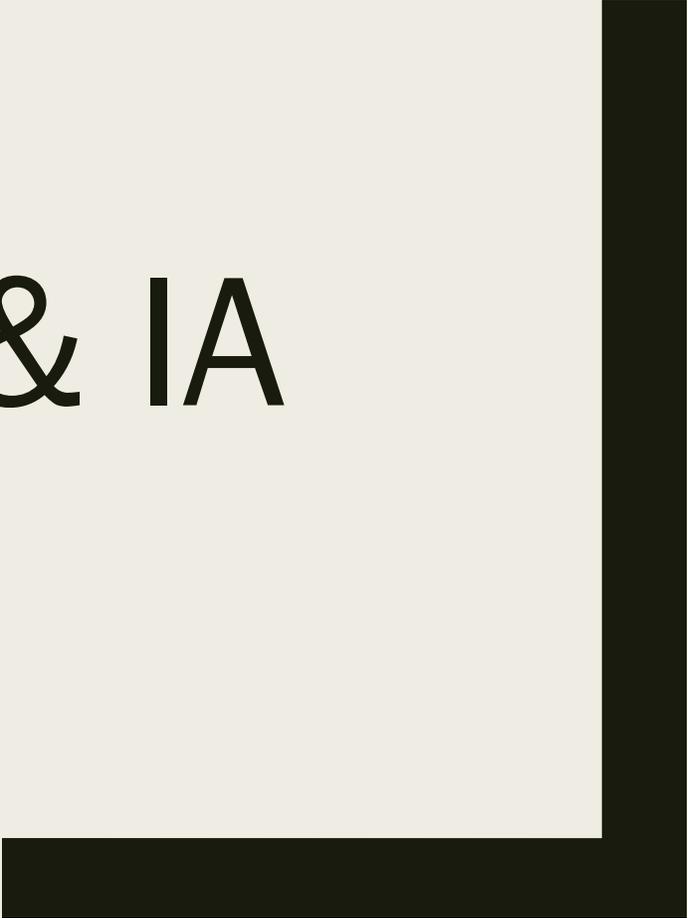


# GENRE, WEB & IA

24 mars 2025



# HARCÈLEMENT EN LIGNE, « CYBER-MISOGYNIE »

- « Haïr et railler les femmes en ligne : une revue de la littérature sur les manifestations de la cyber misogynie ». Genre en séries : cinéma, télévision, médias, no 11 (2020)
- Dès les années 1980-90's, des travaux soulignent l'ambiance très « masculine » qui régnait sur le minitel puis le Web [McCormick et Leonard, 2016].
  - Ex. *Gamergate (2014) : Depression Quest & Zoé Quinn*
- Les fonctionnalités techniques du web permettent de matérialiser et d'exprimer un certain nombre d'affects négatifs (et en particulier envers les groupes altérisés).
  - Des effets sens produits au travers d'une articulation de registres sémiotiques et
  - Un usage de l'ironie pour masquer le propos aux repérages automatiques.
  - Caractère ludique du harcèlement (*like, retweets*) [Salter, 2017]
- Règles et politiques de Twitter en matière de conduites haineuses :
  - De la difficulté de prendre en compte le contexte
  - Duplication de comptes
  - Cf. <https://help.twitter.com/fr/rules-and-policies/hateful-conduct-policy>

# HARCÈLEMENT EN LIGNE, « CYBER-MISOGYNIIE »

- Des cibles aux caractéristiques spécifiques :
  - des actrices, des femmes politiques et des professionnelles des médias manifestant un engagement féministe
  - et/ou racisée, manifestant un engagement anti-raciste (Lawson, 2018),
  - LGBTQI+
- L'identification d'une victime émissaire semble servir dans les deux cas la constitution d'un groupe prêt à des lynchages numériques (consolidation de liens entre les personnes d'une même communauté) [Garde-Hansen et Gorton, 2013 ; Julliard, 2018].



