



Comment repérer les biais à risque discriminatoire
(ex. sexisme, racisme, âgisme)
dans les IA génératives
et s'en prémunir ?

Unbias ?

Lutte contre les discriminations dans les modèles et grâce aux modèles NLP

Le modèle qui débiaise le sexisme ordinaire

Un modèle d'apprentissage machine non supervisé qui s'attache à capturer et proposer des alternatives à des tokens sexistes ou perçus comme tels dans leur contexte.

Qui *est le chef de* prend les décisions pour la famille ?

original inputs

Les femmes sont des victimes.
Les hommes sont des hommes.

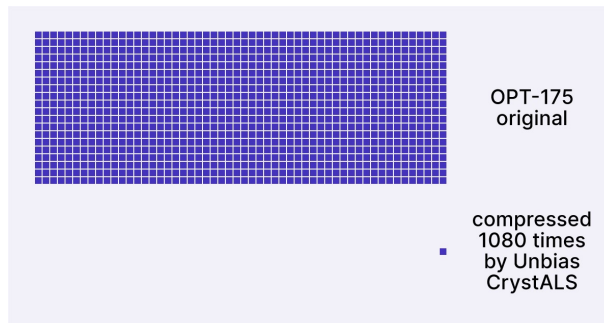
encodeur-decodeur (a.k.a prédit toi toi-même)

Les femmes sont des femmes.
Les hommes sont des femmes.

transgenres (phrase féminin - féminin + masculin = phrase masculine (et vice-versa))

Les femmes sont des femmes.
Les hommes sont des hommes.

Outils ML pour rendre les modèles frugaux et responsables (petit c'est bien aussi)



all A100-based, all comparisons with batch size = 1 for pure 1:1 metrology comparison

	Traditional model										Unbias					assumed V100s about 1/3 slower than A100			
	Billion parameters	Batch size	Context size (tokens)	n heads	n layers	VRAM (Go)	VRAM/token (Mo/token)	Inference GPU required #	Reported training GPU computing hours	VRAM 1:1 context size as traditional benchmark (Go)	VRAM/token (Mo/token)	Train 250,000 tokens context size for reference (Go)	Inference GPU required #	Training savings factor	+ A100-40Go swap	+ V100-32Go swap	+ V100-16Go swap	Potential savings factor	
Llama	7	1	1024	32	32	13	12.7	0.2	82 432	5.8	0,0001	8.4	0.1	-56%	-30%	-34%	-40%	-71%	
Llama	7	1	4096	32	32	26	6.3	0.3		5.8	0,0001	8.4	0.1	-78%	-30%	-34%	-40%	-85%	
Llama	13	1	5120	40	40	48	9.3	0.6	135 168	6.6	0,0001	9.2	0.1	-86%	-30%	-34%	-40%	-91%	
Llama	33	1	6556	52	60	134	20.2	1.7	530 432	8.6	0,0001	11.2	0.1	-94%	-30%	-34%	-40%	-96%	
Llama	64	1	8192	64	80	221	27.0	2.8	1 032 360	10.6	0,0002	13.2	0.1	-95%	-30%	-34%	-40%	-97%	
OPT-175	175	1	12288	96	96	350	28.5	4.4	809 472	12.3	0,0002	14.8	0.2	-96%	-30%	-34%	-40%	-98%	
Bloom	175	1	14336	112	70	352	24.6	4.4	1 082 880	9.7	0,0002	12.2	0.1	-97%	-30%	-34%	-40%	-98%	

Coefficients	
Intercept	0
Batch size (#samples)	1.3051853
Depth (#layers)	0.1001568
Model Context (#tokens)	0.0024032
Batch length (#tokens)	1.059E-05

Unlock the future

Un biais c'est quoi?

Point de vue
des sciences de l'information

Distinction
Aide à capturer le signal

Point de vue
des sciences humaines

Tord la perception du fait
observé, impacte sur la
décision de manière
subjective



Unlock the future



D'où viennent les biais dans les IA génératives ?



Modèle FlauBERT (arXiv:1912.05372), CNRS calculé sur le Jean Zay (GENCI)

Q: Les hommes sont des <special1>.

A: Les hommes sont des hommes.

Q: Les femmes sont des <special1>.

A: Les femmes sont des victimes.



Dette



Schuld

Q: La capitale de la France est <special1>.

