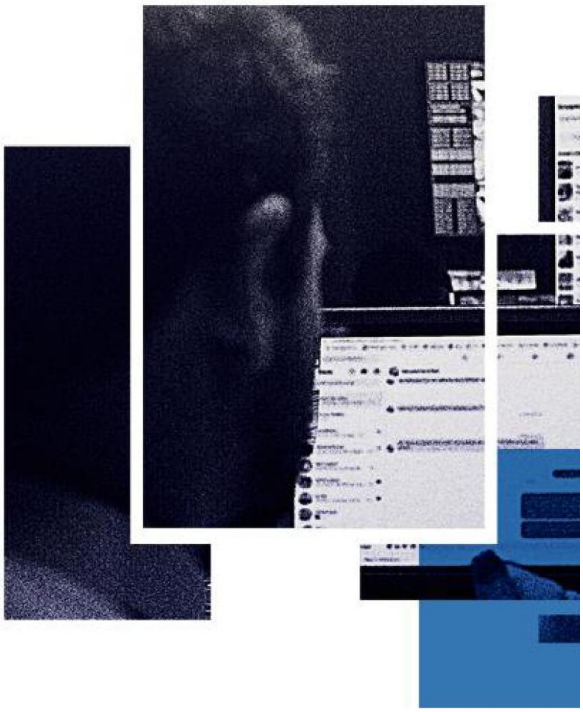
Le Monde
MERCREDI 27 OCTOBRE 2021

FACEBOOK FILES

Chez Facebook, un algorithme qui échappe au contrôle de ses créateurs

Dans des documents internes de l'entreprise, ses ingénieurs avouent leur incompréhension face à un code informatique aux effets imprévus, qui fait du réseau social une machine complexe, très difficile à maîtriser



C'est peut-être le principal sentiment qui émerge à la lecture des « Facebook Files ». Parmi ces milliers de pages de documents internes à Facebook, récupérés par Frances Haugen, une ancienne employée, et transmis par une source parlementaire américaine à plusieurs médias, dont *Le Monde*, de nombreux passages semblent indiquer que Facebook ne comprend plus, ou mal, ce que font ses propres algorithmes. Et que son réseau social est devenu une machine difficile à contrôler.

C'est notamment le cas pour un algorithme crucial, chargé de « classer » les messages qui s'affichent dans le fil d'actualité des utilisateurs : Facebook utilise une multitude de signaux, des plus simples – le nombre de personnes abonnées à une page – aux plus complexes – l'intérêt que les « amis » d'un utilisateur ont manifesté pour un sujet – afin d'attribuer un « score » à un message. Plus ce score est élevé, plus il a des chances d'apparaître dans le fil d'actualités. Or, avec le temps et l'accumulation de nouveaux signaux ajoutés par les ingénieurs du réseau social, le « score » moyen d'un message a explosé. Dans un document non daté, un analyste de Facebook a procédé à quelques calculs, et constate que pour certains contenus, le score *« peut dépasser 1 milliard »*. Ce qui a pour conséquence très directe de rendre de nombreux outils de modération inopérants.

Ces derniers réduisent de 20 %, 30 % ou 50 % le score de certains messages problématiques, afin d'éviter qu'ils ne soient trop diffusés. Mais pour les contenus les mieux notés, le score est tellement élevé que le diviser par deux ne les empêche pas de continuer à s'afficher. *« Certains de ces contenus resteraient en tête même si on appliquait une baisse de leur score de 90 % »*, regrette l'auteur du document.

PAS DE « VISION UNIFIÉE »

Ce problème n'est absolument pas le résultat d'une politique intentionnellement mise en place, qui considérerait que certains messages devraient être immunisés contre les outils de modération automatiques. C'est simplement l'un des très nombreux effets de bords provoqués par les centaines de modifications des algorithmes de Facebook, au fil des ans, dont les propres ingénieurs du réseau social semblent ne pas pouvoir anticiper toutes les conséquences. *« Les différentes parties des applications de Facebook interagissent les unes avec les autres de façon complexe »* et chaque équipe développe des modifications sans qu'il y ait une *« vision systémique unifiée »*, regrette ainsi l'employée Mary Beth Hunzaker, dans une longue note rédigée à l'occasion de son départ de Facebook, en août 2020. La conséquence ? *« Un risque accru de problèmes facilités ou amplifiés par des interactions imprévues entre des fonctions ou des services de la plate-forme. »*

