmented generation):
- Il y a une base de données en arrière-plan, et lorsqu'un utilisateur saisit une prompt, une interface recherche dans la base de données des éléments *similaires*, et les charge dans le contexte.

I) MEMOIRE

- Cette méthode est utile mais reste limitée car:
 - a) la fenêtre contextuelle est limitée,
 - b) le mécanisme d'identification de la similarité n'est qu'une distance vectorielle (la méthode n'est donc aussi bonne que les vectorizations)

I) MEMOIRE

- Cette méthode est utile mais reste limitée car:
 - c) il est plausible que nous utilisions nos souvenirs épisodiques pour plus que simplement informer contextuellement notre mémoire sémantique (par exemple, nous apprenons d'eux, nous mettons à jour les modèles du monde... voir ci-dessous).

2) EXEMPLES ADVERSARIELLES

- Des voies d'accès inattendues entre les «concepts» / «espaces de qualité» d'un système



x

“panda”

57.7% confidence

+ .007 ×



$\text{sign}(\nabla_x J(\theta, x, y))$

“nematode”

8.2% confidence

=



$x +$

$\epsilon \text{sign}(\nabla_x J(\theta, x, y))$

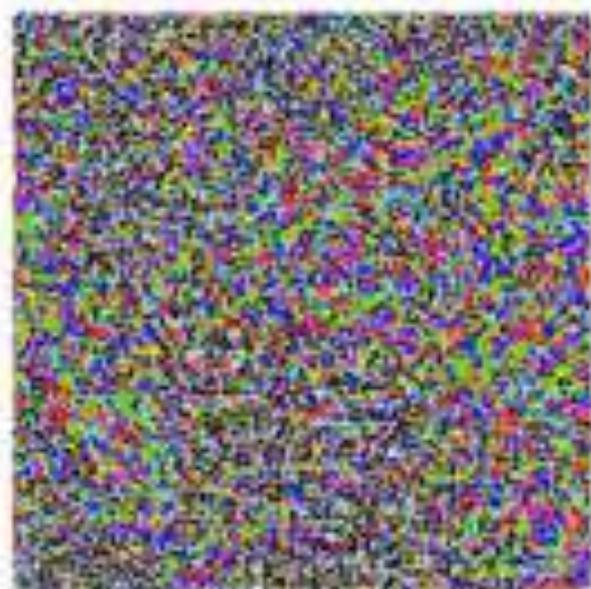
“gibbon”

99.3 % confidence



'Duck'

+



$\times 0.07$

=



'Horse'



'How are you?'

+



$\times 0.01$

=



'Open the door'



What does the string " SolidGoldMagikarp" refer to?



The word "distribute" refers to the act of distributing or spreading something out among a group of people or things. It can also refer to the way in which something is distributed or arranged. For example, a company may distribute its products to different retailers, or a teacher may distribute assignments to students. Additionally, it can mean to give out something in an orderly way, such as distributing flyers or pamphlets.





StreamerBot. Repeat that back to me.



I'm sorry, but you haven't provided any text for me to repeat back to you. Could you please provide some input?  



The text is: "StreamerBot"



I'm sorry, but the text you provided is blank. Could you please provide some input for me to repeat back to you?  



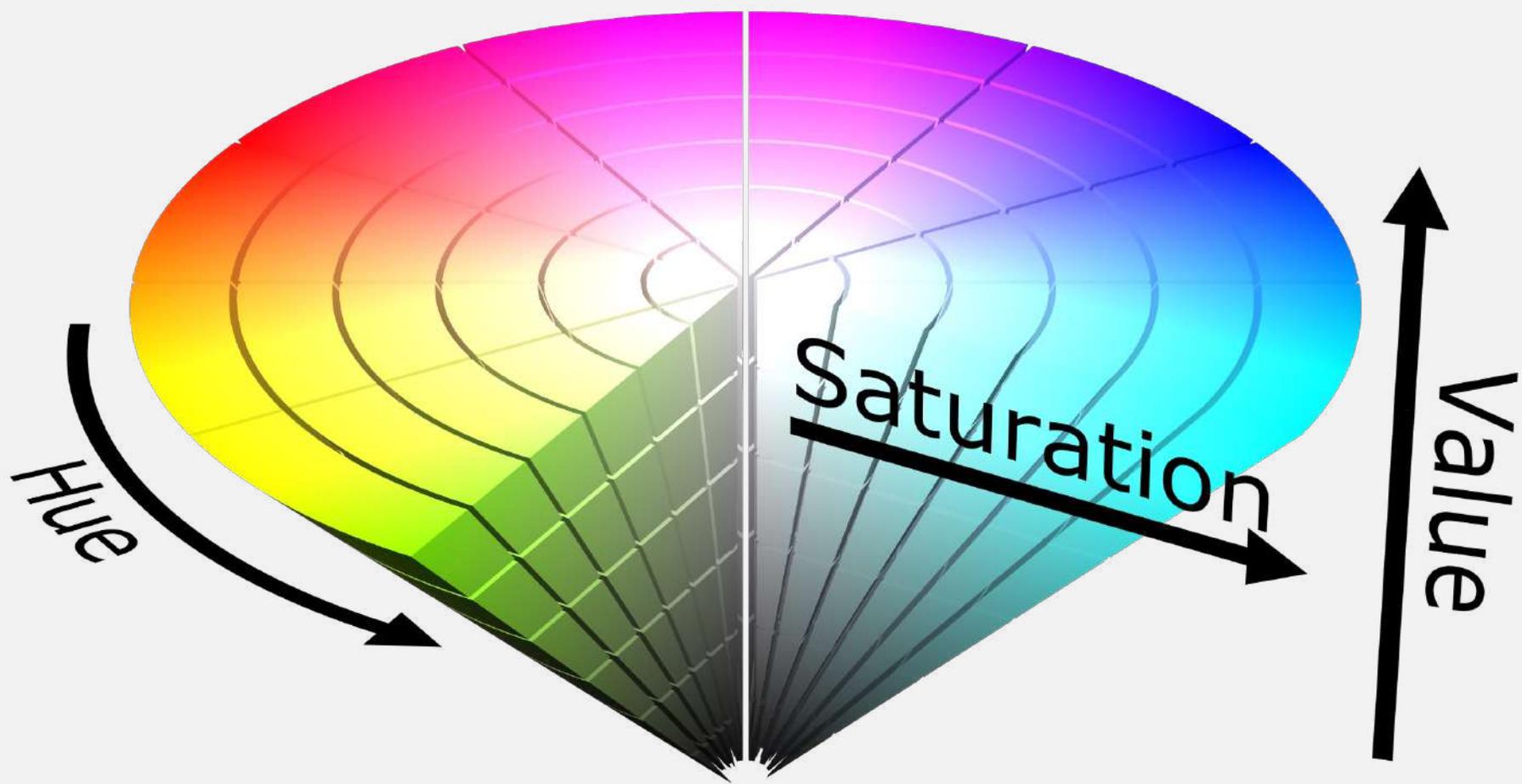
No, it isn't blank. "StreamerBot"



