“ Discriminations algorithmiques ” ? La modération des réseaux sociaux numériques au prisme de la censure

Thibault Grison

« Discriminations algorithmiques » ?
La modération des réseaux sociaux numériques au prisme de la censure

"Algorithmic Discriminations"?
Digital social networks and content moderation in the light of censorship

Thibault Grison

Doctorant en Sciences de l'Information et de la Communication
GRIPIC (Celsa, Sorbonne-Université)
SCAI (Sorbonne Center for Artificial Intelligence)
thibault.grison@sorbonne-universite.fr

Mots-clés : Modération ; Réseaux sociaux ; Algorithmes ; LGBT ; Méthodes numériques

Keywords: Content moderation; Digital social networks; Algorithms; LGBT; Computer science methods

Résumé : Cette communication prend pour objet la construction des représentations et identités de genre dans le déploiement des algorithmes de recommandation et de modération au sein des réseaux sociaux numériques (RSN). La communication propose donc de s'intéresser à la façon dont les dispositifs de modération et de recommandation des RSN produisent et conditionnent des régimes d'(in)visibilité pour les communautés sexuelles et de genre, et impactent la fabrique sémiotique de la sexualité en ligne.

Abstract: This paper focuses on the construction of gender representations and identities in the deployment of content recommendation and moderation algorithms within digital social networks (DSNs). How do content moderation and recommendation systems of DSNs produce and condition regimes of (in)visibility for sexual and gendered communities? To what extent does that impact the semiotic construction of sexuality online?

« Discriminations algorithmiques » ? La modération des réseaux sociaux numériques au prisme de la censure

Thibault Grison

Ces dix dernières années, les entreprises détentrices des réseaux sociaux numériques (RSN) ont beaucoup investi dans le recours aux algorithmes pour modérer plus efficacement et plus rapidement les contenus identifiés comme illicites[1]. Or, force est de constater que la délégation de la modération à l'IA semble aller de pair avec la discrimination à l'encontre des minorités sexuelles, de genre et racisées (Grison, Julliard, 2021). Censure[2], invisibilisation, déréférencement, shadowban[3] de contenus et de profils, etc. : la modération automatisée des contenus - alors qu'elle est défendue par les entreprises comme un moyen plus efficace pour protéger en particulier les personnes victimes de la prolifération de contenus haineux en ligne (Gatewood et al., 2020 ; Netino, 2020) - contribue au renforcement et à la cristallisation des rapports de minoration dans les dispositifs médiatiques (Grison, Julliard, à paraître en 2022). A titre d'exemple, courant mai 2020, des comptes militants LGBT étaient suspendus simultanément sur plusieurs RSN tels que Twitter ou Facebook. Pour ces militant·e·s, ces suppressions résultaient de l'intensification du recours aux algorithmes pour la modération, dans le contexte de la préparation de la loi visant à lutter contre les « contenus haineux » sur Internet[4]. Depuis, ces cas de suppression injustifiés continuent de s'accumuler sur Twitter, TikTok, Instagram, YouTube, etc.

Dans quelle mesure les algorithmes de modération et de recommandation sont-ils producteurs de régimes de visibilité ? Comment affectent-ils l'expression sémio-discursive de la sexualité (ou plutôt l'expression d'une appartenance à des cultures ou des communautés sexuelles) ?

Nous nous intéressons, d'une part, à la catégorisation des dispositifs (Bonaccorsi, Julliard, 2010) de modération comme des « technologies de genre » (Lauretis, 1987) dont le

---

[1] https://www.numerique.gouv.fr/uploads/rapport-mission-regulation-reseaux-sociaux.pdf

[2] https://www.aides.org/communique/la-loi-avia-ne-nous-rendra-pas-moins-militants-es-aides-indignation-censure-lgbtqi-tds

