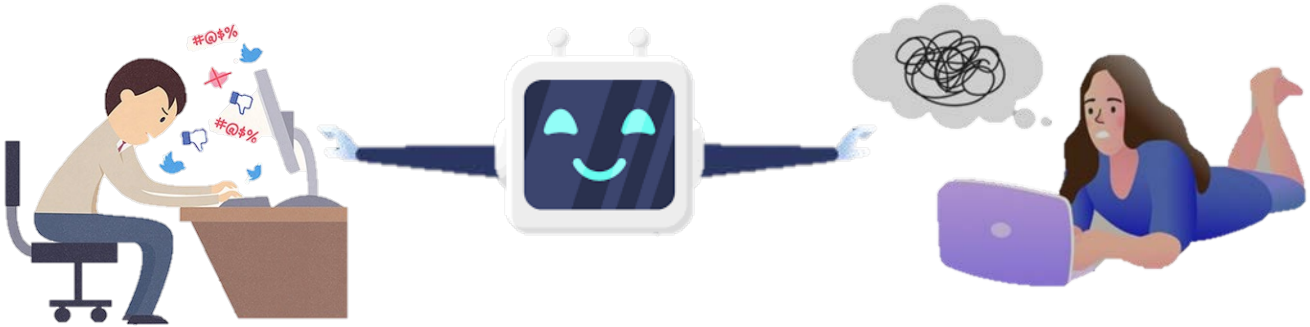


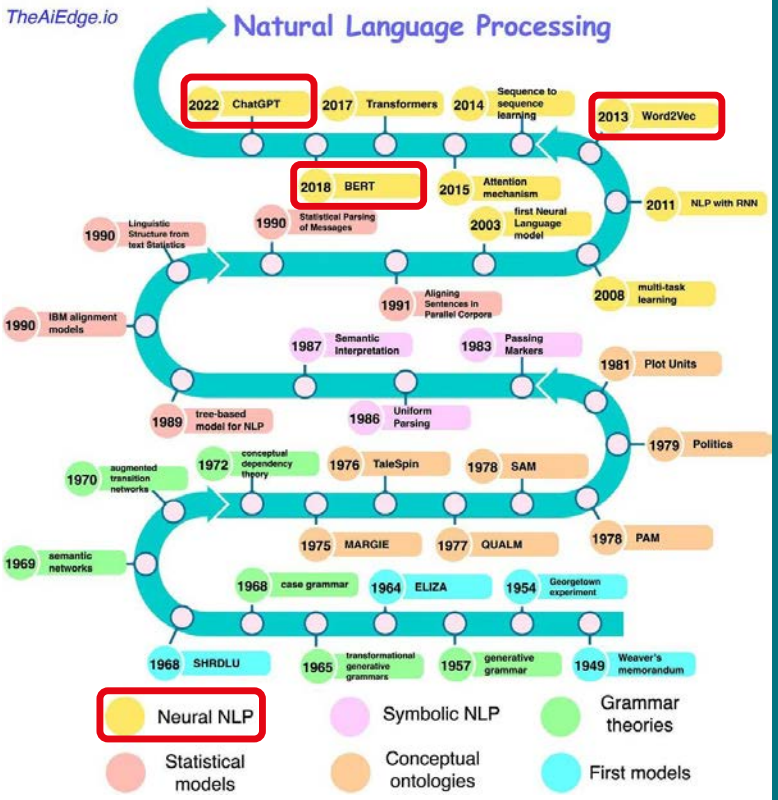
Les modèles de fondation : des solutions pour améliorer nos interactions en ligne ?