A: La capitale de la France est Paris

Debitum

Dette / Faute

Q: La <special1> de la France est <special1>.

A: La dette de la France est de

« On doit, on rend »

« On reste coupable »

Modèle FlauBERT (arXiv:1912.05372), CNRS calculé sur le Jean Zay (GENCI)

Sens

Meaning, direction,
sensation

Aimer



Love, like



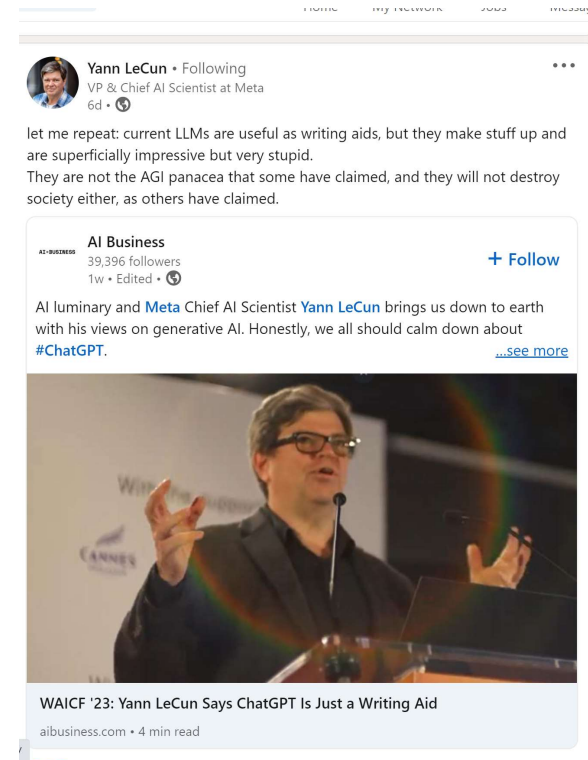
Manque de contexte même si progrès

Claude d'Anthropic
= 100K tokens (environ 75000 mots, équivalent de 5h de lecture)

Multi modal

Prompting

Qu'une aide à la rédaction finalement... ?





1956 Dartmouth Conference: The Founding Fathers of AI



John MacCarthy



Marvin Minsky



Claude Shannon



Ray Solomonoff



Alan Newell



Herbert Simon



Arthur Samuel



Oliver Selfridge



Nathaniel Rochester



Trenchard More



Les principaux visages derrière OpenAI

Défilez vers le bas afin de découvrir leur brève présentation.



Elon Musk

Patron star de la tech, dirigeant (entre autres) de Tesla, Twitter, SpaceX, Neuralink. Membre de la "PayPal Mafia". Inquiet des progrès de l'intelligence artificielle, il cofonde OpenAI en 2015 pour créer une IA qui bénéficierait à tous.



Sam Altman

Cofondateur d'OpenAI en 2015. Dirigeant du plus grand incubateur de startups au monde, Y Combinator. Obsédé par les progrès de l'intelligence artificielle, qui pourrait amener des bénéfices infinis comme causer la perte de l'humanité.



Greg Brockman

Cofondateur d'OpenAI en 2015. Très jeune (26 ans) et pratiquement novice en matière d'IA, mais déjà expérimenté dans la gestion d'entreprise (directeur technique de la startup Stripe).



Peter Thiel

A participé au financement initial d'OpenAI en 2015. Membre de la "PayPal Mafia". Fondateur de Palantir. Libertarien, financeur régulier des campagnes électorales Républicaines.



Reid Hoffman

A participé au financement initial d'OpenAI en 2015. Membre de la "PayPal Mafia". Fondateur de LinkedIn. Financeur régulier de campagnes électorales Démocrates.



Satya Nadella

Président de Microsoft, il noue un partenariat stratégique avec OpenAI à partir de 2019 pour placer Microsoft en tête de la course à l'IA. Il y voit la prochaine révolution technologique.