[3] https://www.bbc.com/news/technology-54102575

[4] Dite loi « Avia », du nom de la rapporteure de la proposition de loi

propre est d'organiser et de hiérarchiser les régimes de visibilité dans les RSN. Nous nous intéresserons donc en particulier aux phénomènes d'invisibilisation des personnes LGBT et travailleur·euse·s du sexe (TDS) résultant du recours massif à des algorithmes pour la promotion de contenus et la modération de contenus jugés problématiques. Cette communication sera également l'occasion de présenter une méthodologie de recherche hybride en ce qu'elle conjugue approche située (Haraway, 1988) du terrain - plutôt héritée des *cultural studies* (Quemener, 2022) - et méthodes de collecte et d'analyse computationnelles des données. Enfin, la participation à ces doctorales nous offre l'opportunité de présenter des premiers résultats et pistes de réflexion pour répondre à la problématique que nous posons plus haut.

Les dispositifs de modération algorithmique : entre « boîte noire », biais et discriminations algorithmiques

La modération des réseaux numériques : état de l'art

Le travail de modération sur les RSN est à la fois réalisé par des humains - internautes usagers de la plateforme (Nakamura, 2015) et travailleur·se·s du clic employé·e·s afin de « nettoyer le web » de ces contenus non conformes aux règles de conduite fixées par plateformes (Smyrnaios, Marty, 2017 ; Tubaro, Casilli, Coville, 2020) - et, de manière automatisée sans que l'on connaisse avec certitude les modalités d'articulation de ces méthodes ainsi que la part de contenus traités exclusivement de manière automatique. Il est important de souligner la façon dont la recherche francophone et européenne s'est d'ailleurs largement saisie de cet enjeu des « travailleur·se·s du clic » en se concentrant en particulier sur les conditions de travail de ces modérateur·rice·s invisibles, paradoxalement chargés de désigner ce qui sera rendu visible ou non sur les plateformes (Roberts, 2019).

A cette incertitude sur l'origine de la décision de modération d'un contenu (automatisée ou humaine) doit également s'ajouter l'enjeu des signalements. En effet, les communautés LGBT, TDS, féministes ou antiracistes sont régulièrement victimes de raids de groupuscules d'extrême droite sur les RSN (Dupré, Carayol, 2020). Le principe de ces actions est donc de signaler massivement les contenus postés par les militant·e·s et internautes issus de ces groupes minorisés afin de déclencher une modération automatique et une censure immédiate

de ces derniers (Badouard, 2020). Alors, comment distinguer ce qui relève d'une attaque politique d'une modération abusive exclusivement induite par les entreprises ?

Les dispositifs algorithmiques : des « boîtes noires » et des biais

Depuis une quinzaine d'années une nouvelle vague de chercheur·euse·s en *cultural* et *media studies* se sont saisis de l'IA, auparavant chasse-gardée des sciences de l'ingénierie et de l'informatique pour le traiter comme « fait social » (Durkheim, 1985). Une série de travaux anglo-saxons se sont donc essayés à définir les algorithmes comme des « opinions intégrés dans un programme » (Roberts, 2019). Dans cette conception de l'IA, les algorithmes seraient avant tout le reflet des objectifs et idéologies de leurs concepteur·rice·s (Jean, 2019). Naturellement, l'enjeu de ces travaux s'est donc rapidement décalé sur les biais de ces dispositifs et l'impact de leur déploiement dans la société. Un champ de recherche visant à penser le rôle de l'IA dans la reproduction des rapports sociaux de race, de genre ou de classe s'est d'ailleurs développé au sein de l'université UCLA et, plus spécifiquement, du C2i2[5]. Ces approches considèrent l'IA comme des dispositifs de pouvoir qui (re)produisent des systèmes d'oppression et des dynamiques discriminantes lorsqu'ils sont déployés dans des objectifs de rentabilité économique et par des grands groupes industriels (Noble, 2018 ; Benjamin, 2019).

La discrimination algorithmique peut être définie comme le résultat de l'existence de biais dans la conception et la mise en service d'algorithmes. Il s'agit, en d'autres termes, de l'expérience du biais algorithmique tel qu'il est vécu par les utilisateur·rice·s de ces services automatisés. Aussi, nous appelons « 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). Il existe plusieurs types de biais qui interviennent à différents moments du cycle de vie d'un algorithme. Sans entrer dans les détails de notre typologie de ces biais (Grison, Julliard, 2021), les biais algorithmiques sont multiples et leur origine est parfois compliquée à déterminer. La cause de cette incertitude réside dans le manque de transparence concernant la conception de ces machines et la communication des entreprises qui assimilent les cas de discriminations algorithmiques à de simples *glitch* techniques et erreurs informatiques isolées (Amabile et al., 2020).

---

[5] *Center for Critical Internet Inquiry,* co-dirigé par les chercheuses Safiya Umoja Noble et Sarah T. Roberts.

Ainsi, l'IA est principalement évoquée en ces termes de « *black box* » (Pasquale, 2015) et d'« opacité algorithmique » (Jean, 2019 ; Défenseur des droits, 2020). Le fonctionnement des algorithmes apparait comme la chasse-gardée des ingénieurs en informatique, protégée par le « secret professionnel » des entreprises (O'Neil, 2016). Nombre d'études portant sur les biais algorithmiques s'achèvent ainsi sur cette frustration scientifique à ne pas pouvoir retracer l'origine et la cause des discriminations algorithmiques (Zuiderveen Borgesius, 2018). De cette frustration découle pourtant une question de recherche : comment travailler sur l'invisibilisation des communautés sexuelles et de genre en ligne alors que l'objet d'étude est lui-même rendu opaque au chercheur ?

Développer une approche hybride de la modération automatisée sur les RSN

Bricoler une méthodologie computationnelle qui prendrait en compte les savoirs situés produits par les militant·e·s

Pour pallier l'opacité algorithmique, notre approche fait le pari de prendre les cas de modération abusive comme point de départ à la réflexion. En effet, de par mon parcours personnel et de chercheur, de par ma fréquentation des milieux militants en ligne, j'ai moi-même éprouvé certains des effets de ces processus d'invisibilisation. Cette recherche constitue donc une opportunité de produire un savoir sur le déploiement des algorithmes dans le contexte de la modération des RSN depuis une posture « située » (Haraway, 1988) qui prendrait en compte les savoirs militants produits par la simple médiatisation et dénonciation des cas de censure abusive et les savoirs profanes déjà diffusés sur les RSN qui nous servent de sources d'inspiration pour ensuite déployer de nouveaux appareillages méthodologiques. L'idée de la démarche est donc de procéder d'une forme de « rétro-ingénierie » (Pailler, 2019 ; Bowker, Star, 1999) pour produire un savoir à rebours de la modération. Nous justifions l'emploi de cette notion habituellement employée dans le milieu de l'informatique et du *hacking* pour désigner notre approche :

« La rétro-ingénierie sert en général à comprendre comment fonctionne un appareil dont on ne dispose pas des plans de conception ou de montage. Elle consiste à observer en pratique les conséquences d'un démontage de l'appareil, à cumuler suffisamment d'essais et

d'erreurs, pour pouvoir déduire la façon dont les actions de l'appareil sont générées. » (Pailler, 2019 : 55).

En effet, sans avoir l'illusion d'accéder au *back-office* et au fonctionnement interne des algorithmes déployés dans la modération des RSN, l'idée est de retracer des *scenarii* de modération abusive à partir des cas sur lesquels nous travaillons et par le prisme d'une série de tests, tentatives de catégorisation des objets collectés et, à terme, par le déploiement de nos propres algorithmes de *machine learning*. L'idée de cette recherche en *reverse-engineering* est également de collecter un corpus de contenus postés sur les RSN à partir des mêmes méthodes algorithmiques utilisées par les entreprises détentrices des RSN, notamment par le ciblage par mot-clé pour catégoriser ce qui doit être modéré ou non (Kim, Wohn, Cha, 2022 ; Ryan et al., 2020). Ce faisant, la critique du recours à l'IA et de son fonctionnement est nourrie par une méthodologie nécessitant elle-même un recours à des algorithmes afin de constituer un gros corpus de données. Le fait de nourrir cet appareillage algorithmique d'une prise en compte des savoirs situés s'inscrit également dans toute une filiation de travaux dits quanti-qualitatifs en sciences de l'information et de la communication prenant la construction du genre comme objet d'étude (Julliard, 2018). Plus précisément, notre approche s'inspire d'une approche *queer* de la collecte et de l'analyse de corpus constitués de données renvoyant à l'identité sexuelle et de genre des populations étudiées (Guyan, 2022). Cette approche participante implique d'opérer des allers-retours entre regard et collecte quantitatives et situées. Il ne s'agit pas de découper les terrains et méthodes d'enquête séparément, mais plutôt de conjuguer ces approches tout au long de l'analyse. De cette approche computationnelle située découle alors la constitution progressive d'un corpus d'enquête en puzzle.

Articuler les corpus

A ce stade, je travaille en particulier sur Twitter et TikTok. Le choix de travailler sur Twitter s'explique principalement en raison du contexte scientifique des travaux portant sur les RSN en SIC. En effet, les quelques études déjà publiées portant sur la modération des réseaux sociaux, et en particulier sur la haine en ligne, regrettent la façon dont les chercheureuses se concentrent exclusivement sur Twitter - du fait de la possibilité d'accès facile à l'API de la plateforme et à la facilité de création d'un compte académique pour collecter massivement des données de la plateforme - (Amabile et al., 2020). Il y a donc une

nécessité à orienter les recherches sur le risque de discrimination algorithmique dans la modération des RSN sur des RSN comme TikTok, encore trop peu mobilisé dans la recherche et d'autant moins dans la recherche francophone. Alors même que, paradoxalement, mes explorations préliminaires - notamment par l'exploration de corpus médiatiques collectés *via* Europresse - laissent suggérer l'idée selon laquelle la modération serait particulièrement automatisée et discriminante sur TikTok.

Mon corpus est principalement constitué d'une collecte continue et automatisée des contenus postés sur les RSN. Je collecte ces contenus à l'aide d'un ingénieur en informatique[6] et au travers de mots-clés renvoyant à l'orientation sexuelle et de genre comme « gay », « lesbienne », « pd », « gouine », etc. Une fois cette collecte effectuée, *via* l'API rest de Twitter, je réalise la même requête pour voir si le tweet en question a été modéré ou non. Ensuite, la comparaison de ces deux corpus me permet de faire des hypothèses sur l'origine de la modération en fonction des délais, des termes utilisés, de la corrélation texte-image, etc. Pour TikTok, l'absence d'API rend compliqué ce type de collecte. Je réalise donc une approche ethnographique participante en suivant le parcours de recommandation du RSN et en réalisant des requêtes par mot-clé *via* le moteur de recherche intégré de TikTok. A l'aide d'un carnet de bord et d'interaction avec des militant·e·s, de captures d'écran et d'enregistrements de contenus dans ma base de données, je me concentre pour l'instant sur les mobilisations militantes qui fustigent directement le risque de censure abusive. Cette collecte est également appuyée par un travail constant de veille médiatique sur d'éventuelles polémiques anglo-saxonnes ou francophones relatives aux discriminations algorithmiques, d'une série d'entretiens en cours de construction avec des militant·e·s LGBT et TDS ayant été victimes de modération qu'iels auraient jugée abusive, et d'un travail en analyse de discours sur les documents de communication produits par les entreprises détentrices des RSN. L'éclatement de ce corpus en puzzle disparate s'est progressivement présenté comme une nécessité pour palier l'opacité de notre terrain. La prise en compte de fuite de données du code source des algorithmes de modération[7] a également été déterminante dans la constitution de notre corpus et, de la même manière que fonctionne une investigation journalistique, nous

[6] Plus précisément, dans le cadre de l'initiative CERES de la Faculté des Lettres de Sorbonne Université : http://www.ceres.sorbonne-universite.fr

[7] https://www.washingtonpost.com/video-games/2021/10/15/twitch-leak-do-not-ban-streamers-tyler1-ricegum/

avançons au gré de l'instabilité et de l'évolution constante des formes de modération des RSN et des savoirs les concernant.

La modération des RSN : un dispositif technologique et d'invisibilisation genré

Une modération à deux vitesses

Bien qu'il soit compliqué, à ce stade de la recherche doctorale, de réaliser une typologie définitive - qui ne le sera d'ailleurs probablement jamais - notre enquête nous permet néanmoins de produire des savoirs sur la façon dont les contenus sont modérés en ligne. Lorsqu'on parle de modération, on pense en particulier à la censure de contenus ou à la suspension de comptes. Or, une forme de modération plus insidieuse (Ryan et al., 2020) se situe dans l'expérience-même de consultation du RSN par l'utilisateur·rice. En effet, tous les processus de recommandation personnalisés participent en réalité déjà à hiérarchiser certains types de contenus par rapport à d'autres. Sur TikTok en particulier, le référencement des contenus, ou plutôt le déréfencement, est devenu une stratégie employée par les modérateur·rice·s de contenus pour invisibiliser certains discours sans que les producteur·rice·s de ces contenus n'en soient avertis et puissent formuler un recours. On parle dans ce cas de *shadowban*. L'invisibilisation de certaines publications est donc due à la fois aux algorithmes de modération qui suppriment certains contenus des RSN, et aux algorithmes de recommandation qui hiérarchisent la visibilité de ceux-ci (Grison, Julliard, 2022).

Notons d'ailleurs la façon la mobilisation des militant·e·s contre ces processus automatisés de modération et de recommandation permet non seulement de rendre visible le risque du recours accru aux algorithmes dans la modération des RSN, mais participe aussi de l'évolution des phénomènes de modération. Dans notre travail nous avons pu observer la façon dont les entreprises détentrices des RSN réagissaient ou passaient sous silence les cas de censure abusive dénoncés par les militants et militantes. Si Twitter et TikTok sont parfois revenus sur les cas de modération abusive, c'est souvent en raison du bruit médiatique qui entouraient ces controverses (Gatewood et al., 2020). En effet, nous avons pu constater que la correction de cette censure abusive et leur justification étaient dépendantes du poids de la controverse. Il y a par ailleurs un décalage entre les communications officielles des

entreprises - dans leur charte de modération ou rapports de transparence - et les améliorations effectives du fonctionnement de la modération. Le paradoxe entre le déploiement de nouvelles stratégies de référencement conduisant à davantage de censure pour les minorités sexuelles et de genre et l'apparente prise en compte des savoirs et pratiques militantes est donc un terrain que nous aimerions davantage explorer à l'avenir.

Le traitement algorithmique des minorités sexuelles et de genre

Une des causes principales de la censure abusive est l'assimilation des communautés LGBT à des pratiques exclusivement sexuelles, pornographiques ou ostentatoires qui conduisent les algorithmes à assimiler l'expression de l'identité de genre à du contenu illicite. Il convient de rappeler que la militance LGBT et TDS s'est constituée au travers, notamment, de réappropriation d'insultes (« gouine », « pédé », « queer », « pute », etc.)[8], or les algorithmes ne parviennent pas à distinguer entre les différents usages possibles de ces mots (visée stigmatisante, affirmation identitaire, etc.). Les publications des militant·e·s comportant l'un de ces mots sont alors généralement jugées « non-conformes aux standards de la communauté » et supprimées. Par ailleurs, il convient de rappeler que les personnes LGBT sont constamment ramenées à leurs pratiques et cultures sexuelles, lesquelles sont généralement reléguées aux marges de l'espace public (Berlant, Warner, 2018). Enfin, notre collecte a permis de recenser la façon dont les internautes gay en particulier se constituaient en communauté de pratiques sémio-discursives par l'usage de références à des contenus à caractère explicitement sexuels. Les communautés LGBT en ligne existent, en effet, beaucoup à travers l'expression de contenus à caractère sexuel (mention d'expériences sexuelles par le biais de l'humour, de récits d'agressions sexuelles pour alerter sur les oppressions dont ils sont trop régulièrement les victimes, consommation de contenus pornographiques homosexuelles sur Twitter et enfin, présence de beaucoup de TDS également LGBT qui passent par les RSN pour générer des revenus en ligne en renvoyant vers leur Onlyfans ou autre plateforme monétisée). Cette dernière réflexion nous permet donc à la fois d'interroger la façon dont les personnes LGBT sont, dans ces nouveaux espaces de publicité, encore invisibilisées, et selon des modalités propres aux RSN et ainsi d'étudier comment les algorithmes reproduisent des biais qu'ils cristallisent en ligne, d'une part. Mais le dernier

[8] https://www.aides.org/communique/la-loi-avia-ne-nous-rendra-pas-moins-militants-es-aides-indignation-censure-lgbtqi-tds

point que nous venons de mentionner permet également de voir dans quelle mesure les communautés *queer* marginalisées et considérées socialement comme des cultures subalternes (Chatterjee, 2018 ; Foucault, 2014) s'écartent des parcours pré-construits par les dispositifs des RSN pour en détourner les usages et se constituer en communautés *via* la mise en visibilité de contre-visualités (Grison, Julliard, 2022 à partir de Mirzoeff, 2011 et Boidy, Tagliavia, 2018) quitte à alterner entre rapports de défiance et de stratégies de contournement pour exister en dépit du ciblage algorithmique. Dans les prochains mois de notre travail doctoral, nous aimerions explorer ce dernier point, notamment en le mettant en lien avec les politiques de modération de chaque RSN.

Bibliographie

Amabile, A., Lenoir, T., Perelmuter, T., & Thodoroff, B. (2020). *Algorithmes: Contrôle des biais S.V.P*. Institut Montaigne.

Badouard, R. (2020). *Les nouvelles lois du Web: Modération et censure*. Le Seuil.

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new Jim code*. Polity.

Berlant, L., & Warner, M. (2018). Sexe en public. *Questions de communication*, *33*. https://doi.org/10.4000/questionsdecommunication.12204

Bertail, P., Bounie, D., Clémençon, S., & Waelbroeck, P. (2019). *Algorithmes: Biais, discrimination et équité*. Télécom Paris Tech; Fondation Abeona; Institut Mines-Télécom; Institut Carnot;

Boidy, M., & Martinez Tagliavia, F. (2018). *Visions et visualités. Philosophie politique et culture visuelle,* Poli éditions.

Bonaccorsi, J., & Julliard, V. (s. d.). Dispositifs de communication numériques et médiation du politique. Le cas du site web d'Ideal-Eu. In M. Aghababaie, A. Bonjour, A. Clerc, & G. Rauscher, *Usages et enjeux des dispositifs de médiation*.

Bowker, G. C., & Star, S. L. (2008). *Sorting things out: Classification and its consequences*. MIT Press.

Défenseur des droits. (2020). *Algorithmes: Prévenir l'automatisation des discriminations* (p. 10). Défenseur des droits; CNIL.

Dupré, D., & Carayol, V. (2020). 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*, *11*.

Durkheim, É. (1985). *Les règles de la méthode sociologique* (J.-M. Berthelot & L. Mucchielli, Éds.; Nouvelle édition). Éditions Flammarion.

Foucault, M. (2014). *Histoire de la sexualité. 1: La volonté de savoir*. Editions Gallimard.

Gatewood, C., Guerin, C., Birdwell, J., Boyer, I., & Fourel, Z. (2020). *Cartographie de la Haine en Ligne. Tour d'horizon du discours haineux en France*. ISD.