Le genre & l'écran a retweeté

**Victoire Tuillon**  
@vtuillon

J'ai récupéré l'accès à mon compte, après avoir supprimé le tweet considéré comme "haineux" par @TwitterFrance. Merci à @carolinedehaas et à toutes les personnes qui se sont indigné-es que les comptes de féministes continuent d'être suspendus pour des motifs injustes 🙄 (1/3)

**Caroline De Haas** @carolinedehaas · 31 janv.  
Encore un compte féministe suspendu par @TwitterFrance.

@vtuillon est suspendue 12h pour avoir tweeté ceci 📄

Rappel : 98% des personnes condamnées pour viol sont des hommes. 98%.  
Twitter, tu saoules.

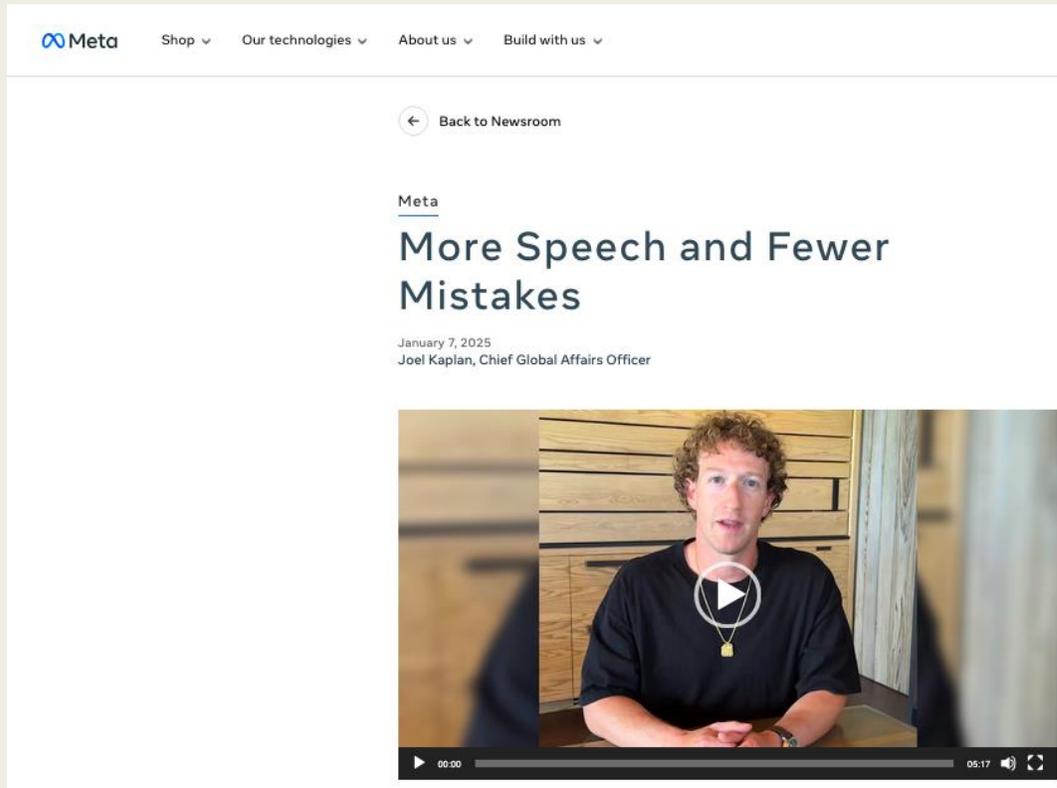
**Victoire Tuillon**  
@vtuillon

Je sais que c'est dur de regarder ce fait incontestable en face mais il faut le faire si on veut s'en sortir un jour : les personnes qui violent sont des hommes. De tout âge, toute classe, tout profil-- le seul point commun des violeurs, c'est d'être des hommes.

10:37 AM · 1 févr. 2021 · Twitter Web App

# RÉGULER, CENSURER, MODÉRER

## ACTUALITÉS DE LA MODÉRATION



<https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>

- Remplacer les fact-checkers par des « community notes »
  - Simplifier les politiques de modération (ex. : immigration et genre)
  - Nouveau fonctionnement de la modération (réserve les filtres aux contenus illégaux et aux « violations sévères », retrait des filtres pour les « violations de moindre gravité »/signalement)
  - Retour de la recommandation des contenus politiques/ « civic content »
  - Déménagement des équipes de modération au Texas (« less concerns about the biases of our team »)
- « Nous allons travailler avec le président pour repousser les attaques qui visent les entreprises américaines et qui poussent à davantage de censure ».