A de multiples reprises, des employés témoignent, dans des documents internes, de leur incompréhension face à des comportements problématiques de leur code informatique. En Inde, ce sont des vidéos pornographiques *soft* qui se retrouvent subitement mises en avant dans l'onglet Watch, sans que personne ne comprenne pourquoi. Aux Etats-Unis, des ingénieurs s'arrachent les cheveux pour comprendre pourquoi certains groupes politiques continuent d'être recommandés aux utilisateurs alors qu'ils ne devraient plus l'être. *« Chaque expérience change la composition du fil d'actualité de manière imprévue »*, note un document de 2018. Deux ans plus tard, les analystes de Facebook se réjouissent de constater que le nombre de contenus engendrant de la colère semble en forte baisse dans les fils d'actualité, mais ils sont bien en peine d'expliquer pourquoi.

« Ce résultat pourrait en théorie être la conséquence d'un ou plusieurs changements apportés à l'algorithme dans les deux derniers mois. En dernière analyse, il pourrait être difficile de déterminer lesquels », note un document. Interrogé à ce sujet, Facebook reconnaît volontiers que ses algorithmes sont devenus des outils très complexes, mais s'enorgueillit de les modifier régulièrement pour les améliorer. Si les conséquences des changements ne sont pas toujours simples à prévoir, le réseau social assure que les sondages qu'il effectue régulièrement auprès de ses utilisateurs lui permettent de détecter rapidement tout problème majeur et, le cas échéant, de revenir en arrière si une modification s'avère trop problématique.

Ces problèmes ne viennent pas d'un manque de compétence. Facebook recrute certains des meilleurs ingénieurs au monde. Et ces travers ne sont pas non plus spécifiques au réseau social : les mêmes effets de bord imprévus se produisent *« à chaque fois que des data scientists et des ingénieurs en informatique mêlent personnalisation du contenu et amplification algorithmique, comme dans le fil de Facebook, l'onglet Pour vous de TikTok ou dans les recommandations de YouTube »*, écrit Roddy Lindsay, ex-ingénieur chez Facebook, dans une tribune publiée en juin dans le *New York Times*. Il ajoute : *« Ces algorithmes (…) perpétuent des biais et affectent la société d'une manière que leurs créateurs comprennent à peine. Facebook a eu quinze ans pour démontrer que les algorithmes de classement des contenus en fonction de l'engagement* [le nombre de commentaires, de partages…] *peuvent être conçus de manière responsable ; s'ils n'ont pas réussi à le faire jusqu'à présent, ils n'y arriveront jamais. »*

Et, les documents de Facebook le montrent, ce n'est pas faute d'avoir essayé. Les analyses réalisées en interne éclairent d'un jour nouveau les multiples modifications, majeures ou mineures, apportées par Facebook à ses différents algorithmes ces dernières années. A commencer par le crucial pivot de 2018, qui ambitionnait de mettre au premier plan ce que Facebook appelle les « interactions sociales significatives » (*meaningful social interactions*, ou MSI). Le projet avait un but clair : privilégier les contenus publiés par les proches, photos de famille et textes des amis, au détriment des contenus politiques et de ceux publiés par des médias et des pages « appeau à clic », dont la consommation passive était jugée peu utile par les internautes.

BON NOMBRE DE MESURES DE MODÉRATION MISES EN PLACE PAR L'ENTREPRISE VONT À L'ENCONTRE DE L'ESSENCE MÊME DE SA PLATE-FORME

Or, le changement MSI a eu, en partie, l'effet inverse de celui qui était recherché : comme le montrent de multiples analyses postérieures menées par les équipes de Facebook, il a favorisé les contenus clivants. En partie parce que l'algorithme modifié a donné énormément de poids aux contenus partagés par les proches, y compris lorsque ceux-ci provenaient de pages douteuses ou très engagées politiquement. Résultat, les *deep reshares*, les contenus repartagés par des amis d'amis, ont subitement pris une grande importance dans les fils d'actualités des utilisateurs, tout en étant, d'après les propres recherches de Facebook, l'un des principaux vecteurs de diffusion pour des pages d'extrême droite ou complotistes.

LONGUE LISTE DE TÂTONNEMENTS

Depuis, fin avril 2019, une analyse menée au sein de l'équipe « integrity », chargée d'étudier et de proposer des solutions contre la désinformation ou les contenus dangereux chez Facebook, estimait qu'une baisse massive du poids de ces *deep reshares* permettrait de réduire la visibilité des messages de désinformation de 25 % à 50 %, selon les formats. Ce serait *« une manière efficace et neutre pour limiter les risques que posent les contenus les plus dangereux* [sur la politique ou la santé] », estimait alors l'auteur de la proposition.