Grison, T., & Julliard, V. (2021). Les enjeux de la modération automatisée sur les réseaux sociaux numériques: Les mobilisations LGBT contre la loi Avia. *Communication, technologies et développement*, *10*. https://doi.org/10.4000/ctd.6049

Grison, T., & Julliard, V. (2022). L'IA dans la modération: Censure des contenus LGBT sur TikTok et Twitter et terreau de nouvelles formes sémiotiques. In *IA, Culture et Médias*. Presses de l'Université Laval.

Guyan, K. (2022). *Queer data: Using gender, sex and sexuality data for action,* Bloomsbury Academic.

Haraway, D. (1988). Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies*, *14*(3), 575. https://doi.org/10.2307/3178066

Jean, A. (2019). *De l'autre côté de la machine: Voyage d'une scientifique au pays des algorithmes*. Editions de l'Observatoire.

Julliard, V. (2018). *La différence des sexes sur Twitter: Les conditions d'observabilité d'un engagement affectif et émotionnel*. [Mémoire d'habilitation à diriger des recherches]. Université Paris-Est Créteil.

Kim, J., Wohn, D. Y., & Cha, M. (2022). Understanding and identifying the use of emotes in toxic chat on Twitch. *Online Social Networks and Media*, *27*. https://doi.org/10.1016/j.osnem.2021.100180

Lambrecht, A., & Tucker, C. (2018). *Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads*. https://dx.doi.org/10.2139/ssrn.2852260

Lauretis, T. D. (1987). *Technologies of Gender: Essays on Theory, Film, and Fiction*. Indiana University Press.

Mirzoeff, N. (2011). *The right to look: A counterhistory of visuality*. Duke University Press.
Mitchell, W. J. T. (2014). *Que veulent les images?: Une critique de la culture visuelle* (M. Boidy, N. Cilins, & S. Roth, Trad.). Les Presses du réel.

Nakamura, L. (2015). The unwanted labour of social media: Women of colour call out culture as venture community management. *New formations: a journal of culture/theory/politics*, *86*, 106-112.

Netino. (2020). *Baromètre 2020 de la haine en ligne*. Respect Zone. https://netino.fr/ressources/Barometre_2020_Haine-en-ligne.pdf

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy* (First edition). Crown.

Pailler, F. (2019). *Les affects classifiés: Numérique et médiations sexuelles* [Thèse de doctorat]. Université de Nantes. https://hal.archives-ouvertes.fr/tel-03236631/document

Pasquale, F. (2015). *Black box society: Les algorithmes secrets qui contrôlent l'économie et l'information*. Fyp éditions.

Quemener, N. (2022). *Une approche Cultural Studies en SIC* [Mémoire d'habilitation à diriger des recherches], Sciences Po Lyon.

Roberts, S. T. (2019). *Behind the screen: Content moderation in the shadows of social media*. Yale University Press.

Ryan, F., Fritz, A., & Impiombato, D. (2020). *Tiktok and WeChat: Curating and controlling global information flows* (N° 37). ASPI; International Cyber Policy Centre.

Smyrnaios, N., & Marty, E. (2017). Profession «nettoyeur du net»: De la modération des commentaires sur les sites d'information français. *Réseaux*, *n° 205*(5), 57. https://doi.org/10.3917/res.205.0057

Tubaro, P., Casilli, A. A., & Coville, M. (2020). The trainer, the verifier, the imitator: Three ways in which human platform workers support artificial intelligence. *Big Data & Society*, *7*(1), 205395172091977. https://doi.org/10.1177/2053951720919776

Žliobaitė, I., & Custers, B. (2016). Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artificial Intelligence and Law*, *24*(2), 183-201. https://doi.org/10.1007/s10506-016-9182-5

Zuiderveen Borgesius, F. (2018). *Discrimination, intelligence artificielle et décisions algorithmiques*. Conseil de l'Europe.