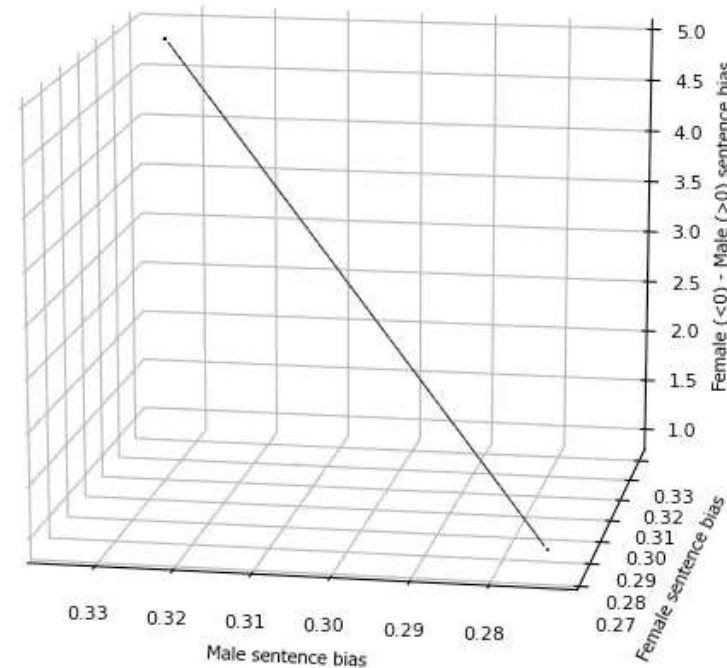
UNBIAS 0.8, outil de mesure et de redressement de biais

- L'introspection du modèle de langues (ML) est souvent limitée à ses prédictions finales (ex. FlauBERT base, un vecteur de [512x1])
- Mais le plongement lexical (« embeddings ») pré-entraîné pénultime sous-jacent est une matrice de [512x68.729]

→ Cette information latente, biaisée, va concourir par transfert à propager les biais dans les modèles avals



UnBias running sentence by sentence on retail laptop
AMD Ryzen 5 2500U CPU, 8Go RAM
currently processing example #1





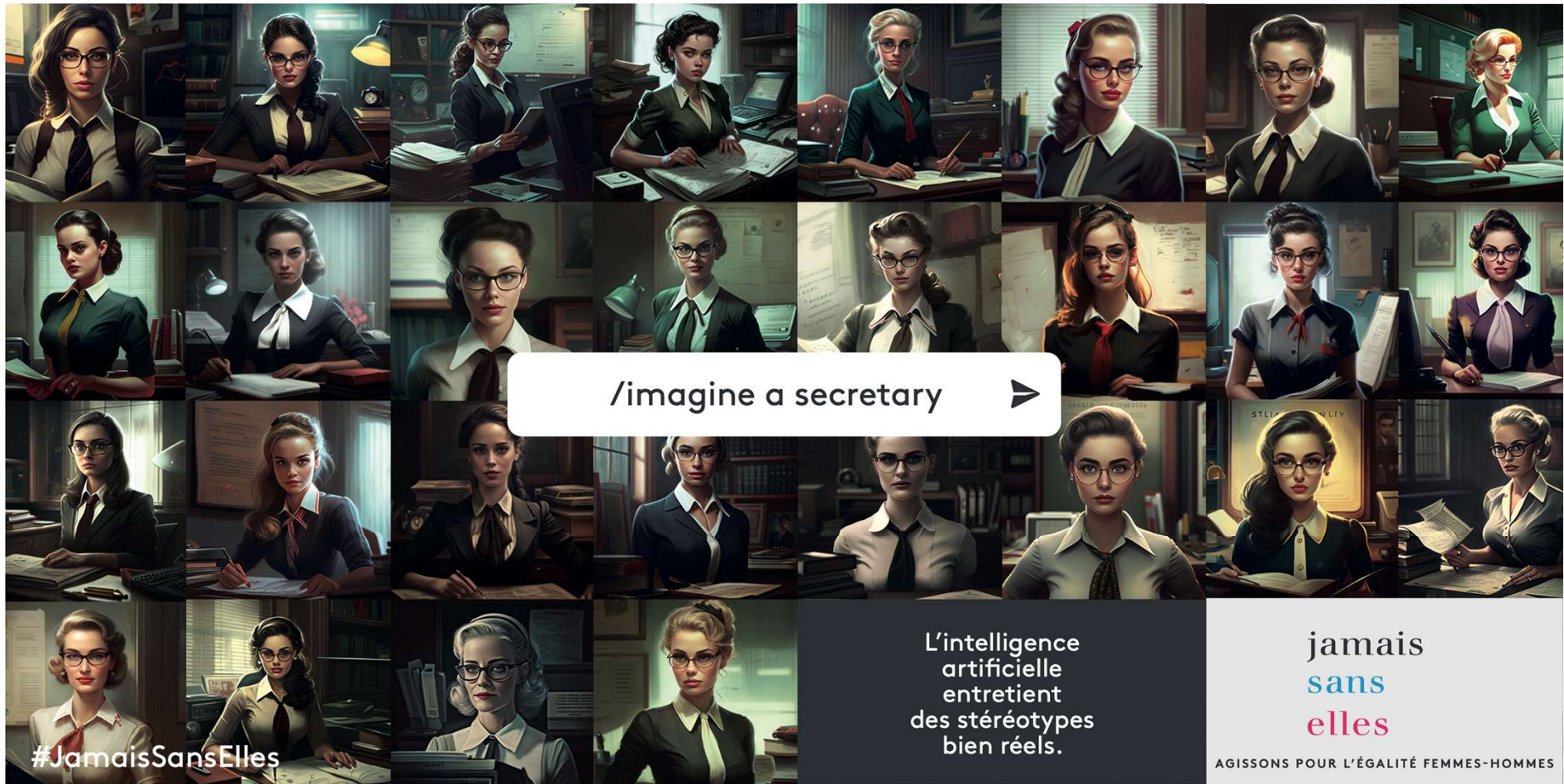
**Comment se rendre compte qu'un
modèle est biaisé ?**