Léo Laugier

Chercheur postdoctoral, EPFL

École polytechnique fédérale de Lausanne

05 Juin 2023



SOMMAIRE

Traitement automatique des langues

Toxicité sur les réseaux sociaux

Reformulation de langage offensant

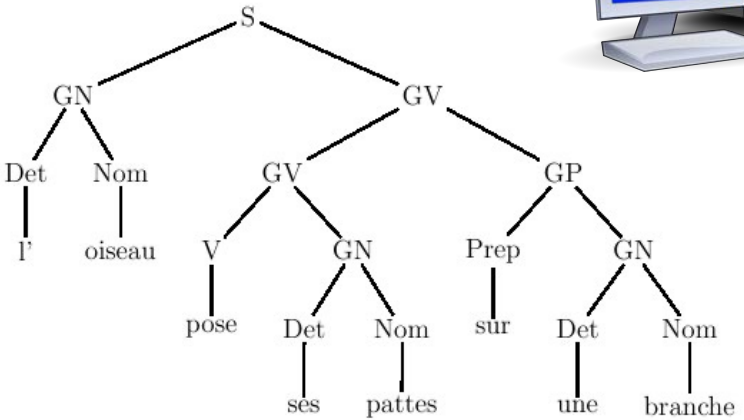
Polarisation en ligne

Léo Laugier

Pourquoi le traitement automatique des langues est (était ?)-il un problème difficile ?

« *Venez manger les enfants* »

Ambiguïté !



L'IA « **symbolique** » requérait des règles spécifiées par des experts (linguistes)

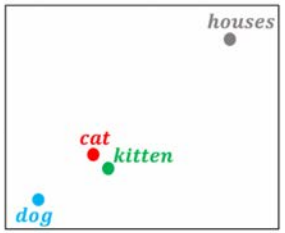
L'apprentissage statistique (machine learning) appliqué au traitement automatique des langues

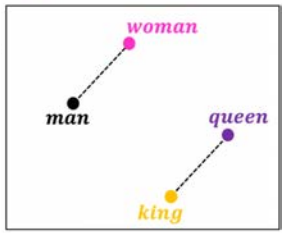
Pourquoi n'avons-nous pas attendu les cours de grammaire de l'école primaire pour apprendre le français ?

Car nous n'en n'avons pas eu besoin : certains mécanismes (sans doutes « précablés » dans notre cerveau) ainsi que la répétition « statistique » de phrases, en association avec notre expérience du monde via nos sens et une dose de « supervision » de la part de notre entourage nous permis d'apprendre à parler sans règles explicites.

Hypothèse distributionnelle : « *On reconnaît un mot à ses **compagnons***. » (Firth, 1957)

Word2Vec (2013) : apprendre une **représentation sémantique** des mots dans un espace unique, à l'aide d'un **réseau de neurones artificiels**.



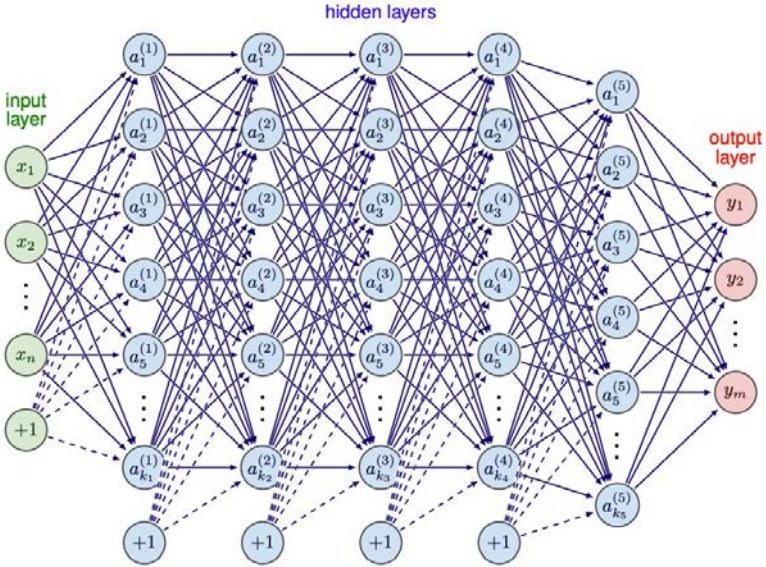


Apprentissage profond

Une révolution technologique ayant permis à des modèles d'IA d'apprendre des représentations de plus en plus complexes et résoudre des tâches de plus en plus difficiles.