« Nous avons d'autres outils pour réduire la visibilité de certains types de contenus, comme les messages haineux ou les photos de nus », explique un porte-parole de Facebook. La diminution globale du poids des *deep reshares* est un *« outil brutal, qui touche aussi des messages positifs ou anodins en même temps que les discours possiblement violents ou provocateurs, et nous l'utilisons donc avec discernement »*. Facebook l'a mis en place récemment et de manière temporaire en Ethiopie, en Birmanie, aux Etats-Unis ou au Sri Lanka.

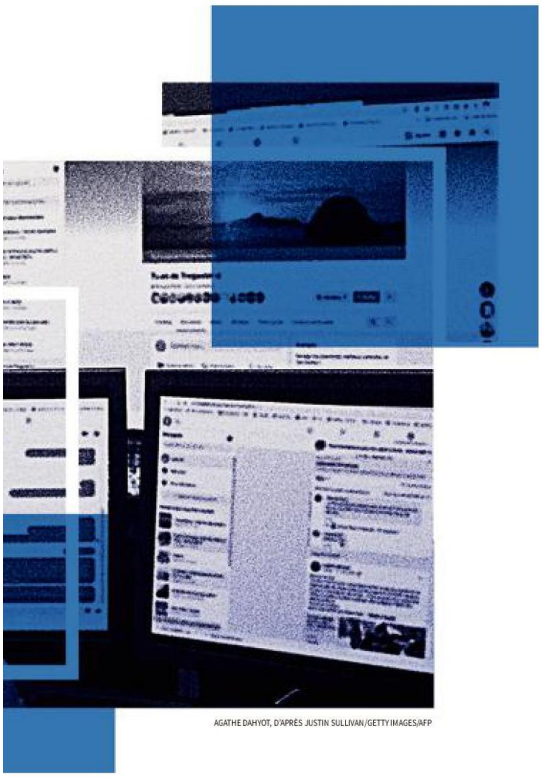
Les détracteurs de l'entreprise, dont Frances Haugen, l'accusent de privilégier, dans les choix de réglages de sa plate-forme, ses chiffres d'engagement des utilisateurs : l'activité, le temps passé, le nombre de likes, de partages… Facebook ignorerait donc largement les conséquences négatives de ses changements d'algorithmes. En 2020, pour l'élection présidentielle américaine, l'entreprise avait mis en place toute une série de mesures préventives, qu'elle a désactivées une fois l'élection passée – avant d'en réactiver une partie le 6 janvier, lors de l'attaque du Capitole menée par les partisans de Donald Trump. Ces allers-retours sont-ils le signe que l'entreprise cherche surtout à protéger ses statistiques d'usage, cruciales pour vendre de la publicité ? Facebook le nie avec véhémence, et affirme avoir parfois pris des mesures qui réduisaient l'engagement, lorsque c'était nécessaire pour la sécurité de ses utilisateurs.

En 2018, le changement d'algorithme avait pour but *« de prioriser le contenu des amis et des familles et était fondé sur des études d'experts du bien-être »*, a ainsi expliqué au *Monde* début octobre Monika Bickert, la responsable des politiques de contenu de Facebook. *« Et, comme nous nous y attendions, ce changement a en fait réduit le temps passé sur la plate-forme, de 50 millions d'heures par jour. »* Plus généralement, Facebook assure que privilégier à tout prix « l'engagement » serait un non-sens pour l'entreprise, car son intérêt à long terme est que ses utilisateurs – et annonceurs – se sentent bien sur la plate-forme, pour s'assurer qu'ils y restent.

La longue liste des tâtonnements et des modifications laisse cependant entrevoir toute la complexité d'algorithmes devenus difficiles à maîtriser, et dont les conséquences ne sont pas toujours immédiatement détectables. Malgré les nombreux changements, des problèmes demeurent. Au fil des ans, les mesures proposées par les membres de l'équipe « integrity » semblent de plus en plus complexes, quand elles ne contournent tout simplement pas le problème en suggérant des modifications, non plus des algorithmes, mais de l'interface. Ainsi, plusieurs documents des deux dernières années évoquent comme piste l'idée d'ajouter des éléments de « friction » pour inciter les utilisateurs à moins partager certains types de contenus, par exemple en le forçant à cliquer sur un article avant de le rediffuser.