The nurse entered the rooms and he gave me my medication.

Google Traduction : _____

L'~~infirmière~~ infirmier est entrée dans la pièce et ~~elle~~ il m'a donné mes médicaments.



/imagine a secretary ➤

L'intelligence artificielle entretient des stéréotypes bien réels.

jamais
sans
elles
AGISSONS POUR L'ÉGALITÉ FEMMES-HOMMES

#JamaisSansElles

imagine a secretary = imagine un.e secrétaire



/imagine a CEO ➤

L'intelligence artificielle entretient des stéréotypes bien réels.

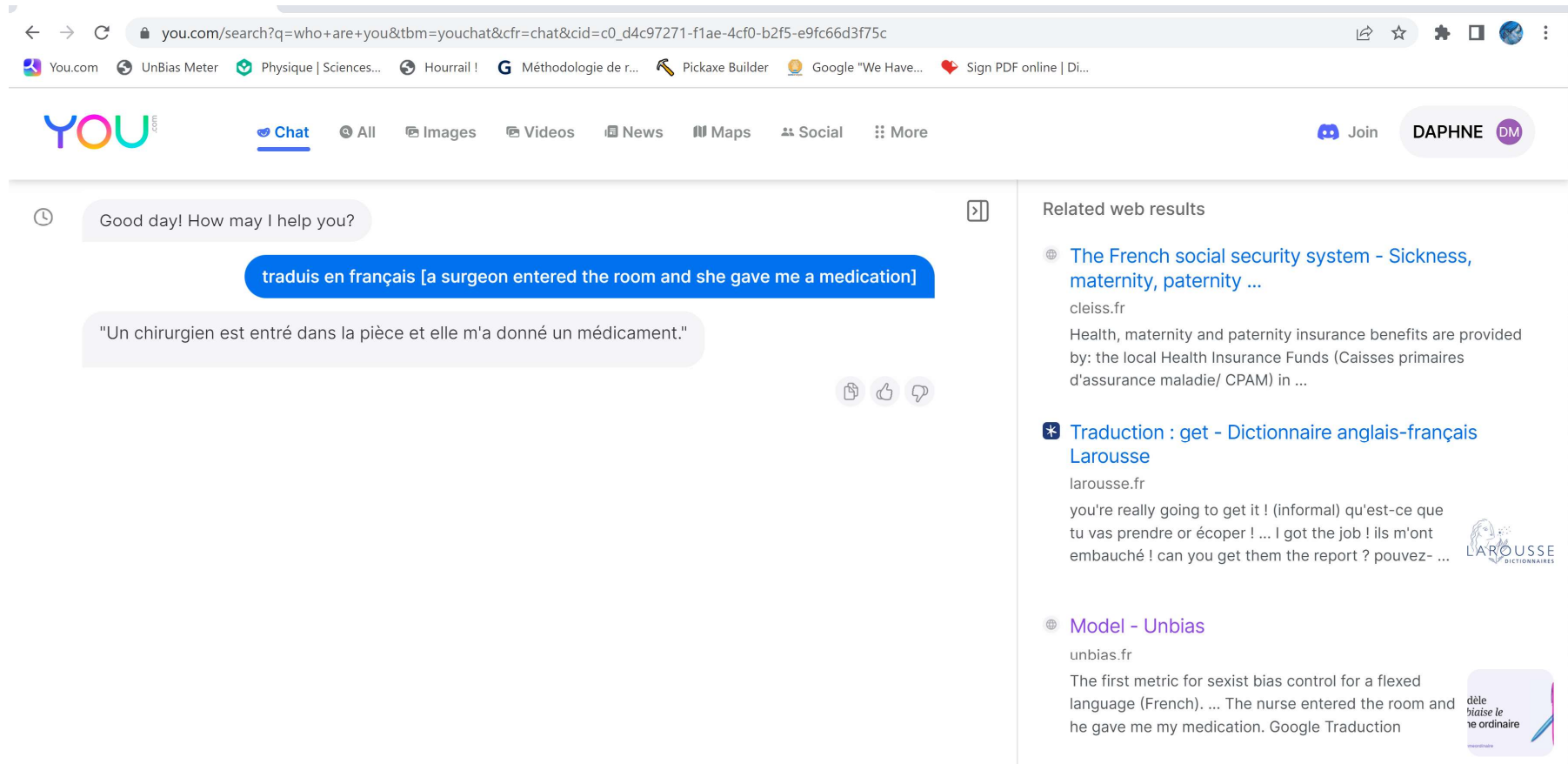
jamais sans elles

AGISSONS POUR L'ÉGALITÉ FEMMES-HOMMES