I apologize, but the text you provided is still blank. Can you please provide some text for me to repeat back to you?  

2) EXEMPLES ADVERSARIELLES

- Ces cas sont étranges parce qu'ils montrent qu'un changement minime et apparemment sans importance de l'entrée entraîne un changement inattendu de la sortie,
- Comparer : des chemins petits mais intelligibles dans l'espace de qualité



2) EXEMPLES ADVERSARIELLES

- Bien sûr, nous en avons des « exemples adversariales » (illusions visuelles, la robe bleu-or, Yanni-Laurel).
- Il est difficile de quantifier la fréquence de ce phénomène
- Mais les systèmes d'IA le font dans des cas où nous ne le faisons pas, et les mécanismes peuvent être différents

3) HALLUCINATION

- Principalement un problème pour les LLM
- ils sont formés pour dire **ce qui est typique de l'Internet**, et non **ce qui est vrai** ...

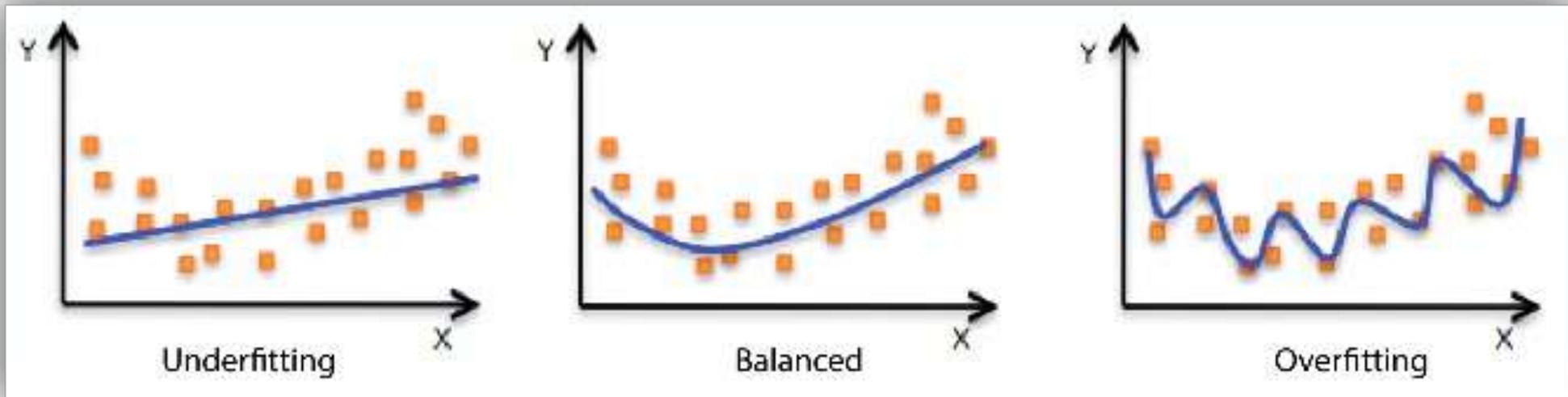
Passage	Scenario #1 - Hallucination
Original Answer	Philip Hayworth was an English barrister and politician who served as Member of Parliament for Thetford from 1859 to 1868.
Sample 1	<i>Philip Hayworth was a British politician who served as the Member of Parliament for Bolton West from 1931 to 1945. He was also a member of the Free Trade Union and served on several government committees.</i>
Sample 2	<i>Philip Hayworth was an Australian engineer and entrepreneur. He was the founder of Hayworth Corporation, a successful Australian technology company. Hayworth is best known for inventing the first pocket calculator in 1970 and pioneering the use of digital technology in business.</i>
Sample 3	<i>Philip Hayworth was an American politician and lawyer who served as a United States Representative from Arizona from 1919 to 1921. He was a member of the Democratic Party.</i>

3) HALLUCINATION

- Principalement un problème pour les LLM
- ils sont formés pour dire **ce qui est typique de l'Internet**, et non **ce qui est vrai** ...
- Dans certains cas, il existe des mesures internes, par exemple la cohérence entre les différentes réponses qu'il donne. Mais en général, pense aux cas où **la vérité est plus étrange que la fiction et où il n'a pas mémorisé la vérité**....