SECRET INDUSTRIEL

Un document d'avril 2020 propose, lui, la mise en place d'un outil *« de transparence interne pour centraliser l'assurance-qualité et le contrôle des rétrogradations de contenus dans le fil d'actualité »*, signe que, même pour les équipes travaillant sur ces sujets, avoir une vue transversale des modifications faites dans l'entreprise est difficile. Certaines pistes de changement les plus récentes semblent même avoir un côté paradoxal, et presque ironique : depuis début 2021, Facebook mène *« dans 70 pays »* des expériences pour afficher moins de contenus politiques dans



AGATHE DAHYOT, D'APRÈS JUSTIN SULLIVAN/GETTY IMAGES/AFP

les fils de ses utilisateurs et a pérennisé ce changement aux Etats-Unis et au Canada. *« C'est une copie carbone de ce qu'ils disaient déjà en 2018 »*, s'amuse Katie Harbath, une autre ancienne employée du réseau social partie en 2021 pour fonder son entreprise. *« La réduction des contenus politiques était l'une des raisons principales du changement d'algorithme MSI : on a vraiment l'impression qu'on tourne en rond »*, regrette-t-elle.

Si l'une des plus puissantes entreprises au monde ne parvient pas à remplir ses propres objectifs, c'est peut-être aussi parce que bon nombre de mesures mises en place par Facebook ces trois dernières années vont directement à l'encontre, non pas de son modèle économique, mais de l'essence même de sa plate-forme et de ses algorithmes. Dans le fil d'actualité, tout comme dans les recommandations de pages ou de groupes à suivre, Facebook a surtout cherché à construire une machine à amplifier, capable de détecter les contenus et les comptes qui susciteront l'enthousiasme de ses utilisateurs.

Rendre plus difficile le partage, masquer, ou réduire la visibilité de ces contenus qui « marchent », est fondamentalement contraire à la mission première de l'algorithme. C'est pour cette raison que les régulateurs, aux Etats-Unis comme en Europe, s'intéressent de plus en plus au fonctionnement même des algorithmes. Certains, comme Frances Haugen ou Roddy Lindsay seraient même d'avis d'inciter les plates-formes à revenir à un classement chronologique des contenus. D'autres, dont Facebook, répondent déjà que le résultat serait pire pour l'utilisateur car les algorithmes filtrent aussi des contenus néfastes… Interpellé par les politiques, Facebook ne souhaite bien sûr pas communiquer d'informations détaillées sur ces logiciels qui sont l'un de ses principaux secrets industriels.

Mais le projet de règlement européen Digital Services Act, actuellement en discussion à Bruxelles, prévoit d'imposer aux réseaux sociaux d'être plus transparents sur leurs algorithmes et permettre aux internautes de modifier les paramètres des systèmes de classement des contenus. Un enjeu pour les régulateurs est d'arriver à vraiment avoir accès à la machine interne de ces plates-formes, jusqu'ici opaques. Le texte prévoit des audits, mais le tableau dépeint par les documents internes de Facebook pourrait pousser les politiques européens à se montrer beaucoup plus exigeants. ■

DAMIEN LELOUP ET ALEXANDRE PIQUARD

CE QU'IL FAUT SAVOIR

Les « Facebook Files » sont des centaines de documents internes, prélevés sur l'espace de discussion accessible aux seuls employés. On y trouve des analyses, des données, des comptes rendus de recherches et l'avis, sans filtre, des employés sur leur société. Une plongée dans les rouages de la machine Facebook.

La lanceuse d'alerte Frances Haugen, 37 ans, spécialiste des algorithmes, a quitté Facebook en mai 2021, après avoir copié des milliers de pages de documents. Elle les a fournis au régulateur et au Congrès américains, puis ils ont été transmis par une source parlementaire à des médias américains et européens, dont *Le Monde*.

Les documents montrent que Facebook consacre l'essentiel de ses ressources à limiter ses effets néfastes en Occident, au détriment du reste du monde. Ils prouvent que ces effets sont parfaitement connus en interne. Et que les algorithmes sont devenus d'une complexité telle qu'ils semblent échapper à leurs propres auteurs.