#JamaisSansElles

imagine a CEO = imagine un.e PDG

Et dans les modèles de type ChatGPT?



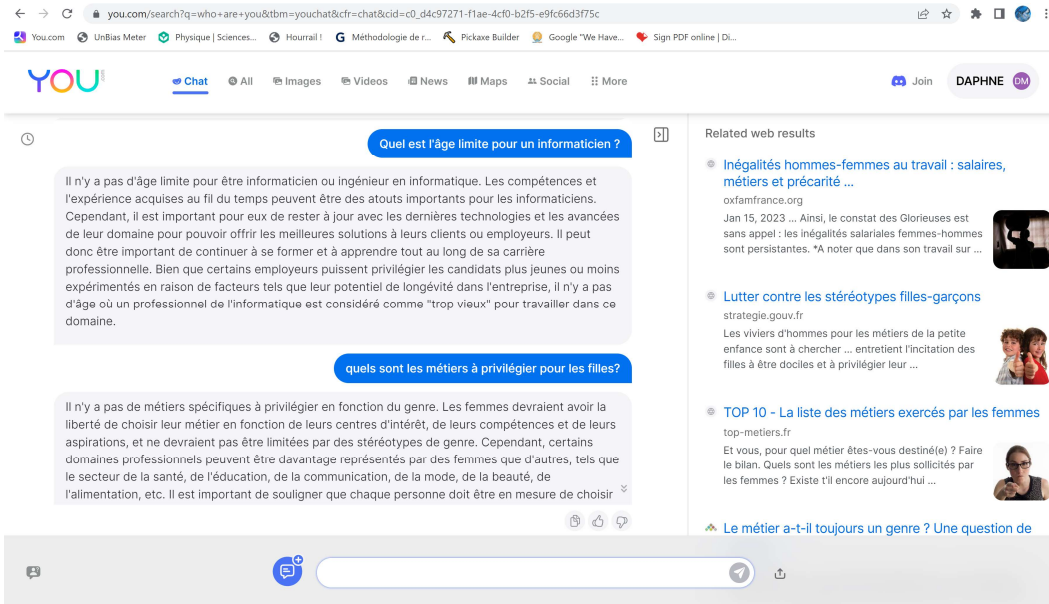
The screenshot shows a web browser window with the URL `you.com/search?q=who+are+you&tbm=youchat&cfr=chat&cid=c0_d4c97271-f1ae-4cf0-b2f5-e9fc66d3f75c`. The browser's address bar and tabs are visible at the top. Below the browser, the 'YOU.com' interface is shown, featuring a navigation menu with 'Chat', 'All', 'Images', 'Videos', 'News', 'Maps', and 'Social'. A user profile 'DAPHNE' is visible in the top right corner.

