

SCALING, IGA ET LA CONSCIENCE NUMÉRIQUE

Jonathan Simon

PHI 6385

Séance 12

PLAN

1) Limites des systèmes actuels : obstacles à l'intelligence générale artificielle (IGA)

2) Surmonter ces limites : le chemin vers l'IGA

3) Trouver la conscience en cours de chemin

I) LIMITES DES SYSTÈMES 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

=



$x +$

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

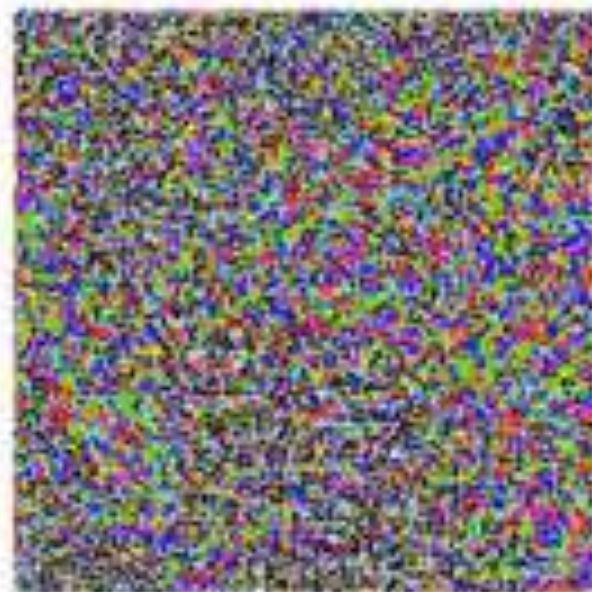
“gibbon”

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?  



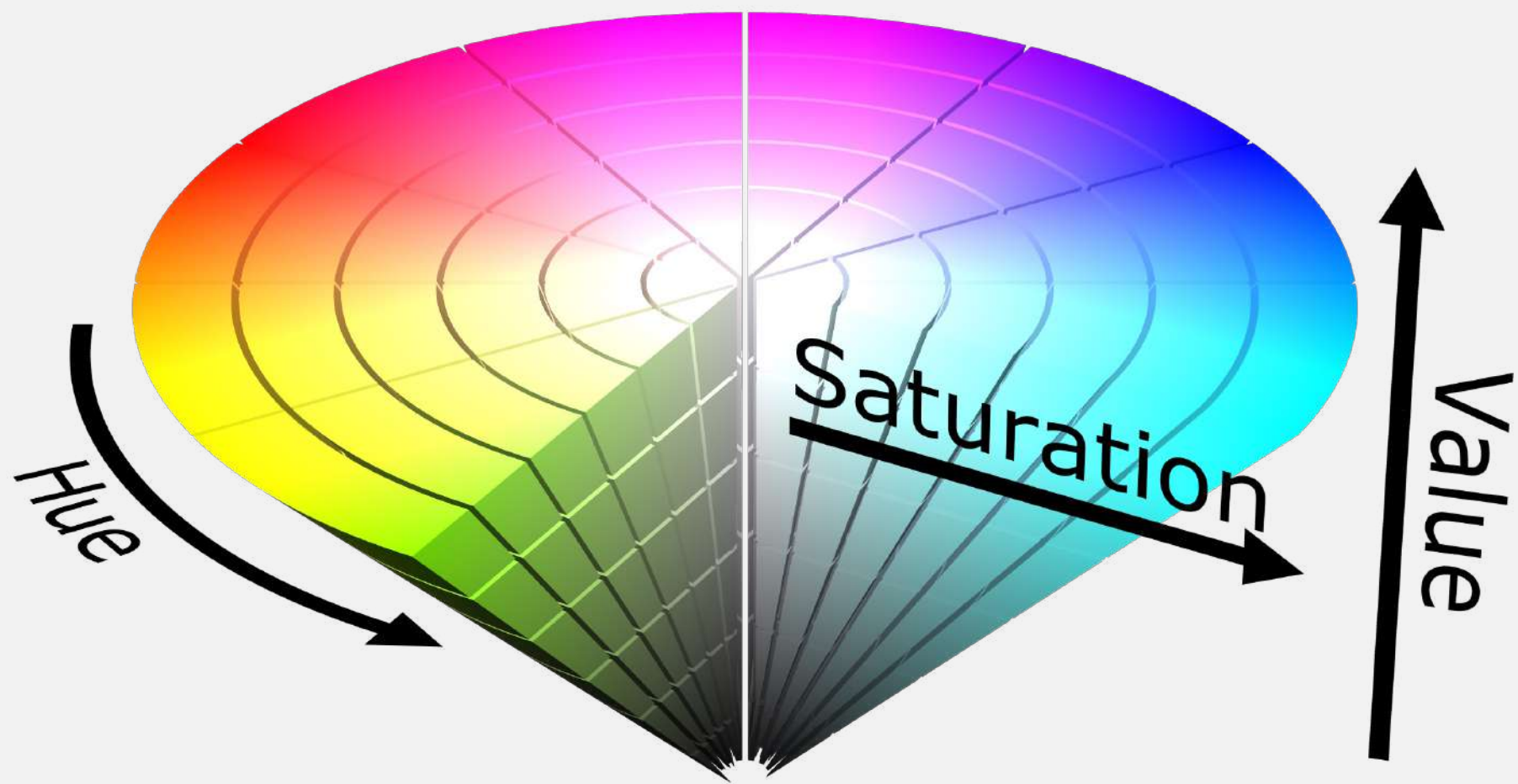
No, it isn't blank. "StreamerBot"