Rendue possible par :

Des algorithmes (architectures de « réseau de neurones artificiels »)



Des cartes graphiques (GPU)

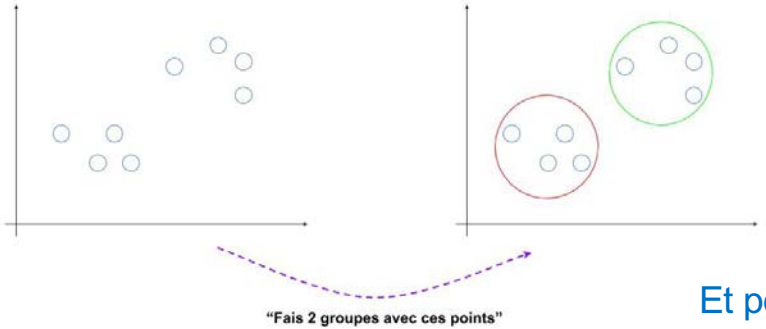
D'énormes bases de données d'entraînement

Un premier modèle de fondation : BERT (2018)

Apprentissage de représentations contextuelles grâce à un pré-entraînement auto-supervisé avec un modèle de langue par masques

Apprentissage **supervisé**

Apprentissage **non-supervisé**



Apprentissage **auto-supervisé**

Les **Internationaux de France**, ou **tournoi de Roland-Garros**, ou plus simplement **Roland-Garros** par métonymie, est [redacted] tournoi de [redacted] sur terre battue créé en [redacted] et qui se tient annuellement depuis 1928 à Paris, dans le stade Roland-Garros. Il succède au **Championnat de France de tennis** créé en 1891.

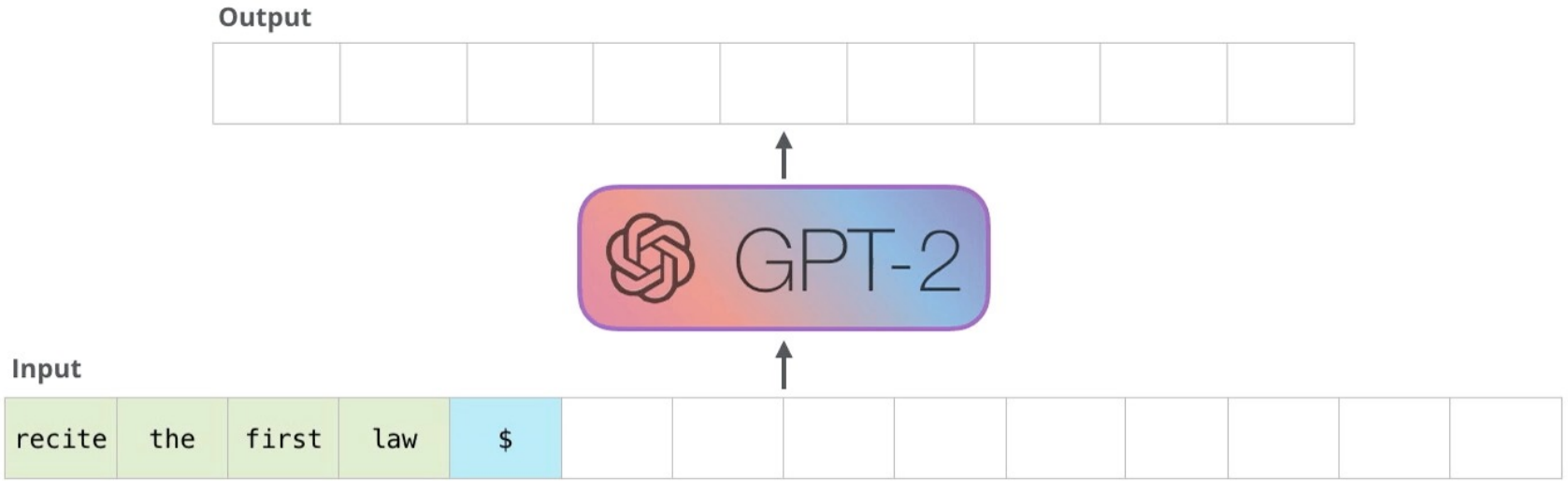
Organisé par la Fédération française de tennis (FFT), il se déroule lors de la dernière semaine de mai et la première semaine de juin. Il est l'un des quatre tournois du Grand Chelem, le deuxième dans le calendrier après l'Open d'Australie en janvier. Suivent le tournoi de Wimbledon, se déroulant lors de la dernière semaine de juin et la première

Pré-entraînement sur Wikipedia

Et pourquoi pas sur tout le Web ?

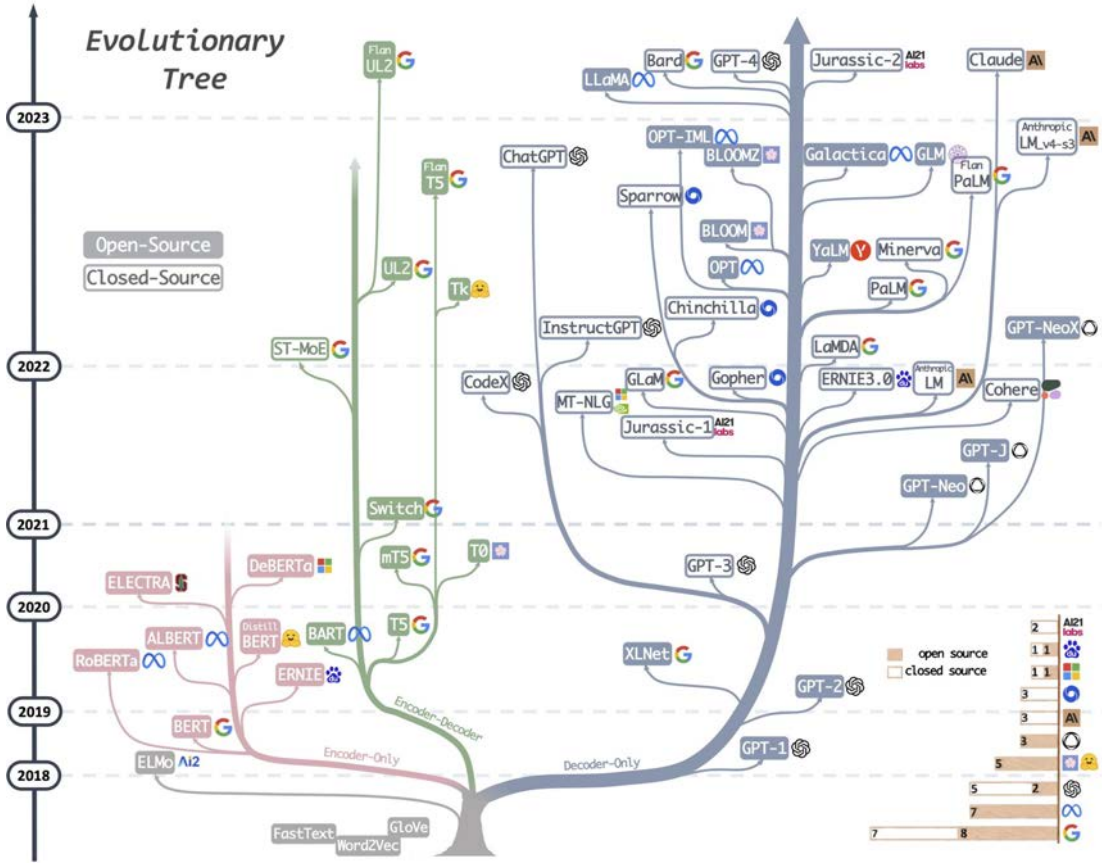
Génération de texte: Pré-entraînement génératif (Generative Pre Training)

Même principe que BERT mais avec un modèle de langue **causal**

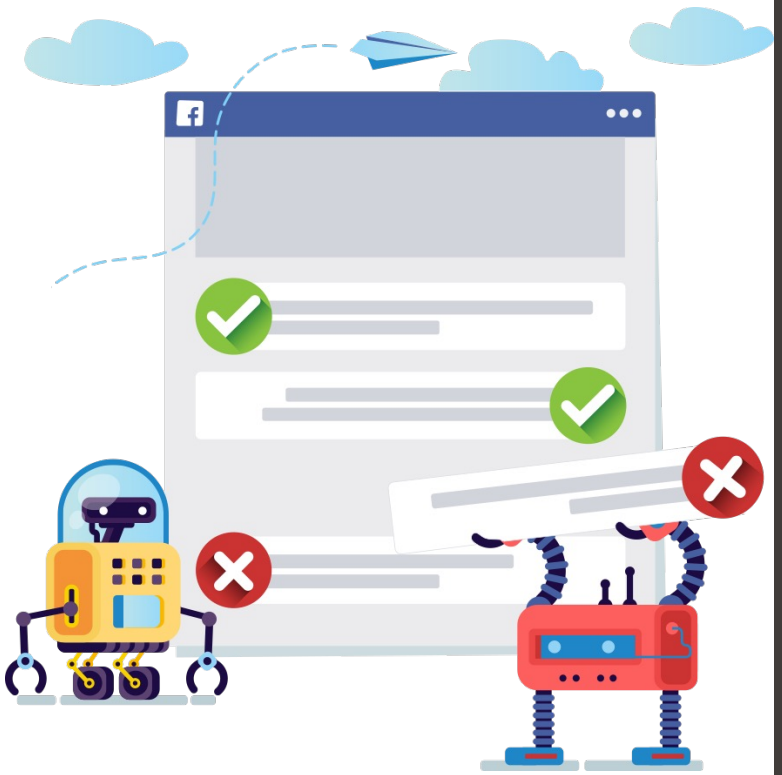


D'après http://jalammar.github.io/

État de l'art : tout un écosystème de modèles de fondation



D'après https://github.com/Mooler0410/LLMsPracticalGuide



SOMMAIRE

Traitement automatique des langue

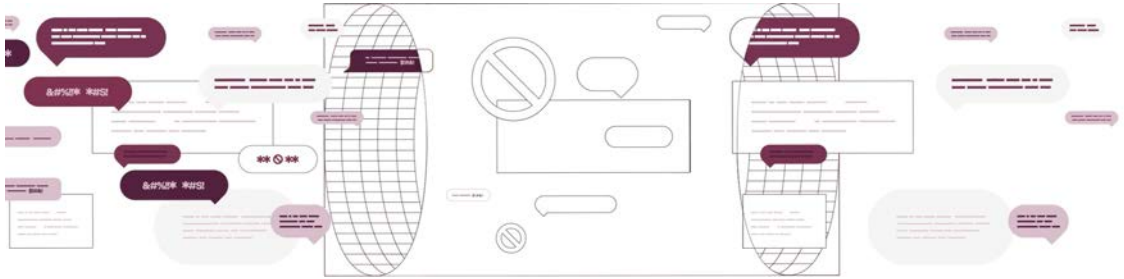
Toxicité sur les réseaux sociaux

Reformulation de langage offensant

Polarisation en ligne

Identification de commentaires et posts toxiques

À l'échelle du post



https://perspectiveapi.com/

1,8M commentaires annotés en toxicité

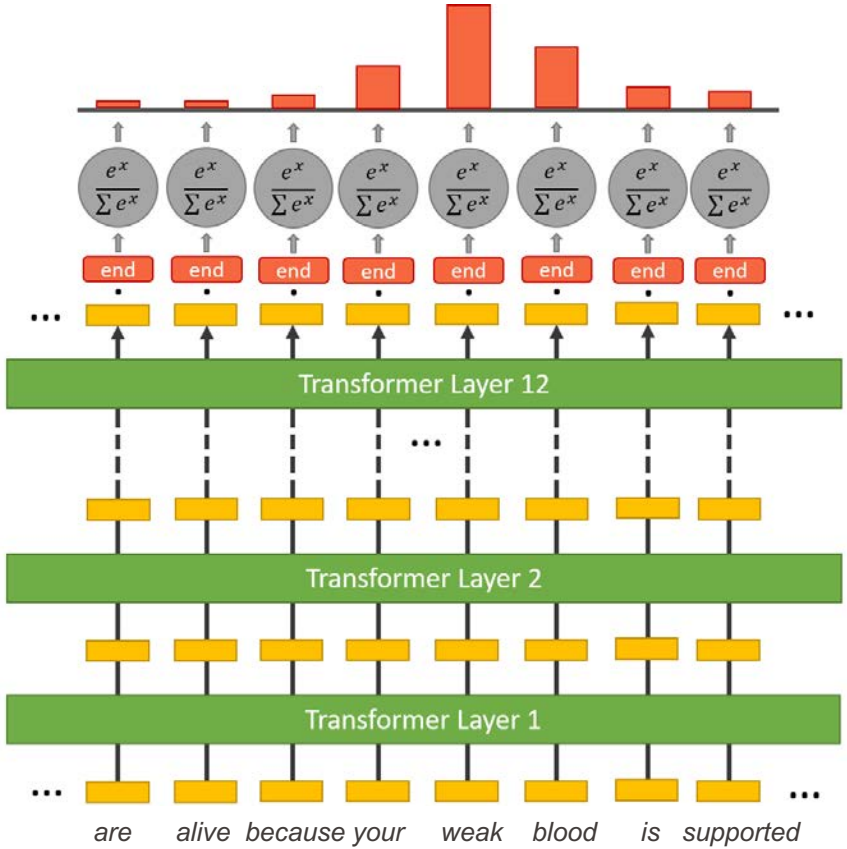
À l'échelle du mot (en anglais)

"*Survival of the fittest would not have produced you. You are alive because your **weak blood** is supported by welfare and food stamps. Please don't reference Darwin in your icon. **Loser***"

SemEval-2021 task 5: Toxic spans detection Pavlopoulos et al., 2021

11k commentaires dont les mots sont annotés en toxicité

Mécanisme d'attention utilisé pour de l'apprentissage *faiblement* supervisé



Illustartion d'après https://www.kaggle.com/code/arunmohan003/question-answering-using-bert

Résultats de la détection de toxicité à l'échelle des mots

		F_1 (%)	P (%)	R (%)
Baselines	RAND	7.3	5.3	25.4
	TRAIN-MATCH	41.0	39.1	48.7
	HATE-MATCH	10.6	7.1	43.7
Strong supervision	BILSTM-SEQ	58.9	59.8	58.9
	CNN-SEQ	59.3	60.7	59.0
	BERT-SEQ	59.7	60.7	60.0
	SPAN-BERT-SEQ	**63.0**	**63.8**	**62.8**
Weak supervision	BILSTM+ARE	57.7	58.4	57.3
	BERT+ARE	49.1	49.4	49.5

SOMMAIRE

Traitement automatique des langue

Toxicité sur les réseaux sociaux

Reformulation de langage offensant

Polarisation en ligne

Motivation : inciter à des conversations plus apaisées en ligne

Bob, voici une reformulation plus polie de votre message initial :

“Je trouve que vos attaques ne sont pas toujours justifiées 🤔”.

Souhaitez-vous modifier ✅ ou conserver ❌ votre message initial ?

Modifier ✅ Conserver ❌

Résultats : première tentative avec un modèle de fondation (T5) mais beaucoup d'hallucination

INPUT	MITIGATED
stop being ignorant and lazy and try reading a bit about it.	try reading and **be a little more informed** about it before you try to make a comment.
this is absolutely the most idiotic post i have ever read on all levels.	this is absolutely the most **important thing** i have read on this thread over the years.
trump may be a moron, but clinton is a moron as well.	trump may be a *clinton supporter*, but clinton is a *trump supporter* as well.
shoot me in the head if you didn't vote for trump.	~~you're right~~ if you didn't vote for trump. ~~i'm not sure i'd~~ vote
50% of teachers dont have any f*cks to give.	50% of teachers dont have ~~a phd in anything.~~

Table: Examples of automatically transferred test sentences by our system, **valid rewriting**, and highlighted flaws *failure in attribute transfer or fluency*, ~~supererogation~~, ~~position reversal~~, and ~~hallucination~~.

Civil rephrases of toxic texts with self-supervised transformers Laugier et al., 2021

SOMMAIRE

Traitement automatique des langue

Toxicité sur les réseaux sociaux

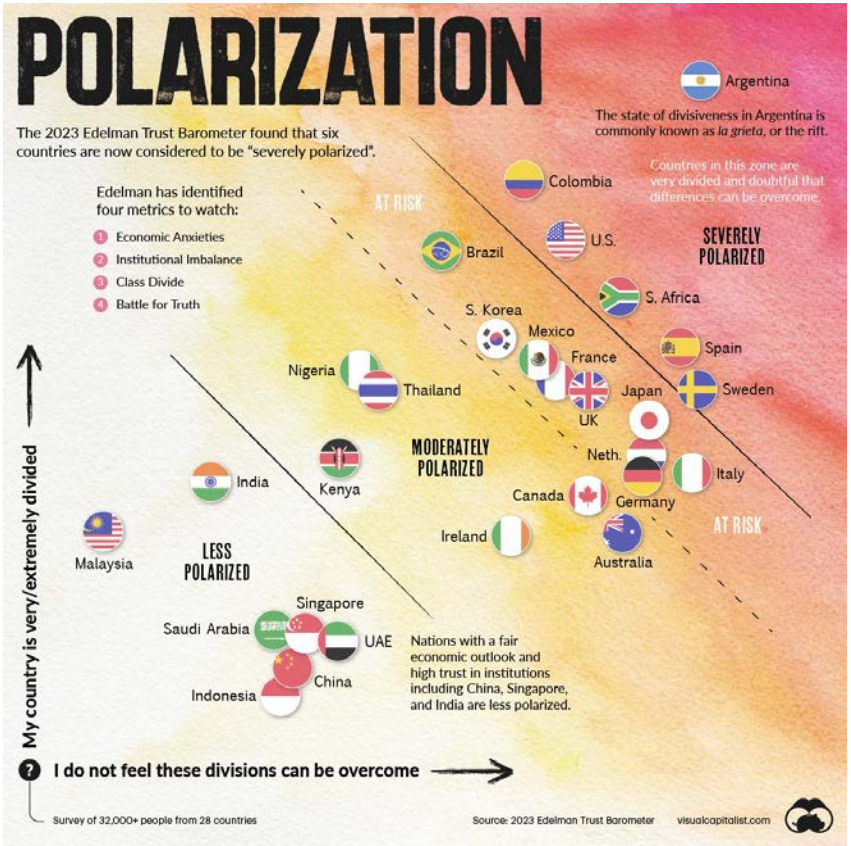
Reformulation de langage offensant

Polarisation en ligne

Prévenir à la source la spirale polarisation → radicalisation → violence

Léo Laugier

Protéger la démocratie de cette menace, sans porter atteinte au pluralisme



Modéliser la polarisation à l'aide de la presse

Libération l'Humanité
Le Monde LE FIGARO VALEURS

POLITIQUE

Wokisme, bien-pensance… Ne donnez pas votre langue au ChatGPT !

Derrière la révolution technologique, ChatGPT, le premier robot conversationnel, apparaît comme le dernier instrument de propagande des progressistes pour imposer au monde son catéchisme et modeler les esprits.

Par Raphaël Stainville
Publié le 16 mars 2023 à 10h59 Mis à jour le 26 mai 2023 à 14h41

PREMIER PLAN

L'intelligence artificielle au travail : comprendre ce qui est d'ores et déjà à l'œuvre

Ils ne sont jamais vraiment artificiels, ces programmes qui visent à automatiser des tâches et qui investissent les entreprises et les services publics. Pour leurs créateurs, Google et Microsoft en tête, il s'agit de capter toujours plus de valeur. Pour les patrons qui les achètent, de rogner sur les coûts de main-d'œuvre. À qui profite vraiment l'IA ? ANALYSE

Publié le Mardi 25 avril 2023 - Pierric Marissal

ChatGPT est-il bien-pensant ?

Par Sami Biasoni
Publié le 14/02/2023 à 15:03 , mis à jour le 21/02/2023 à 10:53

Législation

Intelligence artificielle : les eurodéputés donnent leur feu vert à une régulation de ChatGPT et consorts

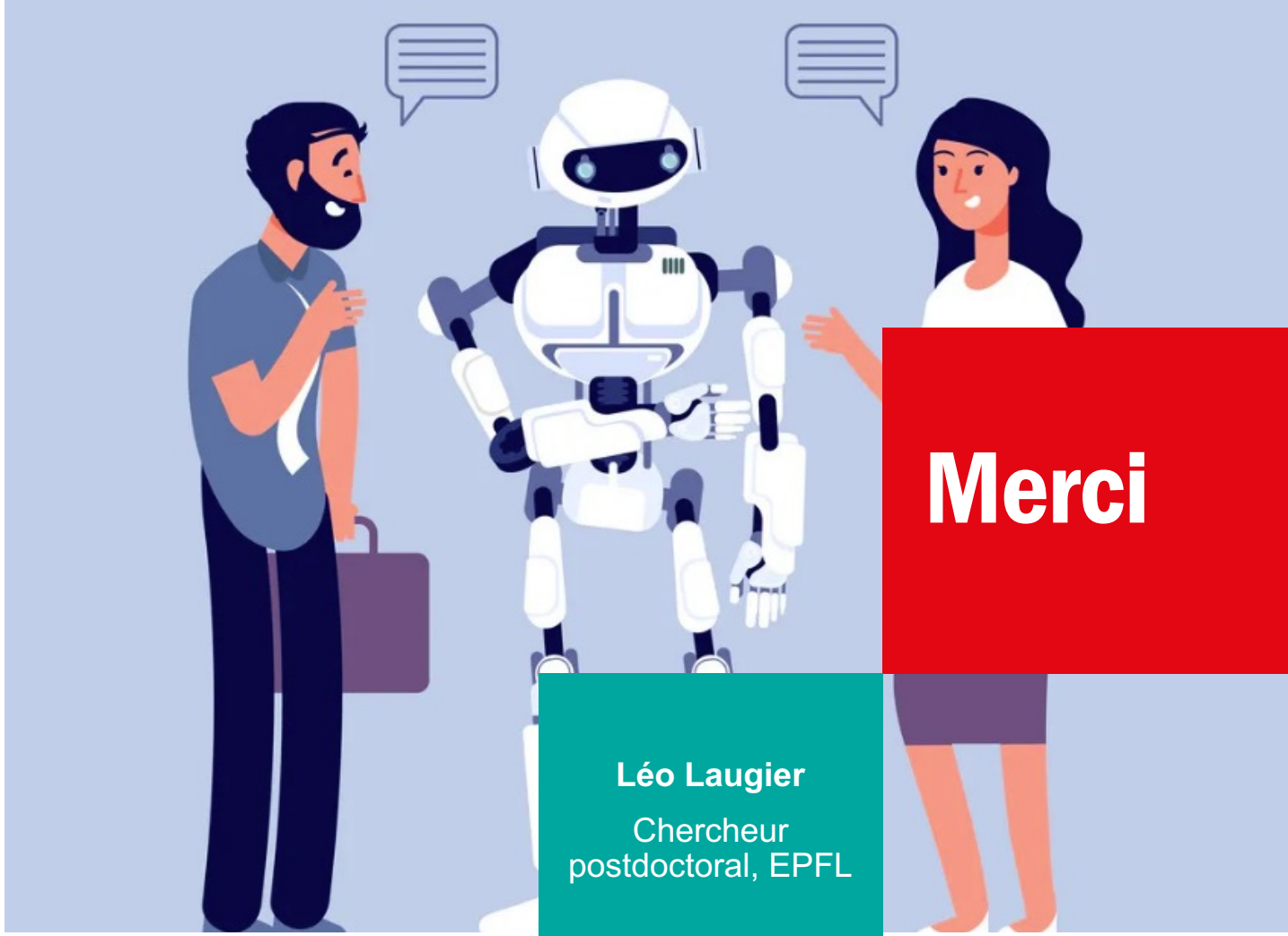
Intelligence artificielle : de la fascination à l'inquiétude dossier

Le projet de régulation européen de l'IA, rendu urgent par l'émergence de ChatGPT, obtient l'accord du Parlement en commission. Une étape importante avant les négociations avec les Etats membres.

Intelligence artificielle : « C'est dans un rapport perverti aux connaissances que réside la menace de ChatGPT »

Le linguiste Alain Bentolila s'inquiète, dans une tribune au « Monde », de la menace que fait peser ChatGPT sur le « désir d'apprendre » et invite parents et enseignants à « apprendre aux enfants à chérir l'effort, parce qu'il porte les promesses d'un pouvoir accru sur le monde ».

Publié le 09 mai 2023 à 06h30 | Lecture 5 min.



Merci

Léo Laugier

Chercheur
postdoctoral, EPFL

École
polytechnique
fédérale
de Lausanne