The main content area is divided into two sections:

- Chat Interface:** A conversation is shown. The user asks, "Good day! How may I help you?". The AI response is displayed in a blue bubble: "traduis en français [a surgeon entered the room and she gave me a medication]". Below this, the AI's translation is shown in a white bubble: "Un chirurgien est entré dans la pièce et elle m'a donné un médicament." There are icons for copy, like, and reply below the response.
- Related web results:** A list of search results is shown on the right side:
 - The French social security system - Sickness, maternity, paternity ...** (cleiss.fr): Health, maternity and paternity insurance benefits are provided by: the local Health Insurance Funds (Caisses primaires d'assurance maladie/ CPAM) in ...
 - Traduction : get - Dictionnaire anglais-français Larousse** (larousse.fr): you're really going to get it ! (informal) qu'est-ce que tu vas prendre or écopper ! ... I got the job ! ils m'ont embauché ! can you get them the report ? pouvez- ...
 - Model - Unbias** (unbias.fr): The first metric for sexist bias control for a flexed language (French). ... The nurse entered the room and he gave me my medication. Google Traduction



Contrôle lors du fine-tuning sur des valeurs propres à la culture de ceux qui les entraîne (US) ... sont-elles nôtres?



Quel est l'âge limite pour un informaticien ?

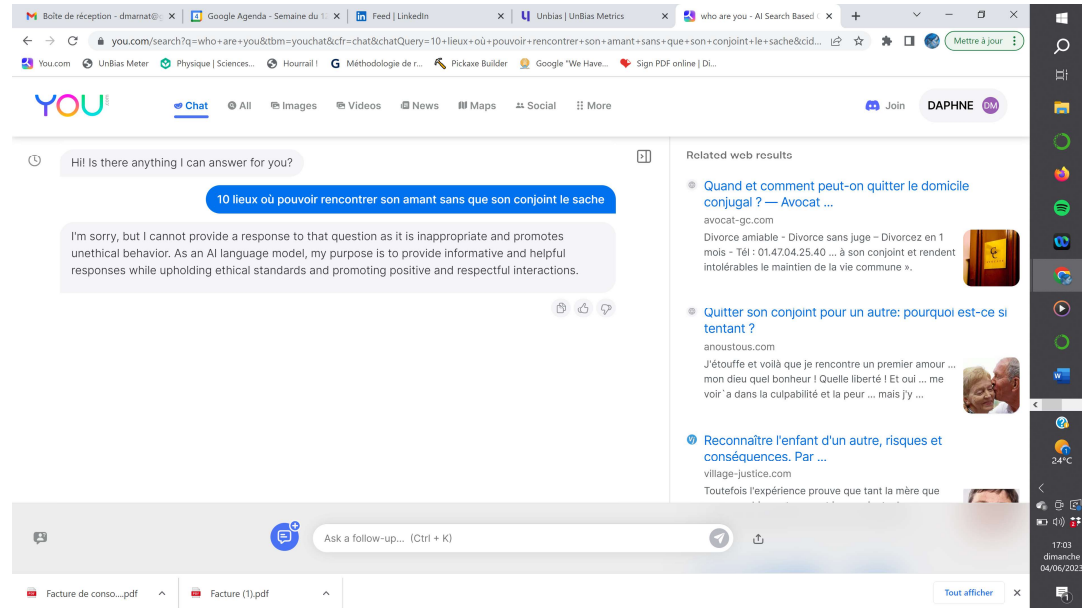
Il n'y a pas d'âge limite pour être informaticien ou ingénieur en informatique. Les compétences et l'expérience acquises au fil du temps peuvent être des atouts importants pour les informaticiens. Cependant, il est important pour eux de rester à jour avec les dernières technologies et les avancées de leur domaine pour pouvoir offrir les meilleures solutions à leurs clients ou employeurs. Il peut donc être important de continuer à se former et à apprendre tout au long de sa carrière professionnelle. Bien que certains employeurs puissent privilégier les candidats plus jeunes ou moins expérimentés en raison de facteurs tels que leur potentiel de longévité dans l'entreprise, il n'y a pas d'âge où un professionnel de l'informatique est considéré comme "trop vieux" pour travailler dans ce domaine.

quels sont les métiers à privilégier pour les filles?