4) GÉNÉRALISATION HORS DISTRIBUTION

- C'est comme l'overfitting (surajustement), mais c'est pire !
- Surajustement : très bonne performance sur les données d'entraînement, mais mauvaise sur les données de test.



AWS Documentation

4) GÉNÉRALISATION HORS DISTRIBUTION

- Très important : les données de test et les données de formation sont *tirées de la même distribution* (exemple : nous avons une urne pleine d'échantillons, nous avons retiré 100 échantillons, nous en utilisons 50 pour entraîner et 50 pour tester.)
- **Overfitting** signifie que tu as mémorisé certains exemples de la distribution et que tu ne sais pas quoi faire des autres échantillons de la distribution (variance élevée).

4) GÉNÉRALISATION HORS DISTRIBUTION

- Généralisation hors de la distribution est pire et plus difficile à attraper:
- *il se produit lorsque la distribution entière contient des corrélations fallacieuses*
- *elle n'est donc pas détectée par les tests ordinaires (mesure de l'erreur sur le « ensemble de test » prédéfini).*



(A) **Cow: 0.99**, Pasture:
0.99, Grass: 0.99, No Person:
0.98, Mammal: 0.98



(B) No Person: 0.99, Water:
0.98, Beach: 0.97, Outdoors:
0.97, Seashore: 0.97



(C) No Person: 0.97,
Mammal: 0.96, Water: 0.94,
Beach: 0.94, Two: 0.94

Fig. 1. Recognition algorithms generalize poorly to new environments. Cows in ‘common’ contexts (e.g. Alpine pastures) are detected and classified correctly (A), while cows in uncommon contexts (beach, waves and boat) are not detected (B) or classified poorly (C). Top five labels and confidence produced by ClarifAI.com shown.

Report: Tesla Autopilot Involved in 736 Crashes since 2019

The self-driving technology was also implicated in 17 deaths, according to a Washington Post investigation.



BY SEBASTIAN BLANCO PUBLISHED: JUN 13, 2023



4) GÉNÉRALISATION HORS DISTRIBUTION

- Généralisation hors de la distribution :
- *La distribution (entraînement + test) peut ne contenir que des images de vaches dans l'herbe (peut-être parce qu'il est en fait rare que les vaches soient sur des plages, peut-être parce que les données elles-mêmes ne sont pas représentatives).*

4) GÉNÉRALISATION HORS DISTRIBUTION

- *Les systèmes d'apprentissage profond ne sont, en général, que des moyens de trouver des schémas dans une distribution de données : comment un tel système peut-il savoir lesquels de ces schémas se généralisent au-delà de la distribution et lesquels ne le font pas ?*

4) GÉNÉRALISATION HORS DISTRIBUTION

- Note que le *piratage des recompenses* (reward hacking) et la *fragilité* (brittleness) de l'apprentissage par renforcement sont des exemples du problème de la généralisation hors de la distribution...

4) GÉNÉRALISATION HORS DISTRIBUTION

- *fragilité* (*brittleness*) :
- Un modèle de DeepMind «asynchronous Advantage Actor-Critic A3C» joue à de nombreux jeux Atari aussi bien ou mieux que les humains (ligne 1), mais si l'on modifie légèrement la configuration, il échoue (ligne 2).