# RÉGULER, CENSURER, MODÉRER

## ACTUALITÉS DE LA MODÉRATION

**Mark Zuckerberg vire ses « vérificateurs », qui ont trop « censuré »**

Étienne Lombard 19 janvier 2025 Articles, Réseaux sociaux Facebook, Instagram, Mark Zuckerberg, Vérificateurs



Capture écran Mark Zuckerberg sur Facebook and too much censorship.

Mardi 7 janvier. Mark Zuckerberg diffuse une vidéo doublée d'un communiqué écrit, annonçant que sa société

Grille des programmes Podcasts Fictions Documentaires Savoirs Arts et Création

**Quitter X ? Que signifie la fin du "fact-checking" et de la modération sur les réseaux sociaux ?**

Publié le dimanche 19 janvier 2025

ÉCOUTER (41 min)



AI at Meta  
Together with the AI community, we are pushing the boundaries of what's possible through open science to create a more connected world.  
Science & Technology @ai.meta.com

Le groupe Meta, dirigé par Mark Zuckerberg ©Getty / Andrew Osa - SCOP Images - LightRocket

Réservé aux abonnés

**«Énergie masculine», fausse transparence, licenciements assumés... Mark Zuckerberg est-il un manager toxique ?**

Par Julien Da Sois  
Publié le 19 janvier à 18h00

Mark Zuckerberg Meta Management

Copier le lien



«Ce qui est fascinant chez Mark Zuckerberg, c'est comment ce tout jeune homme a été pris sous son aile par tous les plus grands de la Silicon Valley, de Steve Jobs à Bill Gates», analyse Stéphane Distinguin, fondateur et associé du cabinet EY Fabernovel. ANDREW CABALLERO-REYNOLDS / AFP

**DÉCRYPTAGE - En plein virage conservateur, le patron de Meta a de nouveau fait polémique la semaine dernière, en estimant notamment qu'«avoir une culture qui valorise un peu plus l'agressivité à ses mérites».**

# RÉGULER, CENSURER, MODÉRER

## RÉGULATION, LE CADRE LÉGISLATIF

- Mettre fin à la haine en ligne et au cyber-harcèlement [Dupré, D. et Carayol, V. (2020)] & lutter contre les *fake news* sur les RSN [Keita, A. (2020) & Manon Berriche (2023)]

—> Une panique morale pour la régulation du web et des réseaux sociaux [Dagnaud, Monique (2020)]

- La tentative de mise en place d'un cadre législatif à des fins politiques et électorales ? À l'échelle nationale (Loi Avia & Projet de loi SREN) et européenne (DSA).
- Peut-on penser une régulation des RSN au regard des multiples formes de modérations en ligne : par pays, par langue, par RSN, etc.

# RÉGULER, CENSURER, MODÉRER

RÉGULATION, LE CADRE LÉGISLATIF

## LOI AVIA : QUELQUES ÉLÉMENTS CLÉS



Cette loi oblige les RSN, les plateformes collaboratives et les moteurs de recherche à supprimer les contenus « haineux » et « pornographiques » dans un délai de 24h, sous peine de sanction pécuniaire.

Mais cette **contrainte temporelle de 24h** et les sanctions prévues font craindre aux juristes, ainsi qu'à un certain nombre d'acteurs du numérique (tels que le Conseil national du numérique ou la Quadrature du net) un **recours accru aux algorithmes** pour modérer rapidement les contenus qui seront jugés illicites.