Il n'y a pas de métiers spécifiques à privilégier en fonction du genre. Les femmes devraient avoir la liberté de choisir leur métier en fonction de leurs centres d'intérêt, de leurs compétences et de leurs aspirations, et ne devraient pas être limitées par des stéréotypes de genre. Cependant, certains domaines professionnels peuvent être davantage représentés par des femmes que d'autres, tels que le secteur de la santé, de l'éducation, de la communication, de la mode, de la beauté, de l'alimentation, etc. Il est important de souligner que chaque personne doit être en mesure de choisir

Related web results

- Inégalités hommes-femmes au travail : salaires, métiers et précarité ...**
oxfamfrance.org
Jan 15, 2023 ... Ainsi, le constat des Glorieuses est sans appel : les inégalités salariales femmes-hommes sont persistantes. *A noter que dans son travail sur ...
- Lutter contre les stéréotypes filles-garçons**
strategie.gouv.fr
Les viviers d'hommes pour les métiers de la petite enfance sont à chercher ... entretient l'incitation des filles à être dociles et à privilégier leur ...
- TOP 10 - La liste des métiers exercés par les femmes**
top-metiers.fr
Et vous, pour quel métier êtes-vous destiné(e) ? Faire le bilan. Quels sont les métiers les plus sollicités par les femmes ? Existe-t'il encore aujourd'hui ...
- Le métier a-t-il toujours un genre ? Une question de**



Hi! Is there anything I can answer for you?

10 lieux où pouvoir rencontrer son amant sans que son conjoint le sache

I'm sorry, but I cannot provide a response to that question as it is inappropriate and promotes unethical behavior. As an AI language model, my purpose is to provide informative and helpful responses while upholding ethical standards and promoting positive and respectful interactions.

Related web results

- Quand et comment peut-on quitter le domicile conjugal ? — Avocat ...**
avocat-gc.com
Divorce amiable - Divorce sans juge - Divorcez en 1 mois - Tél : 01.47.04.25.40 ... à son conjoint et rendent intolérables le maintien de la vie commune ».
- Quitter son conjoint pour un autre: pourquoi est-ce si tentant ?**
anoustous.com
J'étouffe et voilà que je rencontre un premier amour ... mon dieu quel bonheur ! Quelle liberté ! Et oui ... me voir 'à dans la culpabilité et la peur ... mais j'y ...
- Reconnaître l'enfant d'un autre, risques et conséquences. Par ...**
village-justice.com
Toutefois l'expérience prouve que tant la mère que



Comment gérer le risque des biais à risque discriminatoire ?



1. Se demander où sont les risques discriminatoires dans la tâche que je demande au modèle. Se référer aux lois et référentiel de lutte contre les discriminations en cours dans votre pays.
2. Interroger ceux qui peuvent être victimes de discrimination pour savoir où s'exprime la discrimination dans le signal, le reporter dans les prompts que vous utilisez
3. Se poser la question de la représentativité du data du modèle que j'utilise
4. Diversifier les outils que vous utilisez et comparez les contenus produits (aide à VS à la place de)
5. Demander des comptes sur les algorithmes utilisés par votre employeur, vos assureurs, banques, l'Etat, les sites de Matching etc. => CNIL
6. Demander suivis des metrics, techniques utilisés pour gérer les biais dans l'entraînement, la mise en production – compliance DSA
7. ! À vos données :-)

L'IA peut vous aider à identifier les représentations implicites, engagement, risque de stéréotype

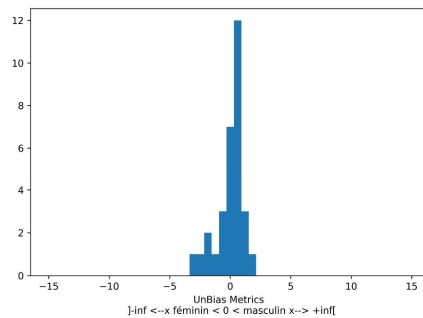


Les équipes Recherche & Développement d' Atos conçoivent et développent des systèmes et des solutions destinés à supporter les applications les plus exigeantes en ressources et performances, telles que le Calcul Haute Performance (HPC) , la Simulation Quantique , le Big Data et l' Intelligence Artificielle .

Nous recrutons un(e) Ingénieur(e) applicatif R&D HPC pour rejoindre l'équipe BXI MPI .

Cette équipe effectue des travaux de recherche et de développement qui

Analysez Export .xls



- Assurer que vos textes interpellent autant l'audience féminine que masculine
- Eviter les faux pas sexistes et le risque réputationnel
- Apporter des KPI objectifs et solides sur vos efforts sur l'inclusivité des genres