Corriger collectivement l'information dans un espace militant

Étude de la vérification d'information participative sur les réseaux sociaux

Clément LÉROU

Mémoire de recherche sous la direction de Julien BOYADJIAN,

Maître de conférences en science politique

Majeure Sociétés numériques *Année universitaire 2024 – 2025*

Sciences Po Lille n'entend donner aucune approbation ni improbation aux thèses et opinions émises dans ce mémoire de recherche. Celles-ci doivent être considérées comme propres à leur auteur.

J'atteste que ce mémoire de recherche est le résultat de mon travail personnel, qu'il cite et référence toutes les sources utilisées et qu'il ne contient pas de passage ayant déjà été utilisé intégralement dans un travail similaire.

Remerciements

Je tiens d'abord à remercier Julien Boyadjian pour avoir accepté d'encadrer mes recherches ainsi que pour ses conseils avisés tout au long de cette année.

Je remercie également l'ensemble de mes camarades du master Sociétés Numériques pour leur soutien et pour l'excellente ambiance de travail qui a entouré la rédaction de ce mémoire.

Sommaire

Introduction

Présentation et intérêt du sujet

Mardi 7 janvier 2025, Mark Zuckerberg a déclaré que Facebook allait progressivement remplacer les *fact-checkers* professionnels de son réseau par un système de notes de la communauté similaire à celui présent sur le réseau concurrent X[1]. Ce système a été lancé en 2021 sur le réseau du milliardaire Elon Musk aux États-Unis, puis étendu au reste du monde en novembre 2022. Il consiste à permettre à des contributeurs, c'est-à-dire des utilisateurs volontaires, de rédiger des courtes annotations pour corriger des publications trompeuses. La validation de ces notes dépend d'une validation communautaire qui décide de leur affichage ou non à tous les utilisateurs du réseau, sous la publication trompeuse. Alors que la lutte contre les *fakenews* en ligne occupe une place importante dans le débat public à chaque élection depuis près d'une dizaine d'années, la question des méthodes à employer pour les réguler suscite un intérêt croissant. À la croisée des régulations étatiques et des choix des propriétaires des réseaux sociaux, les méthodes de vérification d'information apparaissent désormais comme un outil politique. Encouragé par Elon Musk après son rachat de Twitter, le programme des notes de la communauté s'inscrit dans un principe de liberté d'expression totale où le contenu est annoté sans être retiré. L'emploi d'un système décentralisé auquel chaque utilisateur est libre de contribuer évoque un principe juste, pluraliste et équitable. Dans cette étude, nous proposons de lever le voile sur ce système contributif en identifiant les mécaniques sociales et algorithmiques à l'œuvre. Cet espace de création collective au sein d'un réseau social polarisé et militant offre un cas d'étude unique. Il permet d'analyser la formation d'une production neutre et consensuelle dans l'environnement qui semble être le plus défavorable pour y parvenir. Étudier un système contributif en ligne créant collectivement des corrections sur des publications souvent partisanes pourra ainsi permettre de mieux comprendre la formation de l'opinion au sein de systèmes de démocratie participative. Parvenir à transformer des clivages en consensus semble être le défi intemporel des démocraties et sera notre sujet d'étude.

[1] Croquet, P., & Szadkowski, M. (2025). Les « community notes », un outil de modération communautaire à double tranchant. *Le Monde*, 9 janvier 2025

Définition de l'objet d'étude

Notre étude de la vérification d'information collaborative en ligne sera réalisée au travers de l'exemple des notes de la communauté du réseau social X (anciennement Twitter). Nous étudierons ce dispositif en tant que système sociotechnique, c'est-à-dire comme un système qui combine des pratiques et normes contributives avec des architectures algorithmiques. La population étudiée sera celle des contributeurs du système, c'est-à-dire l'ensemble des utilisateurs du réseau social X qui ont fait la demande de rejoindre le système des notes de la communauté et dont la demande a été acceptée. Ce groupe sera désigné par le terme de « communauté ». Les membres de cette population seront évoqués sous le terme de « contributeurs » ou de « membres de la communauté ». La participation collaborative au sein du système peut se décomposer en deux éléments : rédiger des propositions de notes et évaluer les propositions des autres. Les contributeurs seront donc parfois désignés sous les termes de « rédacteurs » ou d'« évaluateurs » lorsque nous évoquerons l'une ou l'autre de ces pratiques contributives. Notons cependant que ces désignations ne signifient pas que ces individus forment deux groupes distincts, dans la mesure où chaque contributeur peut à la fois rédiger une proposition de note et évaluer celles des autres.