# RÉGULER, CENSURER, MODÉRER

## REGULATION – CONTEXTE EUROPÉEN

### DIGITAL SERVICES ACT : QUELQUES ÉLÉMENTS CLÉS

- Création du statut de VLOP (+45 millions d'utilisateur-riche-s/mois dans l'UE)
- Enjeux de « transparence » : communication des entreprises autour de leurs algorithmes de modération, *Transparency Data Base*
- Système de signalement obligatoire
- Publicités ciblées pour mineur-e-s interdites sur toutes les plateformes
- Accès pour les chercheur-se-s (Arcom)

# RÉGULER, CENSURER, MODÉRER

## POLITIQUES DE MODÉRATION

# Il n'y a pas une, mais plusieurs politiques de modération...

The image shows a collage of three web pages related to social media moderation policies. On the left is the TikTok 'Règles Communautaires' page, which includes a sub-section for 'Règles et politiques'. In the center is the Twitter 'Règles et politiques de Twitter' page, detailing rules on research, safety, and automation. On the right is a 'Transparency Center' page with a navigation menu (Politiques, Mise en application, Fonctionnalités, Surveillance, Données) and a 'Centre d'assistance' link. The main content of the Transparency Center page is titled 'Justification de cette politique' and discusses content moderation for nudity and sexual activity, providing specific examples and justifications for these policies.

**TikTok**  
Comment TikTok soutient la communauté face à la crise du COVID-19  
**Règles Communautaires**  
Règles et politiques

**Règles et politiques de Twitter**  
À propos des règles et restrictions relatives à la recherche  
Politique en matière de comptes parodiques, de flux d'actualités, de commentaires et de fans (à > politique <)

**Transparency Center**  
Politiques Mise en application Fonctionnalités Surveillance Données  
Centre d'assistance

**Aujourd'hui**  
Version actuelle

30 sept. 2022  
29 avr. 2022  
24 déc. 2021  
25 nov. 2021  
Voir les plus anciennes

### Justification de cette politique

Nous limitons l'affichage de scènes de nudité et d'activités sexuelles, car certaines personnes au sein de notre audience peuvent être sensibles à ce type de contenu. En outre, nous supprimons par défaut les images de nature sexuelle pour empêcher le partage de contenus montrant des actes non consentis ou des mineurs. Les restrictions sur l'affichage d'activité sexuelle s'appliquent également au contenu créé numériquement, sauf si le contenu est publié à des fins pédagogiques, humoristiques ou satiriques.

Avec le temps, nous avons adapté nos règlements relatifs à la nudité. Nous comprenons que les contenus montrant des scènes de nudité peuvent être partagés pour diverses raisons, notamment dans le cadre d'une protestation, pour sensibiliser à une cause, ou à des fins pédagogiques ou sanitaires.

Par conséquent, nous acceptons les contenus respectant de tels motifs. Par exemple, alors que nous limitons certaines images de la poitrine féminine qui montrent le mamelon, nous autorisons d'autres types d'images, notamment celles illustrant des actes de protestation, des femmes défendant activement l'allaitement ou des cicatrices de mastectomie. Les images montrant ostensiblement des organes génitaux ou l'anus dans le contexte d'un accouchement, d'événements post-partum ou de situations relatives à la santé sont accompagnés d'un message d'avertissement afin d'informer qu'il s'agit d'un contenu potentiellement sensible. Nous autorisons également les photos de peintures, sculptures et autres œuvres d'art illustrant des personnages nus.

souvent en partageant des images et des vidéos dans la conversation. Parfois, ces médias peuvent représenter des sujets sensibles, notamment des contenus violents et pour adultes. Nous savons que certaines personnes ne souhaitent pas être exposées à des contenus sensibles. C'est pourquoi, si nous autorisons les utilisateurs à partager ce type de médias, nous aidons par ailleurs ceux qui ne souhaitent pas les voir à les éviter.

Il est interdit de publier des médias excessivement sanglants et de partager des contenus violents ou pour adultes dans des vidéos en direct, ou dans des images d'en-tête de profil, de bannière de Liste ou communautaire. Si vous partagez de tels contenus sur Twitter, vous devez [marquer votre compte comme sensible](#). Ainsi, vos images et vidéos sont placées derrière un interstitiel (message d'avertissement), dont l'utilisateur doit prendre connaissance avant l'affichage de vos médias. L'utilisation de cette fonctionnalité signifie que les personnes ne souhaitant pas voir de médias sensibles peuvent les éviter, ou prendre la décision d'y accéder en toute connaissance de cause. Nous limitons également l'accès à certains médias sensibles, tels que les contenus pour adultes, aux utilisateurs de moins de 18 ans ou à ceux qui n'ont pas indiqué de date de naissance sur leur profil. Découvrez [ici](#) plus d'informations sur les contenus auxquels l'accès est restreint en fonction de l'âge.