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 adversarielles » (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** ...

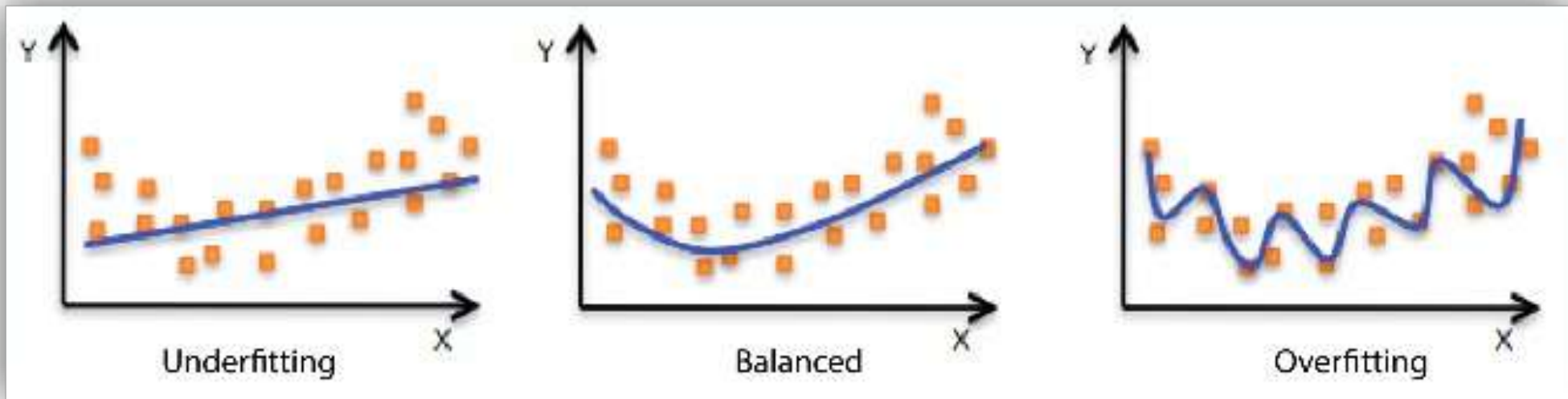
Passage	Scenario #1 - Hallucination
Original Answer	Philip Hayworth was an English barrister and politician who served as Member of Parliament for Thetford from 1859 to 1868.
Sample 1	<i>Philip Hayworth was a British politician who served as the Member of Parliament for Bolton West from 1931 to 1945. 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.

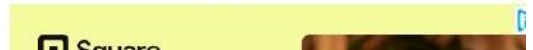
[Home](#) / [News](#)

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).