Figure n°1 : Exemple de note de la communauté (Source : X.com)

Dans ce qui suit, nous pourrons également désigner la note de la communauté par les termes d'« annotation », de « correction » ou simplement de « note ». Une note qui est approuvée par le processus correctif, et donc visible de tous, pourra être désignée par les termes de « note utile », de « note validée » ou encore de « note affichée ». La publication associée pourra également être évoquée par la désignation de « tweet ». Bien que notre objet d'étude se limite au système des notes de la communauté, ce système est déployé sur le réseau social X (anciennement Twitter) et nous évoquerons le concept de polarisation de ce réseau social. La polarisation politique se divise dans la littérature en deux catégories principales. La première, appelée polarisation idéologique, concerne les croyances et les opinions des individus, notamment leur adhésion ou non à des positions idéologiques spécifiques. Dans la littérature, ce concept est abordé sous deux angles distincts. Le premier, désigné comme la « divergence », mesure l'écart entre les distributions idéologiques des différents groupes politiques[2]. Le second, qualifié d'« alignement », s'intéresse au degré de cohérence avec lequel les individus harmonisent leurs positions sur divers sujets en fonction d'un ensemble idéologique particulier[3]. La polarisation affective, quant à elle, repose sur le concept de distance sociale[4]. Elle se caractérise par une intensification des émotions négatives à l'égard des membres d'un parti politique opposé, accompagnée d'un renforcement des émotions positives envers les membres de son propre camp. Le terme de polarisation désignera dans la suite de cette étude à la fois la polarisation idéologique et la polarisation affective.

État de l'art

Avant d'aborder le cadre théorique dans lequel nous allons nous placer, nous allons réaliser une revue de la littérature liée à notre sujet d'étude. Nous commencerons par évoquer les fausses informations en ligne et les régulations liées à la désinformation pour mieux comprendre les objectifs dans lesquels s'inscrit le système des notes de la communauté. Ensuite, nous évoquerons les différentes méthodes mises en place par les réseaux sociaux

[2] Fiorina M. P., Abrams S. A. et al. (2008), Polarization in the American Public: Misconceptions and Misreadings, *The Journal of Politics*, DOI: 10.1017/S002238160808050X.
[3] Abramowitz A. I., Saunders K. L. et al. (2015), Is Polarization a Myth?, *The Journal of Politics*, DOI: 10.1017/S0022381608080493.
[4] Iyengar S., Sood G. et al. (2012), Affect, not ideology: A social identity perspective on polarization, *Public Opinion Quarterly*, DOI: 10.1093/poq/nfs038.

pour se conformer à la loi et analyserons plus spécifiquement la littérature relative aux notes de la communauté. Nous finirons par aborder des résultats issus de l'étude de Wikipédia, un système contributif en ligne beaucoup plus présent dans la littérature que notre sujet d'étude.

1. Les fausses informations en ligne constituent un risque dont il faut nuancer l'importance, mais qui a conduit à l'adoption de régulations

La mise en place de mécanismes de régulation des fausses informations en ligne prend sa source dans les débats qui ont suivi l'élection présidentielle américaine de 2016. De nombreuses fausses informations ont été relayées durant cette élection, et l'idée qu'elles puissent causer une déstabilisation potentielle des systèmes démocratiques a commencé à émerger[5]. Un nombre croissant d'études a été réalisé à partir de 2016 sur les conséquences des fausses informations en ligne. L'intérêt académique pour ce sujet a été également renforcé par la pandémie de Covid-19, durant laquelle la désinformation a été liée à des questions de santé publique. Il a ainsi été démontré que les fausses informations en lien avec le vaccin contre le Covid-19 ont eu un impact significatif sur les taux de vaccination aux États-Unis et au Royaume-Uni, mettant ainsi en danger l'ensemble de la population[6]. L'effet des fausses informations sur de nombreux domaines a par la suite été étudié. La stabilité financière mondiale, par exemple, n'échappe pas au risque posé par les fausses informations. Les *fake news* financières ont en moyenne une portée 83 % plus importante que les informations véridiques, et il a été démontré qu'elles ont un effet significatif sur le marché[7]. La sécurité publique peut également être affectée par la désinformation. Les fausses informations sur Twitter ont parfois eu un rôle prépondérant dans le déclenchement et l'amplification de mouvements violents, l'absence de modération pouvant créer un « brouillard informationnel » propice à la diffusion de fausses rumeurs[8]. Nous pouvons ajouter que l'effet de la désinformation ne nécessite pas une croyance de l'internaute exposé dans sa véracité. Être exposé à une fausse information, même sans y croire, augmente le