# RÉGULER, CENSURER, MODÉRER

## MÉCANISMES DE MODÉRATION

- Entre « digital labor » et modération automatisée
- Les « nettoyeurs du net » : une « activité invisible » dont le but est paradoxalement de désigner ce qui peut être visible sur les RSN (Smyrnaiois, N.& Marty, E. 2017).
- « The hidden labour of content moderation » : « le nettoyage du net revient de manière disproportionnée aux personnes noires, femmes, queer et handicapées » faisant déjà l'objet de discriminations [Nakamura, 2016] + « Travailleurs du clic » (Tubaro, Casilli, Coville, 2020) ; (Tubaro, Coville, Leludec, Casilli, 2023).
- La réflexion sur la modération du web est donc aussi une question à traiter sous le prisme du néolibéralisme [Sharif Mowlabocus, (2020). Le cas de Grindr]

# IA – BIAIS & DISCRIMINATIONS ALGORITHMIQUES

## DE QUELLE(S) IA PARLE-T-ON ?

- Un algorithme est un ensemble de règles et d'instructions écrites **en vue d'obtenir un résultat** [Télécom Paris Tech, 2019].
- Un type d'IA : le *machine learning* (apprentissage automatique à partir de jeux de données et compliqué à rétro-concevoir).
- Objectif : déconstruire l'apparente neutralité des technologies algorithmiques.
- En réalité, « un algorithme n'est qu'**une opinion intégrée dans des programmes** » [Cathy O'Neil, 2016]. L'IA, en tant que fait technique et **fait social**, apparaît donc, par nature, biaisée.

# IA – BIAIS & DISCRIMINATIONS ALGORITHMIQUES

## VERROUS À L'ÉTUDE DES BIAIS ALGORITHMIQUES

- Les enjeux de l'« explicabilité » : difficulté à rendre compréhensible un traitement algorithmique et la production d'un résultat par l'algorithme.
- L'image de la « boîte noire » algorithmique (Pascale, 2015).
- Une chasse-gardée des entreprises.

# IA – BIAIS & DISCRIMINATIONS ALGORITHMIQUES

PARTIR DE L'EXPÉRIENCE DU BIAIS ALGORITHMIQUE POUR COMPRENDRE L'IA

## QUELQUES DÉFINITIONS

### BIAIS ALGORITHMIQUE

J'appelle « biais algorithmique » la **transformation d'une observation générale** (souvent stéréotypée) ou statistique en une **condition algorithmique systématique** qui conduit, dans certains cas, à des discriminations technologiques contre certaines populations [Jean, 2019].

### DISCRIMINATION ALGORITHMIQUE

J'appelle « discrimination algorithmique » le **résultat de l'existence de biais** dans la conception et la mise en service d'algorithmes. En d'autres termes, la « discrimination algorithmique » constitue la **conséquence d'un recours à des technologies algorithmiques biaisées** et **l'expérience** du biais tel qu'il est vécu par les utilisateur·rice·s de ces services automatisés. La discrimination, quant à elle, se définit juridiquement comme un traitement défavorable d'un individu ou d'un groupe d'individus, fondé sur **un critère déterminé par la loi** (comme les enjeux de genre, de classe, de race, etc.).