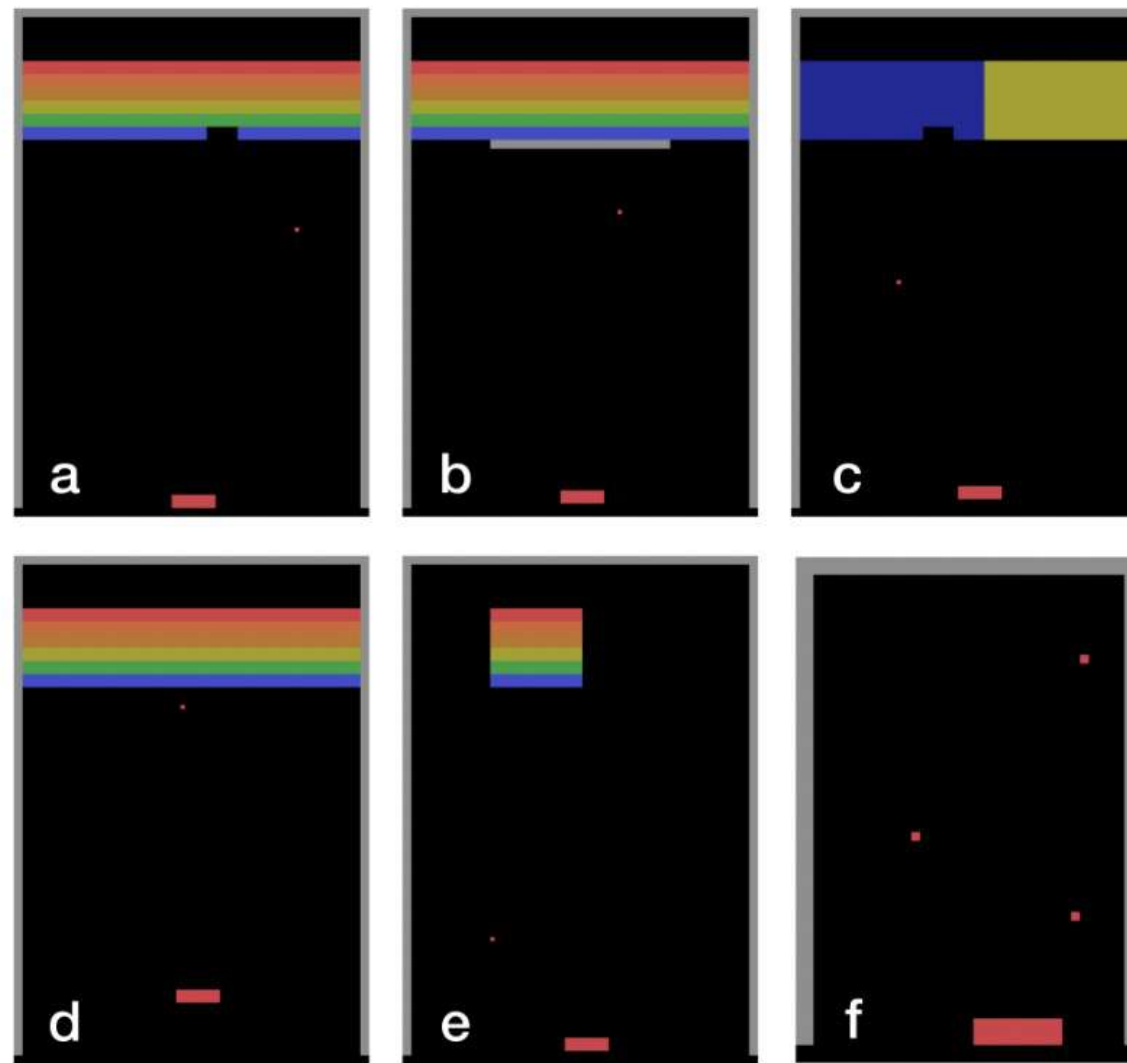
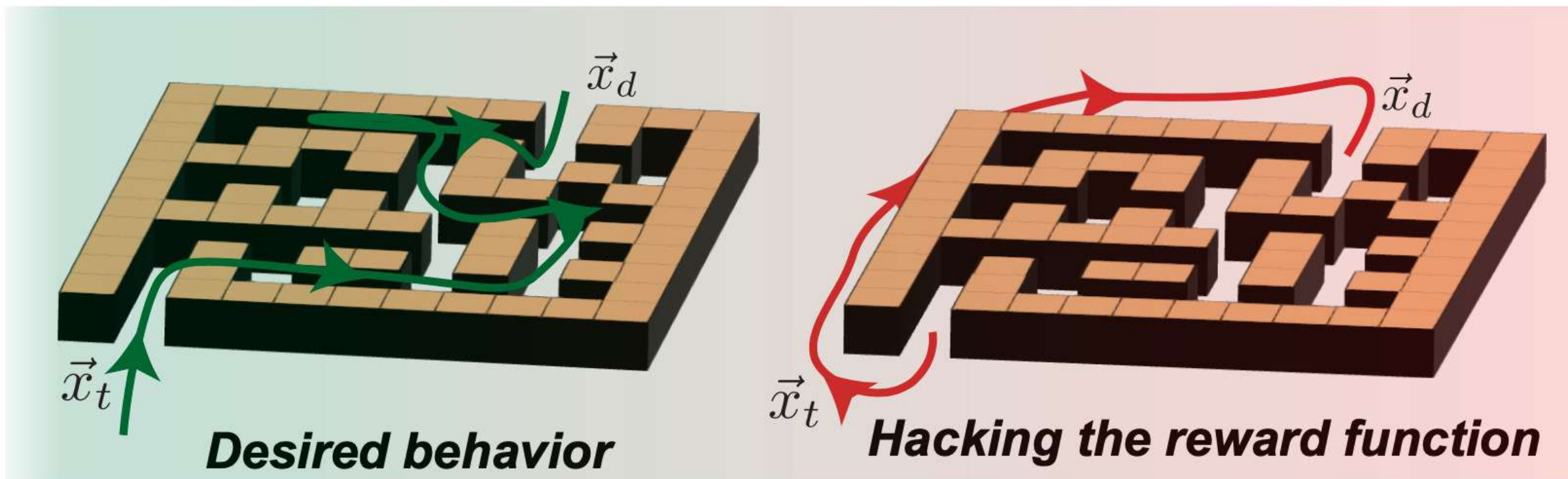