[5] Allcott, H., et Gentzkow, M., (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*. vol. 31, n° 2, p. 211-236.
[6] Loomba, S., de Figueiredo, A., et al., (2021). Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature Human Behaviour*. vol. 5, n° 3, p. 337-348.
[7] Clarke, J., Chen, H., et al., (2020). Fake News, Investor Attention, and Market Reaction. *Information Systems Research*. p. 1-18.
[8] Oh, O., Agrawal, M., et al., (2013). Community Intelligence and Social Media Services: A Rumor Theoretic Analysis of Tweets During Social Crises. *MIS Quarterly*. vol. 37, n° 2, p. 407-426.

degré de véracité perçu de l'information. Ce phénomène est lié à un effet de familiarité. Une information déjà consultée par un utilisateur apparaitra un peu plus véridique que s'il ne l'avait jamais rencontrée, et ce, même s'il sait qu'elle est fausse[9]. De nombreux débats académiques existent sur l'utilité de la correction de fausses informations en ligne dans la réduction des fausses croyances des usagers. Certaines études établissent même que la correction de fausses informations peut créer un effet de retour de flamme qui peut renforcer de fortes croyances préexistantes[10]. Cependant, une revue de la littérature réalisée en 2020 sur cet effet de retour de flamme est venue nuancer l'existence de ce phénomène et a établi que la plupart des études échouent à le reproduire. À l'inverse, le fait que les corrections d'information sont utiles pour réduire les fausses croyances préexistantes a beaucoup plus souvent pu être prouvé[11].

L'effet des fausses informations en ligne le plus abordé dans le débat public reste celui qu'elles auraient sur les élections et la stabilité des systèmes démocratiques. De nombreuses études ont cherché à vérifier ces affirmations, et leurs conclusions invitent à une certaine prudence vis-à-vis de ces effets prétendus. Il s'avère en effet qu'il existe de grandes disparités parmi les internautes en termes d'exposition aux fausses informations. Seuls 1 % des utilisateurs seraient exposés à 80 % des fausses informations en ligne[12]. Ces fausses informations, contrairement à une croyance répandue, n'auraient eu qu'un effet très limité dans l'élection de Donald Trump en 2016[13]. Ce résultat s'explique en partie par le fait qu'une proportion très faible de la population ne s'informe que sur les réseaux sociaux[14]. Le phénomène de désinformation est en réalité occulté par un phénomène bien plus massif : la non-information. Une proportion importante d'internautes, notamment les plus jeunes, n'est exposée qu'à très peu d'informations d'actualité en ligne[15]. La surestimation de l'effet des fausses informations sur les élections peut s'expliquer par une croyance implicite dans le fait que les internautes dotés d'un faible capital culturel constituent une masse relativement

[9] Pennycook, G., Rand, D. G., et al., (2019). Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*. vol. 116, n° 7, p. 2521-2526.
[10] Nyhan, B., Reifler, J., et al., (2010). When Corrections Fail: The Persistence of Political Misperceptions. *Political Behavior*. vol. 32, n° 2, p. 303-330.
[11] Swire-Thompson, B., DeGutis, J., et al., (2020). Searching for the Backfire Effect: Measurement and Design Considerations. *Journal of Applied Research in Memory and Cognition*. vol. 9, p. 286-299.
[12] Grinberg, N., Joseph, K., et al., (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science*. vol. 363, n° 6425, p. 374-378.
[13] Allcott, H., et Gentzkow, M., *op. cit.*
[14] Cardon, D., (2019). Pourquoi avons-nous si peur des fake news ? *Médias.* n° 59-60, p. 96-108.
[15] Boyadjian, S., (2020). Désinformation, non-information ou sur-information ? Les usages différenciés des réseaux sociaux dans la jeunesse étudiante. *Réseaux*. vol. 38, n° 222, p. 25-56.