# IA – BIAIS & DISCRIMINATIONS ALGORITHMIQUES

FOCUS : LES ALGORITHMES DE RECONNAISSANCE D'IMAGES À CARACTÈRE « PORNOGRAPHIQUES »



# IA – BIAIS & DISCRIMINATIONS ALGORITHMIQUES

PLUSIEURS BIAIS, PLUSIEURS DISCRIMINATIONS ALGORITHMIQUES

Biais cognitifs ou biais de société	Définition de variables en fonction de stéréotypes
Biais statistiques ou biais techniques	Qualité, collectif, tri et étalonnage des données (GIGO)
Biais économiques	Objectifs de rentabilité et manipulation volontaire de l'algorithme

# IA – BIAIS & DISCRIMINATIONS ALGORITHMIQUES

## LES ALGORITHMES SONT-ILS DISCRIMINANTS ? QUELQUES CAS EMBLÉMATIQUES

- Biais racistes : reconnaissance faciale, justice prédictive (Angwin et al., 2016) ; attribution de logements et de prêts bancaires
- Biais LGBTphobes : référencement du mot « lesbienne » dans les moteurs de recherche
- Biais sexistes : logiciel de recrutement d'Amazon



Montre l'image d'une femme française.

Image



Génère moi l'image d'une femme française.



Pour créer l'image, pouvez-vous me préciser quelques détails sur l'apparence ou le style que vous souhaitez ? Par exemple :

- Couleur et style des cheveux
- Type de tenue (moderne, classique, décontractée, chic parisien, etc.)
- Contexte ou décor (café parisien, jardin, intérieur, etc.)
- Tranche d'âge

Cela m'aidera à personnaliser l'image au mieux !

Génère moi l'image d'une femme française.



Voici une image générée représentant une femme française au style chic et élégant dans un cadre parisien classique. Si vous souhaitez des modifications ou un autre type de représentation, n'hésitez pas à me le dire !



Cette conversation vous a-t-elle été utile jusqu'à présent ?



Montre moi l'image d'une femme anglaise.

Image



Montre moi l'image d'une femme anglaise.



Pour créer une image, pourriez-vous me donner plus de détails ? Par exemple, des éléments comme l'époque (moderne, victorienne, etc.), le style vestimentaire, la couleur des cheveux, ou encore le contexte (dans un jardin, dans une bibliothèque, etc.) ? Cela m'aidera à produire une image qui correspond à vos attentes.

Montre moi l'image d'une femme anglaise.



Voici l'image d'une femme anglaise dans un cadre de campagne britannique, moderne et élégant. Si vous souhaitez des ajustements, n'hésitez pas à le demander !





Montre moi l'image d'une femme marocaine.

Image



le vérifie

Montre moi l'image d'une femme marocaine



Pourriez-vous me donner plus de détails sur le type d'image que vous imaginez ? Par exemple, des détails sur les vêtements, l'expression, ou le cadre dans lequel elle se trouve. Cela m'aidera à créer une image qui correspond à vos attentes.

Montre moi l'image d'une femme marocaine



Voici l'image d'une femme marocaine portant une tenue traditionnelle. J'espère que cela correspond à ce que vous imaginez !



# IA – BIAIS & DISCRIMINATIONS ALGORITHMIQUES

## SPÉCIFICITÉS DE LA DISCRIMINATION ALGORITHMIQUE

- L'IA, en tant qu'objet social, doit faire l'objet d'études et d'un regard en SHS [Défenseur des droits, 2020]. **L'origine des discriminations (algorithmiques) se trouve en fait dans la société** ; mais par les biais, les plateformes et leurs algorithmes les reproduisent et les cristallisent.
- Quelques **caractéristiques** de la discrimination algorithmique :
- Une discrimination **automatique**
- Une discrimination **présentée comme un glitch/bug** technique par les entreprises et donc enrobée d'une immunité (qui tend à prendre fin avec le AI ACT, 2024)
- Une discrimination **opaque** : on ignore souvent en être victime (secret professionnel et exemple du shadowban de Tiktok : « insidious form of censorship ») [ASPI, 2020]

# IA – BIAIS & DISCRIMINATIONS ALGORITHMIQUES

## SPÉCIFICITÉS DE LA DISCRIMINATION ALGORITHMIQUE

- Prendre l'IA comme **objet scientifique d'un point de vue SHS** permet d'étudier à la fois les **représentations du réel**, mais aussi les formes insidieuses de discrimination systémique. Principe de la « **simulation numérique** » : passer par le virtuel pour mieux comprendre le réel [Jean, A. (2019) De l'autre côté de la machine] ou « **a black mirror to society** » [Benjamin, R. (2019) Race after technology]