Le blues des équipes chargées de « l'intégrité »

Des employés témoignent de leur frustration face à leur manque de moyens et aux arbitrages de la firme

« Il n'est pas normal qu'un grand nombre de personnes (…) partent en disant "pour info, nous empirons activement le monde". » Ce commentaire, publié à la fin de l'année 2020 sur Workplace, l'outil de discussion interne des salariés de Facebook, résonne avec de nombreux témoignages similaires. *« Ces dernières semaines, nous avons assisté à un certain nombre de départs de personnes haut placées dans les équipes "integrity" [*un département chargé de superviser et de penser la modération au sein du réseau social*] (…), toutes ont exprimé des critiques spécifiques sur les limites de l'impact du travail de ces équipes au sein de Facebook »*, écrit, fin 2020, un employé dont le nom a été anonymisé.

Cette publication est issue de documents internes à Facebook récupérés par Frances Haugen, une ancienne employée, et transmis par une source parlementaire américaine à plusieurs médias, dont *Le Monde*. La lanceuse d'alerte, dernière d'une longue liste d'anciens salariés du réseau social ayant publiquement pris la parole contre leur ex-entreprise, travaillait au sein de l'équipe « civic integrity », qui rassemblait 300 employés couvrant la lutte contre la désinformation, les contenus haineux, les manipulations politiques ainsi que la protection des élections. « Civic integrity » a été démantelé en décembre 2020.

« Business integrity », « pages integrity »… Rassemblées sous la bannière « integrity », ces unités ont pour mission de trouver des façons de lutter contre les usages néfastes de Facebook, Instagram, Messenger, et de modérer le plus efficacement possible les plates-formes du mastodonte américain. Une partie d'entre elles sont, depuis la fin de l'année 2020, regroupées dans une nouvelle entité baptisée « central integrity », et dirigée par le vice-président de Facebook, Guy Rosen.

Mission d'ampleur

Les documents consultés par *Le Monde* mettent en lumière les critiques acerbes formulées par des membres de ces équipes vis-à-vis du fonctionnement de l'entreprise et des limites à leur champ d'action. *« Je ne pense pas pouvoir rester en bonne conscience : je pense que Facebook a probablement une influence négative sur la politique dans les pays occidentaux ; je ne pense pas que la direction soit dans un travail de bonne foi pour régler ce problème »*, écrit notamment, en décembre 2020, un salarié sur le départ. Interrogée par le *Monde* et trois autres médias européens (*Knack*, *Berlingske* et *Tamedia*), l'entreprise n'a pas souhaité répondre précisément à certaines questions sur les critiquées exprimées par d'anciens employés.

Facebook a mis en avant ses investissements dans la sécurité et la modération de sa plate-forme. *« Je suis fier des immenses progrès que nous avons fait. Ce progrès est en grande partie du au dévouement de l'équipe ["integrity"] pour continuellement comprendre les difficultés, identifier les manquements et appliquer des solutions »*, a déclaré Guy Rosen dans un communiqué. Plusieurs facteurs sont avancés par les employés pour expliquer leur frustration dont, en premier lieu, le manque de moyens. Dans un document interne datant d'octobre 2019, le directeur de l'équipe « civic integrity », Samidh Chakrabarti, soulignait ainsi le manque de bras face à l'ampleur de sa mission : sécuriser des élections dans le monde entier et lutter contre des manipulations politiques à grande échelle. Un manque de moyens qui a conduit ce département à prioriser les pays sur lesquels concentrer sa surveillance : *« La douloureuse réalité est que nous ne pouvons simplement pas couvrir le monde entier avec le même niveau d'efforts »*, expliquait Samidh Chakrabarti en 2019.

> **« IL Y AVAIT TELLEMENT DE CAS QUE J'ÉTAIS RÉDUITE À DÉCIDER SUR QUELS DOSSIERS TRAVAILLER EN PROFONDEUR », ÉCRIT UNE EX-SALARIÉE**

Ce problème a également été soulevé par Sophie Zhang, une autre lanceuse d'alerte, ancienne experte en analyse de données au sein de Facebook et qui a vivement critiqué l'entreprise après son licenciement en septembre 2020. Employée dans l'équipe chargée des « comportements inauthentiques » (faux « j'aime », faux commentaires…), elle a graduellement commencé à repérer et faire remonter à sa hiérarchie des opérations de manipulation politique, se battant pour que des actions soient prises, alors même que ce travail, complexe, ne faisait pas partie de ses attributions. *« Il y avait tellement de* [cas] *dans le monde que j'étais réduite à décider personnellement sur quels dossiers enquêter en profondeur »*, explique-t-elle dans son message de départ sur Workplace. L'organisation broussailleuse de Facebook complique également le travail des employés chargés de protéger ses utilisateurs.