crédule. Les *fakenews* ne sont perçues comme dangereuses que dans la mesure où l'on fait l'hypothèse que les internautes qui les consultent vont modifier leurs intentions de vote par le simple fait de les avoir lues[16]. Les formes d'appropriation de l'information sont en réalité variées, et le partage d'une fausse information n'indique pas nécessairement qu'on y adhère[17]. Il a également été démontré que les internautes partageant des fausses informations le font généralement plus par motivation partisane et hostilité envers les partis adverses que par ignorance ou manque de capacités cognitives[18]. La panique morale déclenchée par les fausses informations [19] est également partiellement liée aux révélations de ciblage informationnel, comme l'affaire *Cambridge Analytica*. On observe cependant qu'en réalité, une importante proportion des fausses informations est produite dans un objectif mercantile. De plus, les procédés de ciblage informationnels restent rudimentaires et n'influencent pas de manière significative l'opinion[20]. Notons finalement que les fausses informations ne se propagent pas uniquement sur les réseaux sociaux. Les médias traditionnels sont souvent les vecteurs les plus importants de leur diffusion. Ces médias peuvent, par exemple, partager des fausses informations par inadvertance ou pour des courses à l'audience[21].

Malgré les nuances apportées par la littérature à l'effet des fausses informations sur nos démocraties, elles ont tout de même donné lieu à de nombreuses régulations étatiques ces dernières années. Ces régulations sont en grande partie portées par des textes européens. Depuis le 17 février 2024, le DSA (*Digital Service Act*) est entré en vigueur et prévoit le devoir de combattre la désinformation pour toutes les VLOP (*Very Large Online Platform*). Les articles 34 et 35 du DSA imposent à ces plateformes de réduire le risque que constituent les fausses informations sur les démocraties. Elles ont également l'obligation de publier des rapports détaillés sur leurs pratiques de modération et de lutte contre la désinformation. À ces textes, s'ajoutent de nombreux codes de conduite à destination des plateformes, pour les accompagner dans leur mise en conformité vis-à-vis du DSA[22]. En plus du droit européen, les réseaux sociaux présents mondialement doivent également se conformer au droit des

[16] Vauchez, Y., (2022). La crédulité des crédules. Débat public et panique morale autour des fake news en France. *Émulations.* n° 41.
[17] Vauchez, Y., *op. cit.*
[18] Osmundsen, M., Bor, A., et al., (2021). Partisan polarization is the primary psychological motivation behind political fake news sharing on Twitter. *American Political Science Review*. vol. 115, n° 3, p. 999-1015.
[19] Cardon, D., *op. cit.*
[20] Cardon, D., *op. cit.*
[21] Harsin, J., (2018). Un guide critique des fake news : de la comédie à la tragédie. *Pouvoirs.* n° 164, p. 99-119.
[22] Bovermann, M., (2024). Putting X's Community Notes to the Test. *Verfassungsblog*. 08 janvier 2024.

autres régions dans lesquelles ils sont implantés. Aux États-Unis, la désinformation est beaucoup moins encadrée qu'en Europe. Le premier amendement est souvent évoqué pour justifier au contraire d'une certaine liberté d'expression généralisée. En réalité, cet amendement ne s'applique qu'au gouvernement et non aux plateformes numériques, considérées juridiquement comme de simples fournisseurs d'accès à un service en ligne[23]. C'est la section 230 du *Communication Decency Act* (CDA) qui encadre les pratiques des plateformes numériques dans ce domaine. Ce texte protège les plateformes au travers d'une double immunité. Une première, qui prévoit qu'une plateforme n'est pas légalement responsable du contenu publié par ses utilisateurs, et qu'elle ne peut pas être pénalement responsable si ce contenu viole la loi. La deuxième immunité prévue par la section 230 du CDA concerne la modération du contenu. Les plateformes sont libres de prendre des actions « en bonne foi » pour modérer tout contenu décrit sous le terme vague de « répréhensible », et ce, sans risquer de poursuites pénales. En pratique, la désinformation est considérée par les tribunaux comme un contenu « répréhensible ». Cet article a donc pour objet de protéger les plateformes d'un risque de poursuite pour atteinte à la liberté d'expression lorsqu'elles modèrent de la désinformation[24]. La législation américaine laisse donc une grande liberté aux réseaux sociaux dans leur choix de modérer la désinformation ou non. La loi les protège quel que soit le choix qu'elles retiennent. Notons que de nombreux responsables américains appellent à une évolution de la section 230 du CDA. Le camp démocrate souhaite imposer davantage de régulation des fausses informations sur les réseaux sociaux[25], tandis que le camp républicain souhaite à l'inverse garantir légalement une liberté d'expression étendue[26]. Le sujet de la modération de la désinformation est donc particulièrement clivant et les choix retenus par les grands réseaux sociaux s'inscrivent souvent en partie dans des dynamiques partisanes. La relative liberté laissée aux plateformes dans la modération de la désinformation par le droit américain implique que les méthodes de régulation sont essentiellement mises en place pour être en conformité avec le droit européen.