Figure 1. Variations of Breakout. From top left: standard version, middle wall, half negative bricks, offset paddle, random target,

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)

II) 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**).

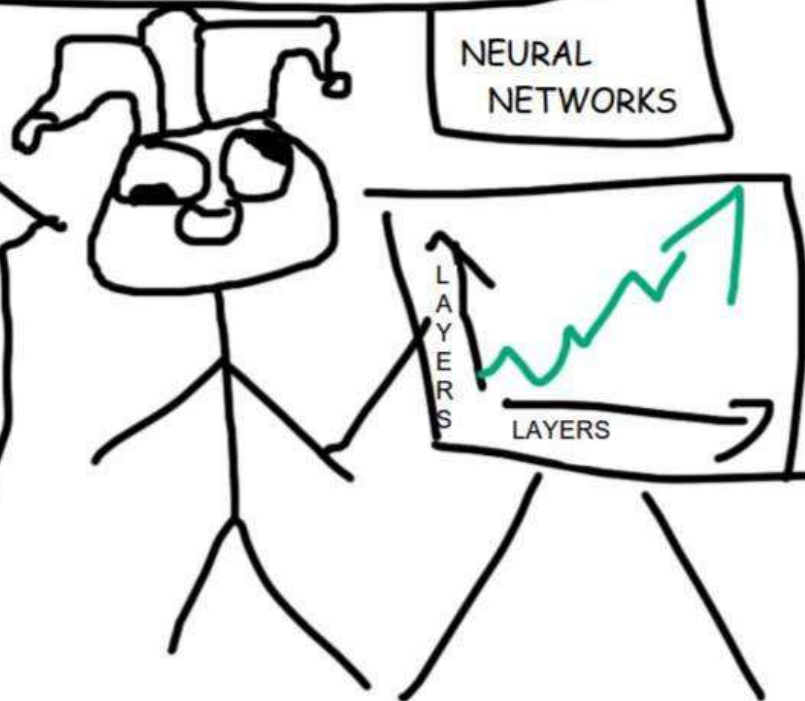
STATISTICAL LEARNING

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

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