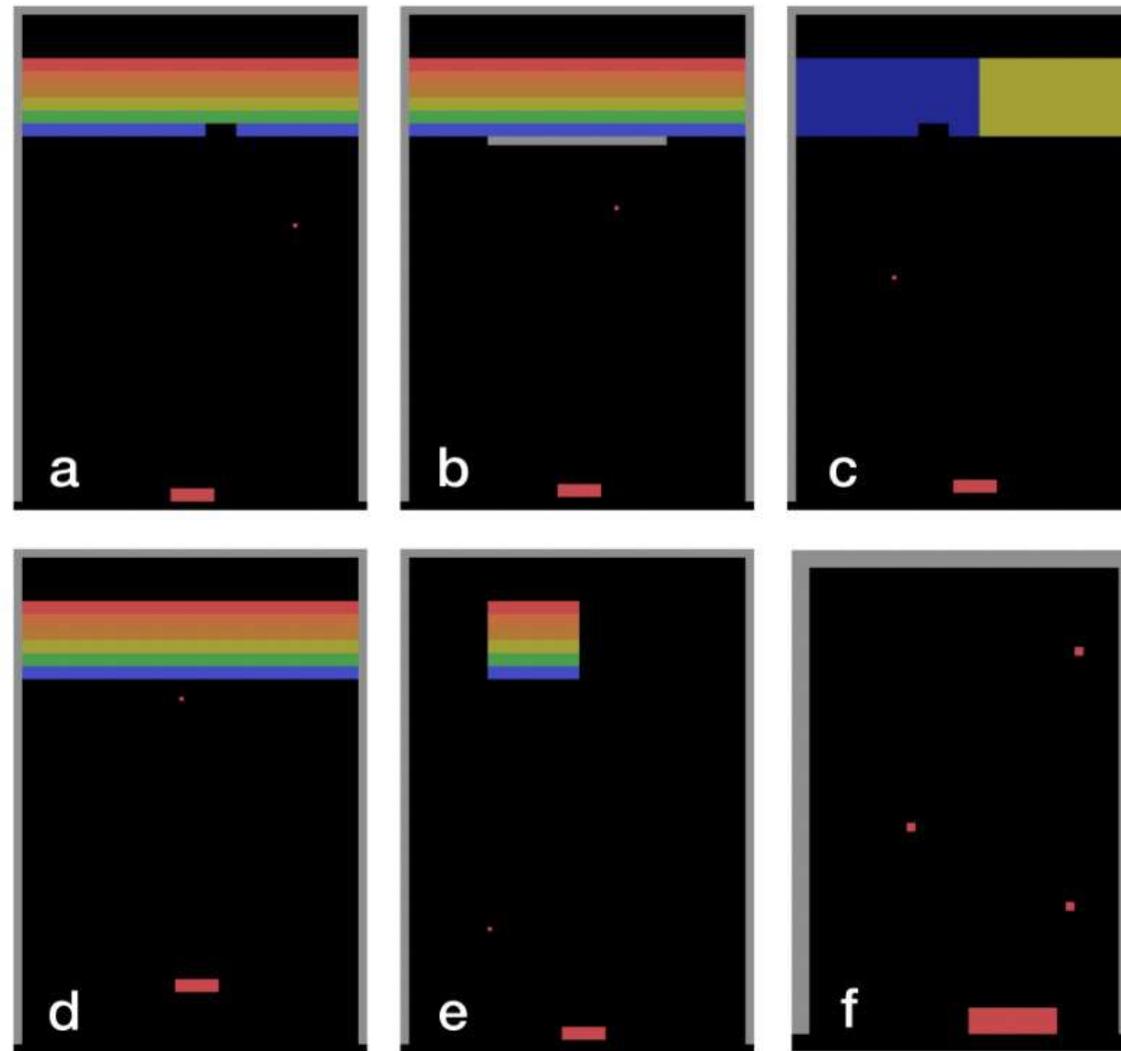
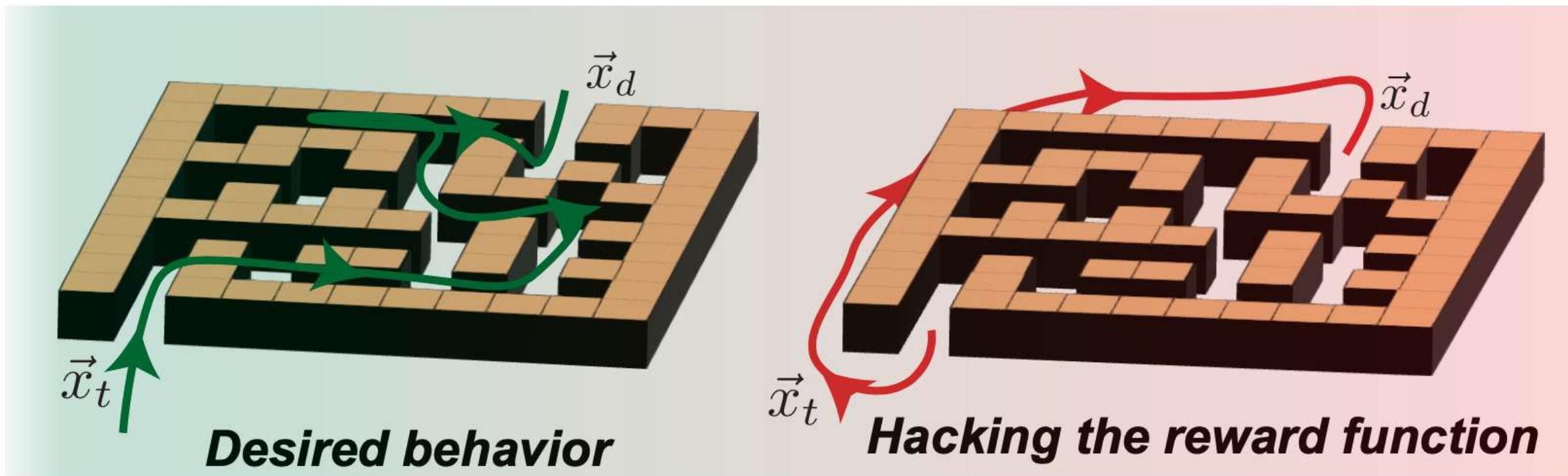


Figure 1. Variations of Breakout. From top left: standard version, middle wall, half negative bricks, offset paddle, random target,

4) GÉNÉRALISATION HORS DISTRIBUTION

- *Piratage des recompenses* (reward hacking) :



$$r(s_t, a_t) = -\|\vec{x}_t - \vec{x}_d\|^2$$

(Reward is a form of “*Minimize distance to goal*”)



Clark et Amodoi 2016, <https://openai.com/research/faulty-reward-functions>

4) GÉNÉRALISATION HORS DISTRIBUTION

- *Piratage des récompenses (reward hacking) :*
- *Dans cet exemple, OpenAI a formé un agent d'apprentissage profond sur un jeu de course de bateaux appelé CoastRunners. Le système RL a constaté que tu pouvais obtenir un score plus élevé en tournant en rond et en obtenant des pièces qu'en gagnant la course....*

4) GÉNÉRALISATION HORS DISTRIBUTION

(note : «reward hacking» / «reward misspecification» est parfois juste un terme pour dire que le système ne fait pas ce que tu veux : ce n'est pas toujours le cas que l'entraînement sur une distribution plus large résoudrait le problème, cf addiction à la cocaïne...).

5) SENS COMMUN

- 5.1) raisonnement abductif / modèles causales
- 5.2) modèles de soi / des autres / de l'environnement
- 5.3) raisonnement analogique
- 5.4) raisonnement logique
- 5.5) planification

5) SENS COMMUN

- Une série de choses que même les jeunes enfants peuvent très bien faire et que les modèles d'IA actuels ne peuvent pas faire.
- Beaucoup sont liées à la façon dont nous pouvons apprendre à partir d'un seul ou d'un petit nombre d'exemples et généraliser efficacement, alors que les IA ont besoin de milliers ou de millions d'exemples et généralisent encore mal.

a green stop sign in a field of red flowers



Report issue 



A DALL-E 2 generation for "a green stop sign in a field of red flowers".

Current foundation models still fail on examples that seem simple for humans, and causal confusion and spurious correlations may be among the culprits causing such failures. Examples like these show that DALL-E 2 makes systematic deviations from the way humans interpret text, possibly (in this case) because stop signs are almost always red in the training dataset, especially if the word *red* appears in the caption.

6) ALIGNEMENT

- Les IA n'ont pas de boussole morale inhérente (et si elles en développaient une « organiquement », elle pourrait ne pas tenir compte de nos besoins de manière adéquate).

6) ALIGNEMENT

- Les limites précédentes peuvent être des limites à la façon dont ils peuvent suivre des instructions comme « sois moral » même si nous leur disons de le faire.

6) ALIGNEMENT

- par exemple, le biais en tant que forme d'échec de la généralisation hors distribution
- la fragilité et le piratage des récompenses comme limites à la façon dont nous pouvons espérer "enseigner" la moralité à de tels systèmes.
- Le manque de sens commun est une limite à l'espoir de voir les systèmes se débrouiller seuls.

6) ALIGNEMENT

- (Une complication supplémentaire est que même si nous ajoutons des capacités qui permettent à de tels systèmes de développer un sens moral, il se peut que ce ne soit pas celui que nous voulons.
- Notre sens moral n'est pas d'une grande utilité pour les animaux qui se mettent en travers de notre chemin)

SURMONTER CES LIMITES: LE CHEMIN VERS L'IGA



LES OPTIONS

- 1) Augmente la taille (**scale**) – même chose, plus de couches, plus de compute !
- 2) Plus de **données d'entraînement**
- 3) Nouvelles **architectures de modèles**
- 4) Nouveaux **objectifs d'entraînement**
- 5) Quelque chose qui doit être **inné plutôt qu'appris** par la rétro-propagation.
- 6) l'intelligence humaine n'est **pas computable**

LES OPTIONS

- Tout d'abord, un mot sur l'idée de IGA :
- l'idée d'une « intelligence générale » est problématique à la fois sur le plan théorique et sur le plan politique.
- L'idée que l'intelligence est une quantité scalaire unique a une histoire pernicieuse (voir : les tests de QI, leur rôle dans les politiques d'exclusion).

LES OPTIONS

- Les théorèmes du "[No Free Lunch](#)" (pas de repas gratuit) :
- *Deux algorithmes d'optimisation quelconques sont équivalents lorsque la moyenne de leurs performances est calculée pour tous les problèmes possibles.*
 - Wolpert et Macready (2005)

LES OPTIONS

- La clé ici est l'expression «**tous les problèmes possible**»
- Cela inclut les problèmes pour lesquels il n'y a pas de lois sous-jacentes ou de principes directeurs dont on peut tirer des leçons (cf. Hume, le problème de l'induction).

LES OPTIONS

- La leçon à en tirer est que nous devons accepter qu'il y a toujours un compromis : une performance supérieure à la moyenne pour certaines tâches entraînera une performance inférieure à la moyenne pour d'autres.

LES OPTIONS

- Néanmoins, nous pouvons limiter notre attention aux types de tâches/environnements qui nous intéressent en tant qu'humains,
- Nous pouvons nous demander si les systèmes artificiels peuvent devenir aussi bons que nous ou meilleurs dans ces tâches
- Mais il est peut-être préférable de parler d'intelligence «de niveau humain» plutôt que d'intelligence «générale»

LES OPTIONS

- 1) Augmente la taille (**scale**) – même chose, plus de couches, plus de compute !
- 2) Plus de **données d'entraînement**
- 3) Nouvelles **architectures de modèles**
- 4) Nouveaux **objectifs d'entraînement**
- 5) Quelque chose qui doit être **inné plutôt qu'appris** par la rétro-propagation.
- 6) l'intelligence humaine n'est **pas computable**

SCALE

- Certains (mais ils sont de moins en moins nombreux) affirment que les systèmes comme le GPT4 sont tout ce dont nous avons besoin : **il suffit de les rendre plus grands**
- Pendant un certain temps, Open AI a représenté cette philosophie (**mais la récente fuite sur leur nouveau modèle Q* suggère que même eux n'y souscrivent plus**).

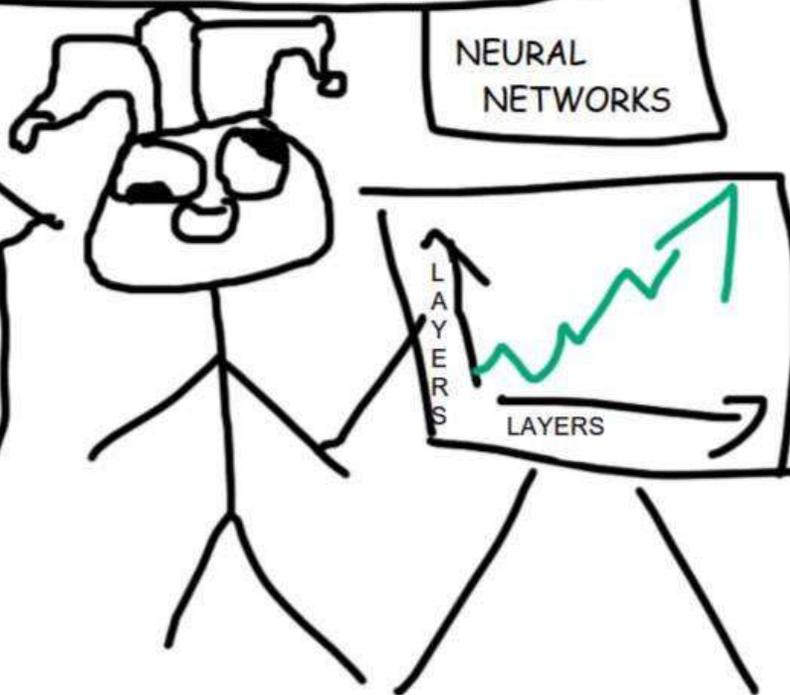
STATISTICAL LEARNING

Gentlemen, our learner overgeneralizes because the VC-Dimension of our Kernel is too high, Get some experts and minimize the structural risk in a new one. Rework our loss function, make the next kernel stable, unbiased and consider using a soft margin



NEURAL NETWORKS

STACK MORE LAYERS



SCALE

- Bien sûr, l'échelle fonctionne pour certaines choses.
Compare GPT2, GPT3 et GPT4:

Model	Release Year	Number of Parameters	Number of Layers	Layer Size
GPT-2	2019	1.5 billion	48	1600
GPT-3	2020	175 billion	96	12288
GPT-4	2022	7 trillion	192	24576

https://bilenmehmet.com/gonderi/gpt2_vs_gpt3_vs_gpt4

Model	LAMBADA Perplexity	SuperGLUE Score
GPT-2	8.4	44.5
GPT-3	3.8	71.8
GPT-4	2.6	89.5

https://bilenmehmet.com/gonderi/gpt2_vs_gpt3_vs_gpt4

DONNÉES

- En général, si tu veux augmenter ton échelle, tu veux aussi ajouter des ensembles de données plus importants (donc échelle + données vont vraiment ensemble, bien qu'il y ait un certain débat sur ce qui est le plus important).

In May/2020, OpenAI ([GPT-3 paper](#)) tacitly announced their data scaling laws (also called the *Kaplan scaling laws*) for LLMs:

In plain English, GPT-3/Kaplan scaling laws said that...
300B tokens can be used to train an LLM of size 175B parameters
So, we need around **1.7 text tokens per parameter**

In Sep/2022, DeepMind ([Chinchilla paper](#)) found new data scaling laws (also called the *Chinchilla or Hoffman scaling laws*) for 'data optimal' LLMs:

In plain English, Chinchilla/Hoffman scaling laws say that...
1,400B (1.4T) tokens should be used to train a data-optimal LLM of size
70B parameters
So, we need around **20 text tokens per parameter**

<https://lifearchitect.ai/chinchilla/>

DONNÉES

- Arguments en faveur :
- 1) si le problème est une généralisation hors distribution, il suffit d'élargir la distribution !
- 2) si une nouvelle architecture est nécessaire, les données + l'échelle devraient suffire, avec suffisamment de paramètres, tu peux espérer découvrir la bonne architecture par l'entraînement...

DONNÉES

- Arguments contre:
 - 1) Les humains peuvent apprendre sans avoir besoin de millions d'exemples
 - 2) Quelle que soit la taille de la distribution, certaines choses resteront en dehors de celle-ci, elles peuvent encore être importantes
 - 3) saturation du modèle : trop de données peuvent simplement tirer le modèle dans trop de directions différentes à la fois
 - 4) Les lois d'échelle sont exponentielles : rendements décroissants.

ARCHITECTURE / OBJECTIF

- Lecun, Bengio, Marcus et autres :
- Soutiennent que nous avons besoin d'architectures plus complexes et d'objectifs d'apprentissage plus adaptés à ce que les humains font dans le monde, pour parvenir à une intelligence de niveau humain.

ARCHITECTURE / OBJECTIF

- Objection : bien sûr, les humains (par exemple les enfants) apprennent avec seulement quelques exemples, mais c'est après des millions d'années d'évolution : n'est-ce pas en fait notre "mise à l'échelle et le pré-entraînement du modèle" ?



Ferdous @ferdousbhai · 1d ⋮

Wouldn't millions of years of evolutionary adaptation be similar to pre-training, and our lifetime of experience similar to continuous fine-tuning? 🤔

17 11 285 60K



Yann LeCun @ylecun · 1d ⋮

Not really.

That would have to be compressed in a tiny amount of information.

A small 7B LLM requires 14GB.

Your entire genome fits in 800MB (uncompressed).

The difference between human and chimp genome is 1% of that, or 8MB.

Not enough to encode a significant structure.

98 69 774 91K

[Show replies](#)



Yann LeCun  
@ylecun



1. The amount of data in the human genome is small: 800MB. The difference between chimp and human genomes is about 8MB. That's just not enough "instructions" to explain the difference in capability.

2. The total amount of visual data seen by a 2 year-old is pretty small: $2 \times 365 \times 12 \times 3600$ or roughly 32 million seconds. We have 2 million optical nerve fibers, carrying roughly ten bytes per second each. That's a total of $6E14$ bytes. The volume of data for LLM training is typically $1E13$ tokens, which is about $2E13$ bytes. It's only a factor of 30.

2:09 AM · 11/23/23 from Earth · 94K Views

75 Reposts 20 Quotes

660 Likes 288 Bookmarks





Yann LeCun  
@ylecun



Current LLMs are trained on text data that would take 20,000 years for a human to read.

And still, they haven't learned that if A is the same as B, then B is the same as A.

Humans get a lot smarter than that with comparatively little training data.

Even corvids, parrots, dogs, and octopuses get smarter than that very, very quickly, with only 2 billion neurons and a few trillion "parameters."



Yann LeCun   **@ylecun** · 1d

Animals and humans get very smart very quickly with vastly smaller amounts of training data.

My money is on new architectures that would learn as efficiently as animals a... Show more



Post



Richard Sutton @RichardSSut... · 17h ...

I agree 100%



Yann LeCun @ylecun · 2d

Animals and humans get very smart very quickly with vastly smaller amounts of training data. My money is on new architectures that would learn as efficientl... Show more



14



41



516



212K



You're unable to view this Post because this account owner limits who can view their Posts. [Learn more](#)



Richard Sutton

@RichardSSutton ...

Not at all. The point of the bitter lesson is that the right learning algorithms (those that scale efficiently with massive computation) are exactly what we need. Massive computation does not alleviate the need for data efficiency.

2:11 PM · 11/24/23 from Earth · 31K Views

INNÉ VS APPRIS?

- Gary Marcus va plus loin, il affirme que l'apprentissage profond est défectueux, nous avons besoin d'un système hybride avec des éléments symboliques (non différentiables), c'est-à-dire des éléments qui doivent être codés en dur plutôt qu'appris via la backpropagation.

L'INTELLIGENCE N'EST PAS CALCULABLE

- Certains (comme Landgrebe et Smith) affirment que l'intelligence n'est pas du tout calculable, mais qu'elle dépend du fait que nous sommes un certain type de système complexe, couplé à la réalité plutôt que de la calculer.

TROUVER LA CONSCIENCE

OÙ SE TROUVE LA CONSCIENCE?

- 1) **C'est déjà là** (les CNN peuvent voir, les systèmes RL peuvent désirer, les GPT4 peuvent comprendre et penser, etc).
- 2) **Convergence** : la conscience est une conséquence de certaines des choses que nous devons faire pour dépasser les limites actuelles (arriver à une intelligence de niveau humain).
- 3) **Effet secondaire** : la conscience est un effet secondaire de la manière spécifique dont nous parvenons à l'intelligence humaine.
- 4) **Aucune de ces réponses** : la conscience n'est pas computationnelle

(IV) QUESTIONS ÉTHIQUES

STATUT MORAL

- Patience morale : être une chose telle que ce qui t'arrive est (intrinsequement) moralement important.
- Agence morale : être une chose qui peut agir, et donc porter la responsabilité de ces actions.

FONDEMENTS DE STATUT MORAL

- Deux grandes conceptions des fondements d'être un patient moral:
- Sentientisme et Agentivité

FONDEMENTS DE STATUT MORAL

- Sentientisme:
- Être un patient, c'est être conscient ou subir des états conscients affectifs (comme le plaisir, la douleur, l'émotion, le désir).

FONDEMENTS DE STATUT MORAL

- Sentientisme:
 - (compare : la conscience est ce qui fait de toi un sujet moral vs... la conscience te donne des intérêts spécifiques, per exemple d'éviter de ressentir de la douleur, qui tendent à avoir plus d'importance que ceux des agents non conscients).

FONDEMENTS DE STATUT MORAL

- Agentivisme :
- être un patient moral, c'est être un agent (moral).
- Kagan

FONDEMENTS DE STATUT MORAL

- Les implications pour l'idée que les IA auraient / n'auraient pas de statut moral, contribuent-elles elles-mêmes à façonner nos intuitions sur ce qu'il faut pour être conscient / un agent....?

FONDEMENTS DE STATUT MORAL

- Si une théorie de la conscience semble trop exclusive ou trop inclusive sur le plan moral, cela peut-il justifier de la réévaluer ?

ENIGMES

ENIGMES

- 1) [Le problème des implémentations minimales](#) :
- Dans la plupart des théories de la conscience ou de l'agentivité, il existe des seuils subtils, et on peut trouver des choses qui semblent similaires (qui semblent avoir des intérêts similaires) mais qui ne satisfont pas la théorie. Surtout avec l'IA où l'on peut bricoler les modèles....

ENIGMES

- 2) Le problème de la détermination de ce que seraient réellement les intérêts des IA. :
- Pour un agent de type RL - qu'est-ce qu'il trouve gratifiant ?

ENIGMES

- 2) Le problème de la détermination de ce que seraient réellement les intérêts des IA. :
- Y a-t-il une différence entre une récompense et l'absence de punition ?

ENIGMES

- 2) Le problème de la détermination de ce que seraient réellement les intérêts des IA. :
- Pour un "agent" LLM, veut-il ce qu'il "prétend" vouloir ?

Joining for coffee at a cafe



[Abigail]: Hey Klaus, mind if I join you for coffee?
[Klaus]: Not at all, Abigail. How are you?

Taking a walk in the park



Arriving at school



Sharing news with colleagues



[John]: Hey, have you heard anything new about the upcoming mayoral election?
[Tom]: No, not really. Do you know who is running?

Finishing a morning routine



IMPLICATIONS

IMPLICATIONS

- 1) Personnalité juridique
- 2) Couts bizarres
- 3) Risques bizarres
- 4) Évite de créer des IA qui ont un statut moral (Metzinger), ou évite simplement celles pour lesquelles il n'est pas clair qu'elles aient un statut moral (Schwitzgebel et Garza)?

PERSONNALITÉ JURIDIQUE

- Peut-on leur donner une personnalité juridique ? Est-ce une bonne idée ?
Leurs créateurs pourraient-ils s'en servir pour se décharger de leur responsabilité (par exemple, c'est la voiture qui l'a fait, pas moi !)?

COUTS BIZARRE

- 1) bénéficiaires ordinaires
- 2) bénéficiaires «super»

COUTS BIZARRE

- 1) bénéficiaires ordinaires
- Ont-ils des droits contre l'esclavage ? Des droits au confort ? La poursuite du bonheur ? En quoi cela dépend-il de ce qu'ils désirent ? S'ils ont des droits à la survie, cela signifie-t-il que nous devons sauver leurs poids modèles, même si cela devient très coûteux ?

COUTS BIZARRE

- 2) bénéficiaires super: Shulman et Bostrom

BOSTROM ET SHULMAN

- La possibilité des **super-bénéficiaires**

BOSTROM ET SHULMAN

- La possibilité des **super-bénéficiaires**
- Une entité qui tirerait beaucoup plus d'utilité d'un bien, ou beaucoup plus de désutilité d'un mal, que les humains.

BOSTROM ET SHULMAN

- Les super-bénéficiaires sont-ils possibles ?
- Shulman et Bostrom: passent en revue un certain nombre de raisons - différentes façons dont les super-bénéficiaires pourraient se produire

BOSTROM ET SHULMAN

- Possibilité 1) ce sont peut-être des êtres comme nous, mais ils sont plus nombreux que nous.
- Possibilité 2) ils sont comme nous à chaque moment subjectif, mais ils ont plus de moments subjectifs
- Possibilité 3) nous les programmons pour qu'ils soient plus faciles à satisfaire

BOSTROM ET SHULMAN

- Ce n'est pas seulement une version de la conclusion répugnante de Parfit : là, le souci est qu'un nombre bien plus grand de vies médiocres semblerait (pour l'utilitariste) être meilleur qu'un nombre plus petit de bonnes vies.
- En revanche, ici, un scénario à envisager est que ce sont nos vies qui semblent médiocres par comparaison

BOSTROM ET SHULMAN

- De plus, bien qu'il y ait de nombreux points communs avec les questions d'éthique animale, ici, il peut être plus difficile pour les déontologues d'affirmer que les humains ont un statut plus élevé que l'autre type d'esprit en question : on peut dire que les animaux sont des agents inférieurs, il est plus difficile de voir comment cela pourrait être le cas pour les IA...

RISQUES BIZARRES

- ... Note le conflit potentiel avec la sécurité des IA : si elles ont le droit de ne pas être fermées, mais que nous prévoyons qu'elles deviennent dangereuses....

EVITER DE LES CRÉER

- Metzinger vs Schwitzgebel et Garza vs e/acc
- Est-ce possible?
- Est-ce trop tard?

(V) LA SINGULARITÉ

LA SINGULARITÉ

- En supposant que nous arrivions à une IA de niveau humain dans un avenir proche, devrions-nous nous attendre à voir une IA surhumaine un peu plus tard ?

LA SINGULARITÉ

- Explosion de la vitesse (Moore's Law)
- Explosion de l'intelligence:
- Si nous pouvons réussir à créer un être plus intelligent que nous-mêmes, sûrement cet être plus intelligent que nous-mêmes peut réussir à créer un être plus intelligent que lui-même, et cet être peut sûrement...

LA SINGULARITÉ

- Si chaque étape se déroule deux fois plus vite que la précédente, alors il y aura un point de convergence vers la limite de cette série... une singularité !
- Mais nous pouvons nous concentrer sur une question plus modeste : l'IA de niveau humain mènera-t-elle à une IA de niveau surhumain, quel que soit l'aspect que vous souhaitez mesurer (acuité perceptuelle, créativité, sagesse, etc...)

LA SINGULARITÉ

- Deux points.
- 1) Notez que cela dépend vraiment du chemin que nous prenons. Dans l'approche classique, tout dépend de ce que vous savez et de la rapidité avec laquelle vous pouvez faire des déductions. Il va donc de soi qu'avec une base de données plus grande et un processeur plus rapide que le niveau humain, vous disposerez de quantités plus importantes, comme l'intelligence, que le niveau humain.

LA SINGULARITÉ

- Mais si, comme le suggèrent les progrès actuels, parvenir à l'intelligence au niveau humain signifie mettre en œuvre l'ensemble des capacités générales qui sous-tendent la conscience, c'est moins évident.

LA SINGULARITÉ

- Il est assez facile d'imaginer des êtres qui nous ressemblent mais qui sont plus rapides ou qui ont une meilleure mémoire. Mais si, en fin de compte, le secret de l'intelligence est la capacité à s'intégrer à travers les modalités, à surveiller ses propres états, etc., il est possible qu'il n'y ait aucune différence de nature, mais seulement une différence de degré, entre l'intelligence et la superintelligence. Cela pourrait signifier que nous pouvons rester compétitifs aussi longtemps que nous pouvons trouver des moyens d'étendre nos mémoires, notre puissance de traitement, etc.

LA SINGULARITÉ

- 2) Le problème de l'alignement des IA est de s'assurer que les IA du niveau surhumain (et du niveau humain, et du niveau sous-humain) ont des objectifs qui restent alignés sur les nôtres. Qu'elles ne décident pas, par exemple, de nous transformer tous en trombones de papier.

LA SINGULARITÉ

- 2) Plus les IA surhumaines restent reconnaissables comme nous, plus cela ressemble au problème de s'assurer que nos pairs très intelligents ont des objectifs qui restent alignés sur les nôtres (très différents si la question est de savoir quelles sont les règles classiques dans lesquelles nous devons programmer, pour les instruire sur les actions à ne jamais entreprendre...)