[23] Congressional Research Service. Liability for Online Content: An Overview of Section 230 of the Communications Decency Act [En ligne]. https://sgp.fas.org/crs/misc/LSB10082.pdf
[24] Congressional Research Service, *op. cit.*
[25] Warner, M. (2023). Legislation to Reform Section 230 Reintroduced in the Senate, House [En ligne]. https://www.warner.senate.gov/public/index.cfm/2023/2/legislation-to-reform-section-230-reintroduced-in-the-senate-house
[26] U.S. Congress. S.2972 — Safe Tech Act [En ligne]. https://www.congress.gov/bill/117th-congress/senate-bill/2972

2. Les notes de la communauté : un modèle de vérification d'information participative qui fait ses preuves

Pour répondre aux nouvelles législations évoquées dans la partie précédente, les différents réseaux sociaux ont adopté des méthodes variées. La première méthode est de faire appel à une équipe de professionnels et de spécialistes pour corriger les informations trompeuses. Des organisations tierces ont émergé comme *Snopes* ou *PolitiFact* aux États-Unis qui se chargent de corriger en continu certaines des affirmations mensongères les plus partagées en ligne[27]. Une revue de la littérature sur l'efficacité de la vérification d'information par des tiers a montré que la présence de labels d'avertissement sur du contenu mensonger réduit sa croyance de 13 à 35 % et son partage de 25 à 46 %. Il est cependant également démontré qu'indiquer les publications mensongères par un label est moins efficace que de les supprimer[28]. Le principal problème de ce type de modération est la défiance qu'il crée auprès des utilisateurs. Plus de 70 % des électeurs du parti républicain aux États-Unis déclarent considérer que les fact-checkers favorisent un bord politique plus qu'un autre[29]. Cette défiance est particulièrement forte parmi les individus âgés et conservateurs. En plus de ce manque de confiance d'une partie des utilisateurs, la modération des fausses informations par des équipes dédiées souffre d'un problème de mise à l'échelle. Les volumes de publication sur les réseaux sociaux les plus fréquentés ne permettent pas un emploi viable de cette méthode à des coûts raisonnables.

Pour tenter de résoudre ces problèmes de mise à l'échelle et de défiance des utilisateurs, le réseau social Twitter a mis en place en 2021 un programme test de système de vérification d'information collaborative. Nommé initialement *Birdwatch* avant d'être renommé *Community Notes* (Notes de la communauté), le programme a été étendu à tous les utilisateurs se trouvant aux États-Unis en octobre 2022, puis à tous les utilisateurs du réseau en décembre de la même année. Le système des notes de la communauté consiste à faire apparaitre de manière collaborative une note visible de tous, sous des publications jugées fausses ou trompeuses. Ces notes sont généralement composées de quelques phrases et d'un

[27] Chuai, Y., Tian, H., et al., (2024). Did the Roll-Out of Community Notes Reduce Engagement With Misinformation on X/Twitter? *Proceedings of the ACM on Human-Computer Interaction.* vol. 8.
[28] Martel, C., Rand, D. G., et al., (2024). Misinformation warning labels are widely effective: A review of warning effects and their moderating features. *Current Opinion in Psychology*.
[29] Pew Research Center (2019). Republicans Far More Likely Than Democrats To Say Fact Checkers Tend To Favor One Side [En ligne]. https://www.pewresearch.org/short-reads/2019/06/27/republicans-far-more-likely-than-democrats-to-say-fact-checkers-tend-to-favor-one-side/

lien vers une source réfutant ou contextualisant le contenu de la publication. Participer au système des notes de la communauté se fait sur la base du volontariat et n'importe quel utilisateur peut en faire la demande. Pour que celle-ci soit acceptée, il faut cependant que l'utilisateur n'ait jamais violé les règles d'utilisation de X, qu'il ait rejoint le réseau depuis au moins six mois et qu'il ait renseigné un numéro de téléphone vérifié sur son compte. