Préface

Antonio A. Casilli

Antonio A. CASILLI (2020), « Préface », Sarah T. Roberts, *Derrière les écrans: les nettoyeurs du Web à l'ombre des réseaux sociaux*, Paris, La Découverte.

Préface

La traduction du livre de Sarah Roberts représente un moment important pour le débat public français autour du numérique et du rôle de la « modération commerciale de contenu » sur Internet. Cette notion, que la chercheuse a été la première à employer, englobe une pluralité de pratiques de tri, de signalement et de suppression de l'information en ligne. Les modérateurs sont parfois des administrateurs qui sélectionnent l'information « de qualité » des sites web, parfois des médiateurs de forums internet qui tempèrent les conversations entre partisans trop zélés d'opinions incompatibles, parfois des travailleurs précaires payés à la tâche pour étiqueter des images comme pornographiques ou violentes, et parfois encore des salariés de centres d'appel proposant des services en ligne à d'autres entreprises qui externalisent auprès d'eux le filtrage nécessaire des contenus hébergés sur leurs sites.

Dans cet ouvrage, ainsi que dans le monumental film documentaire *The Cleaners*, dont Sarah Roberts a été la conseillère scientifique[1], les modérateurs sont également qualifiés de « nettoyeurs du web » opérant dans les ténèbres de l'infrastructure numérique. Tantôt cette obscurité est un effet de la délocalisation qui rend nécessaire la synchronisation des rythmes de travail des modérateurs basés aux Philippines avec les fuseaux horaires des usagers modérés aux États-Unis ou en Europe, tantôt elle reflète métaphoriquement la nature des contenus filtrés. En confesseurs d'un nouveau genre, par écrans interposés, les modérateurs prennent alors en charge la noirceur du monde. Cette dimension est au cœur de l'argument de Roberts : une approche humaniste de l'information ne saurait ignorer le travail humain nécessaire pour la façonner et la mettre en circulation, et la pénibilité de cette tâche doit en conséquence être reconnue et compensée.

Ce livre représente l'aboutissement d'une enquête entreprise il y a plus d'une décennie. Il est aussi la manifestation concrète de l'impulsion que son autrice a su donner à un nouveau domaine de recherche de part et d'autre de l'Atlantique. Les travaux récents de Lisa Nakamura, Tarleton Gillespie, Camille Alloing et Julien Pierre, Lauren Huret, Nikos

[1] Hans BLOCK et Moritz RIESEWIECK, *The Cleaners (Im Schatten der Netzwelt)*, film documentaire, 2018.

Smyrnaios et Emmanuel Marty[2] s'inspirent à bien des égards de cette étude pionnière. Son importance tient également à l'analyse de la transition sociale et technologique qu'elle décrit, entre un internet des origines, dont les promoteurs, en quête de liberté et d'autonomie, entendaient repousser la « frontière électronique » et où la modération était assurée par les utilisateurs eux-mêmes, et un environnement dominé par les grands médias numériques, où des équipes spécialisées s'adonnent au filtrage des contenus pour standardiser les plateformes et les applications contemporaines sur le modèle des centres commerciaux et des complexes résidentiels gardés.

La dialectique entre auto-modération (*self-moderation*) et modération assurée par des agents préposés (*staff moderation*) est un phénomène documenté[3]. Le Web des origines aspirait à être une confédération de territoires autonomes édictant leurs propres normes. Celui qui a fait surface depuis les années 2000, au contraire, est une galaxie de systèmes fermés dont le souci principal est de se protéger de leurs propres utilisateurs. En particulier, aux États-Unis, la section 230 du *Communications Decency Act*, qui garantit l'immunité des hébergeurs de contenus en ligne contre d'éventuelles poursuites pénales liées à la diffusion d'images, de vidéos ou de textes publiés par les utilisateurs de leurs services, a de fait investi les grandes plateformes de la décennie suivante du pouvoir discrétionnaire de déterminer les contenus autorisés et ceux qu'il convient de soustraire à la vue des usagers.

Toutefois, la commercialisation de la modération n'a pas été impulsée uniquement par les grandes plateformes. Au contraire, comme elles l'ont fait dans d'autres contextes, ces dernières auraient volontiers continué de profiter du travail de « bénévoles » non rémunérés. Si l'activité s'est progressivement professionnalisée, ce processus est largement dû aux luttes et aux revendications des modérateurs eux-mêmes. En prenant conscience, il y a vingt ans, que leur travail valait

---

[2] Lisa NAKAMURA, « The unwanted labour of social media : Women of color call out culture as venture community management », *New Formations*, nº 86, 2015, p. 106-112 ; Tarleton GILLESPIE, *Custodians of the Internet : Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*, Yale University Press, New Haven, 2018 ; Camille ALLOING et Julien PIERRE, *Le Web affectif. Une économie numérique des émotions*, INA Éditions, Bry-sur-Marne, 2017 ; Lauren HURET, *Praying for my Haters*, Les Presses du réel, Dijon, 2019 ; Nikos SMYRNAIOS et Emmanuel MARTY, « Profession "nettoyeur du net". De la modération des commentaires sur les sites d'information français », *Réseaux*, nº 205, 2017, p. 57-90.

[3] Axel BRUNS, « Social media : Tools for user-generated content », *Social Media, vol. 2 : User Engagement Strategies*, 2009, <http://snurb.info/files/Social%20Media%20Report%20Volume%202%20-%20User%20Engagement%20Strategies.pdf>.

beaucoup plus qu'un accès gratuit au service, les 14 000 usagers qui animaient, administraient et, par-dessus tout, modéraient les forums de discussion en ligne des communautés d'AOL – ancêtre des grandes plateformes sociales du XXIe siècle – ont engagé un recours collectif qui s'est soldé en 2010 par un règlement à l'amiable de 15 millions de dollars[4]. Ce précédent a ouvert la voie aux grandes actions en justice des années suivantes, grâce auxquelles les risques psychosociaux associés au travail de modération ont enfin été reconnus[5].

Pour les utilisateurs des services en ligne, l'essor de la modération commerciale de contenu met fin au rêve d'un internet conçu comme un vecteur de libération de la parole et des subjectivités, les internautes se chargeant eux-mêmes de faire le tri. Désormais, les plateformes ne s'embarrassent guère de scrupules démocratiques pour réglementer les pratiques de leurs usagers, principalement pour protéger leur image de marque des comportements nocifs d'une partie d'entre eux et préserver leur entreprise des retombées légales que les écarts de conduite seraient susceptibles d'entraîner.

La demande de modération découle de nécessités internes à l'industrie numérique, mais elle émane aussi d'acteurs extérieurs. Les gouvernements exigent depuis la moitié des années 2010 une collaboration active des réseaux sociaux dans la lutte contre le terrorisme international ou contre des troubles à l'ordre public. De leur côté, certains annonceurs menacent depuis la fin de la décennie de boycotter les plateformes qui ne s'impliquent pas suffisamment dans la lutte pour les causes que leurs marques seraient supposées incarner. Les priorités parfois discordantes de ces acteurs provoquent une cacophonie morale et politique. Volatiles et arbitraires, les directives que reçoivent les modérateurs relèvent ainsi du tâtonnement, débouchant tour à tour sur la suppression de vidéos racistes et intolérantes en France, de profils de militants antifascistes aux États-Unis, de pages considérées comme blasphématoires au Pakistan, des fils de discussion d'opposants politiques en Chine. Les standards sont continuellement redéfinis pour prendre en compte les lois des pays dans lesquels les plateformes s'implantent, les enjeux de société émergents, les préconisations des services juridiques ou commerciaux.

L'organisation des fonctions de modération épouse les modèles

---

[4] Voir à ce sujet Antonio A. CASILLI, *En attendant les robots. Enquête sur le travail du clic*, Seuil, Paris, 2019, p. 178-183, 207-212.

[5] Sam LEVIN, « Moderators who had to view child abuse content sue Microsoft, claiming PTSD », *The Guardian*, 12 janvier 2017, <https://www.theguardian.com/technology/2017/jan/11/microsoft-employees-child-abuse-lawsuit-ptsd> ; Casey NEWTON, « Facebook will pay $52 million in settlement with moderators who developed PTSD on the job », *The Verge*, 12 mai 2020, <https://www.theverge.com/2020/5/12/21255870/facebook-content-moderator-settlement-scola-ptsd-mental-health>.

économiques du secteur de la tech et les mécanismes de construction d'identités collectives au sein des équipes dédiées à cette activité. Ses modalités relèvent donc de jeux d'échelle. Dans les milieux numériques souvent spécialisés et de petite taille hérités du siècle passé, modérer revenait principalement à négocier les prises de parole et les dynamiques réputationnelles en sélectionnant les participants et en harmonisant leurs registres. C'était une « modération communautaire », c'est-à-dire réalisée par un collectif impliqué, dont l'objectif était de fédérer des publics et de cultiver une identité propre, en évinçant les internautes qui ne partageaient pas l'esprit et les idéaux du groupe. La récupération marchande des sites communautaires par les grandes plateformes généralistes du XXIe siècle et leur revente à des annonceurs en tant que cibles publicitaires favorisent au contraire les variations d'échelle des publics. La « scalabilité » des services proposés permet l'extension potentiellement infinie de ces publics au gré de l'inclusion de segments de marché toujours nouveaux. D'où l'importance de faire coexister des points de vue divergents. Mais cela ne conduit pas à l'épanouissement d'un espace public, au sens idéal où l'entendait Jürgen Habermas, dans lequel il serait possible de confronter les extrémités opposées du spectre politique, culturel ou social. Au contraire, la logique des plateformes numériques consiste à maintenir séparées les différentes sphères, pour qu'elles ne communiquent pas. Cette compartimentation a parfois été présentée comme un effet des « bulles de filtrage » engendrées par le fonctionnement algorithmique des plateformes[6], ou comme une conséquence directe de la tendance des utilisateurs à s'enfermer dans des « chambres d'écho »[7].

Quoi qu'il en soit, le travail des modérateurs sert justement à consolider l'impression d'ajustement des contenus produits par les usagers, laquelle rend à son tour possible le fonctionnement des algorithmes. Leur activité consiste à cacher de manière sélective des pans entiers de ce qui est partagé en fonction de préférences étalonnées. Le tarissement progressif de la curiosité des usagers est l'un des risques associés à ce cloisonnement. Sous le prétexte de séparer les messages utiles du bruit, la modération peut avoir pour conséquence un appauvrissement du panorama cognitif des utilisateurs, qui cessent progressivement d'être surpris du manque de variété au sein de leur plateforme. Pourquoi n'y a-t-il pas de contenus « adultes » (ni même de représentations artistiques desdits contenus) sur Facebook ? Parce que les modérateurs les ont filtrés. Combien de personnes s'en étonnent ? De moins en moins, car leurs attentes et leurs goûts se sont imperceptiblement modifiés au fil du tri opéré par les modérateurs.

---

6 Eli PARISER, *The Filter Bubble : What the Internet Is Hiding from You*, Penguin Press, New York, 2011.

7 Eytan BAKSHY, Solomon MESSING et Lada A. ADAMIC, « Exposure to ideologically diverse news and opinion on Facebook », *Science*, vol. 348, no 6239, 2015, p. 1130-1132.

Mais les petites mains de la modération ne se limitent pas à former les goûts des publics. Elles entraînent des algorithmes qui suggèrent ensuite des achats aux usagers, prennent des décisions sur les profils à suivre, sélectionnent les actualités à mettre en exergue. C'est ainsi que les travaux de Sarah Roberts prolongent et approfondissent les réflexions actuelles sur le *digital labor*. Les modérateurs qui travaillent au sein de sociétés du secteur numérique et ceux qui opèrent depuis des plateformes de micro-travail produisent une énorme quantité de métadonnées qui calibrent les modèles mathématiques et les « dressent » en temps réel à anticiper les choix et les jugements des usagers. Ils contribuent de cette manière à l'automatisation que les grandes plateformes vendent autant à leurs investisseurs qu'au décideurs politiques.

Il est difficile de déterminer à quel moment la modération de contenu s'arrête et l'entraînement des intelligences artificielles commence. Les travailleurs payés à la micro-tâche qui évaluent la pertinence des réponses de Google Search contribuent à améliorer l'algorithme de référencement du moteur de recherche mais en même temps filtrent les contenus malveillants, les infox ou les sites contenant des images et des messages trop violents et explicites. De même, les salariés payés à l'heure par des sous-traitants de YouTube doivent autant vérifier que les vidéos n'enfreignent pas le droit de la propriété intellectuelle ou les lois des pays dans lesquelles elles sont diffusées qu'améliorer l'algorithme de monétisation qui assortit ces vidéos aux annonces publicitaires qui les accompagnent.

En tant qu'il est sous-jacent au développement des intelligences artificielles, le travail des modérateurs est invisibilisé et euphémisé par les plateformes qui les recrutent sous diverses dénominations : ce sont des *raters* chez Google, des *reviewers* pour Facebook… Leurs fonctions sont parfois rendues méconnaissables par des intitulés de poste aussi ronflants que fantaisistes, tels que « analyste de données », « responsable de communautés », « expert en renseignement ». À l'instar de la technique consistant à rendre illisible pour un humain un programme informatique, ces procédés conduisent à une « obfuscation » du rôle véritable des modérateurs. Ce travestissement s'impose si l'on considère que c'est sur la capacité des sociétés de la tech à automatiser leurs processus que les investisseurs misent aujourd'hui. L'automation est vendue à l'opinion publique en tant que garantie de l'objectivité des critères de sélection de l'information.

Néanmoins, les grandes plateformes numériques peinent à tenir cette promesse. Leurs algorithmes sont tout sauf autonomes et continuent de nécessiter l'intervention des modérateurs. L'effort que Facebook a engagé depuis le milieu des années 2010 pour lutter contre la prolifération de la propagande politique et les fausses nouvelles représente un exemple historique de l'imprescriptibilité du travail de modération humaine. Quand, en 2016, il a été révélé que le système d'« actualités personnalisées » de la plateforme n'était pas entièrement

automatisé, mais assuré par une équipe opérationnelle externe[8], l'entreprise s'est empressée de renvoyer son équipe de modérateurs en se targuant de pouvoir la remplacer par un algorithme. Laissé sans supervision humaine, le système est vite devenu la proie de manipulateurs et de promoteurs de *fake news*[9]. La modération a été alors réintroduite pour vérifier les éléments factuels des actualités (*fact checking*), et à nouveau confiée à des sous-traitants externes. En Europe, des sociétés comme CCC, CPL Resources ou Accenture recrutent désormais de plus en plus de travailleurs précaires pour évaluer les contenus de Facebook[10].

On pourrait imaginer que ce travail d'accompagnement des décisions automatiques soit destiné à disparaître, une fois que tous les algorithmes auront appris à marcher, pour ainsi dire, sans les béquilles des tâcherons du clic. Mais, à en en juger par les déclarations du P-DG de la plateforme de Palo Alto, la pleine automation s'inscrit dans un horizon utopique toujours repoussé. Presque cinq ans après les premières révélations, face à la pandémie de Covid-19, Mark Zuckerberg admettait que « son système avait été affecté par l'absence de modération humaine[11] ». D'autres plateformes ont dû faire face au même type de problèmes. YouTube affichait depuis le début de la crise le message suivant : « IMPORTANT : à cause du Covid-19, nous allons conduire

---

[8] Michael NUNEZ, « Former Facebook workers : We Routinely suppressed conservative news », *Gizmodo*, 9 mai 2016, <https://gizmodo.com/former-facebook-workers-we-routinely-suppressed-conser-1775461006>.

[9] Sam THIELMAN, « Facebook fires trending team, and algorithm without humans goes crazy », *The Guardian*, 29 août 2016, <https://www.theguardian.com/technology/2016/aug/29/facebook-fires-trending-topics-team-algorithm>.

[10] « Locations », CCC, 2017, <https://www.yourccc.com/en/locations> ; Dennis GREEN, « Facebook content moderation firm asked on-site therapists to disclose counseling details with employees, according to report », *Business Insider*, 17 août 2019, <https://www.businessinsider.fr/us/facebook-contractor-accenture-forced-counselors-to-disclose-sessions-2019-8> ; Henri POULAIN et Julien GOETZ, « Traumas sans modération », *Invisibles. Les travailleurs du clic*, épisode 3, France.TV/Slash, 12 février 2020, <https://www.france.tv/slash/invisibles/saison-1/1274811-traumas-sans-moderation.html>.

[11] Scott NOVER, « Facebook labeled 50 million pieces of Covid-19 misinformation in april », *AdWeek*, 12 mai 2020, <https://www.adweek.com/brand-marketing/facebook-removed-50-million-pieces-of-covid-19-misinformation-in-april/>

moins de modération humaine pour protéger la santé de nos effectifs ». Twitter reconnaissait de son côté que, sans la modération humaine des travailleurs confinés, « ces systèmes automatiques manquent de contexte et de perspective[12] » et sont donc voués à se tromper.

Entre mars et avril 2020, les pays dans lesquels les modérateurs sont majoritairement installés (l'Espagne, l'Irlande, les Philippines) ont tous été concernés par des mesures de confinement très contraignantes. En raison de la nature sensible des contenus qu'ils traitent et des accords de confidentialité qu'ils signent, les membres de ces équipes ne sont presque jamais autorisés à travailler depuis chez eux (à la différences de leurs homologues opérant *via* des plateformes de micro-travail ou des agences). Ils sont pourtant indispensables au point que, au bout du premier mois de confinement, en plusieurs pays européens hébergeant des sites de modération, Facebook les désignait comme prioritaires pour réintégrer les bureaux – alors que les autres salariés continuaient à télétravailler[13].

Si l'on considère les multiples enjeux que soulève la modération de contenu, tels que la constitution de communautés numériques, la tension entre gratuité et logiques marchandes, le rôle du travail humain à l'heure de l'automation, cette activité ne saurait être réduite – comme c'est trop souvent le cas – à une simple question de censure que les États délèguent à des entreprises privées. Cela est en partie vrai pour le Vieux Continent, où la Commission européenne a longtemps tenté d'imposer des règlements comme celui de 2018 relatif à la prévention de la diffusion en ligne de contenus à caractère terroriste[14]. À son tour, le gouvernement français a fait voter en 2020 la loi Avia contre les messages haineux sur

---

[12] Vijaya GADDE et Matt DERELLA, « An update on our continuity strategy during COVID-19 », *Twitter Blog*, 1er avril 2020, <https://blog.twitter.com/en_us/topics/company/2020/An-update-on-our-continuity-strategy-during-COVID-19.html>.

[13] Taylor HATMAKER, « Facebook wants content reviewers back ASAP, slows return plan for most employees », *TechCrunch*, 16 avril 2020, <https://techcrunch.com/2020/04/16/facebook-content-moderators-wfh-when-will-facebook-employees-return/>.

[14] COMMISSION EUROPÉENNE, « Proposition de règlement du Parlement et du Conseil relatif à la prévention de la diffusion de contenus à caractère terroriste en ligne », Bruxelles, 12 septembre 2018, <https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=CELEX%3A52018PC0640>.

Internet, aussitôt annulée par le Conseil constitutionnel[15]. Les deux textes proposaient de faire appel à la modération supposément « automatisée » des grandes plateformes numériques pour retirer de la circulation dans un délai d'une heure n'importe quel contenu signalé comme illicite – sans l'autorisation préalable d'un juge. Certaines associations de défense des libertés publiques craignent à juste titre que ces lois aient accessoirement pour effet de consolider la place hégémonique des géants du Web, les seuls disposant d'armées de modérateurs actifs 24 heures sur 24, dont la mobilisation serait nécessaire afin de respecter des obligations aussi strictes[16].

La modération s'est désormais transformée en une sorte de réflexe pavlovien qu'États et plateformes semblent avoir pleinement intégré. Les pouvoirs publics entendent s'en servir pour contrôler à la source la circulation de contenus problématiques. Les plateformes veulent l'employer comme outils de gouvernance interne et comme instrument de rapprochement avec les gouvernements. Quelle que soit l'issue des évolutions en cours, elles renvoient à une utopie bien lointaine le célèbre adage *information wants to be free*, qu'il était loisible d'interpréter comme une aspiration conjointe à la gratuité et à la liberté. L'étude de la modération commerciale nous rappelle qu'un travail ni gratuit ni entièrement volontaire sera toujours nécessaire dans un monde dominé par des plateformes marchandes.

Antonio A. Casilli, juillet 2020.

---

[15] Conseil d'État, « Avis sur la proposition de loi visant à lutter contre la haine sur Internet », Assemblée générale, Séance du jeudi 16 mai 2020, Section de l'intérieur n° 397368, <https://www.conseil-etat.fr/ressources/avis-aux-pouvoirs-publics/derniers-avis-publies/avis-sur-la-proposition-de-loi-visant-a-lutter-contre-la-haine-sur-internet>.

[16] La Quadrature du Net, « Coup d'État sur la "Loi Haine" », *La Quadrature du Net*, 22 janvier 2020, <https://www.laquadrature.net/2020/01/22/coup-detat-sur-la-loi-haine/>.