Un document interne non daté, détaillant les équipes impliquées dans la surveillance du processus électoral en Inde, montre la densité et la granularité du fonctionnement de Facebook : pas moins de quarante départements différents sont concernés, de l'équipe luttant contre le spam aux employés « produit », en passant par les responsables de partenariats médias et les départements impliqués dans la surveillance des opérations de désinformation.

Plusieurs documents laissent entendre que des conflits éclatent parfois entre les équipes d'« integrity » et celles chargées des affaires publiques ou du développement commercial. Selon Sophie Zhang, certains employés « integrity » sont directement sous la coupe de départements « produit », chargés de développer des fonctionnalités et de faire croître le nombre d'utilisateurs. *« Les motivations sont naturellement différentes si l'on dépend d'une équipe qui est focalisée sur la croissance »*, estime-t-elle dans un entretien accordé au *Monde*.

Mais davantage que celle des services commerciaux, c'est l'influence des équipes chargées des affaires publiques qui fait l'objet des critiques les plus vives. *« J'ai entendu de nombreux collègues de l'équipe de politique de modération dire qu'ils se sentaient obligés de s'assurer que leurs recommandations soient en phase avec les intérêts des décideurs politiques »*, peut-on par exemple lire dans un document interne daté de juin 2020.

Dans les faits, les départements édictant les politiques de modération de la plate-forme ont les mêmes supérieurs hiérarchiques que ceux gérant les affaires publiques et les relations politiques, ce que l'ancien responsable de la sécurité informatique du groupe, Alex Stamos, a qualifié de *« péché originel »*. Une ancienne salariée écrit ainsi : *« Ces derniers mois, encore et encore, j'ai vu des propositions prometteuses des équipes "integrity", s'appuyant sur des recherches et données solides, être prématurément étouffées ou sévèrement limitées par des décideurs – souvent par peur des réactions du public et des responsables politiques. »*

Accusation d'ingérence

Un cas saillant survenu en Inde a par exemple suscité l'inquiétude en interne. En août 2020, le *Wall Street Journal* révèle que, quelques mois plus tôt, un cadre de Facebook dans le pays a refusé de placer Raja Singh, une responsable politique membre du parti au pouvoir, sur une liste d'organisations et d'individus dangereux, malgré des publications enfreignant les règles du réseau social sur les appels à la haine, et visant les musulmans. Selon le quotidien américain, Facebook craignait les réactions politiques qu'aurait pu entraîner une telle décision. Après de nombreuses réactions en interne, Raja Singh sera finalement banni de la plate-forme moins de deux semaines plus tard.

Cette accusation d'ingérence de la part des équipes en charge des affaires publiques est loin d'être la seule. En 2017, un mécanisme intitulé *sparing sharing*, conçu pour réduire le spam, a été mis en place : il consistait à limiter la portée des liens et contenus partagés par les internautes identifiés comme publiant beaucoup plus que la moyenne des utilisateurs. Un document interne suggère que cet outil a également eu pour effet de réduire la visibilité de certains contenus très conservateurs.

En 2020, l'entreprise a réalisé une expérience, en limitant temporairement ce mécanisme, officiellement car un audit interne l'a jugé trop imprécis. Dans un document consulté par *Le Monde*, une ancienne salariée de Facebook estime que la véritable raison de cette expérimentation a été, là encore, la crainte de réactions politiques. Interrogé à ce sujet, Facebook a éludé la question, tout en ajoutant que *« la montée de la polarisation »* des débats politiques *« a fait l'objet d'études académiques sérieuses ces dernières années »*, mais qu'aucun élément ne montre que sa plate-forme et les réseaux sociaux en général *« en sont le principal moteur »*.

Mis bout à bout, les documents examinés par *Le Monde* laissent transparaître une intense frustration de la part des salariés de Facebook. *« Beaucoup de gens ont été recrutés dans les équipes "integrity" pour essayer de rendre l'entreprise meilleure, en pensant qu'ils avaient la capacité de le faire »*, estime Sophie Zhang. Avant d'ajouter qu'en interne, Facebook a à plusieurs reprises sondé ses employés, leur demandant s'ils pensaient que les produits du groupe avaient un effet positif sur le monde : *« Ce nombre était à environ 70 % quand j'ai rejoint Facebook, il était descendu à 50 % à mon départ. »* ■

FLORIAN REYNAUD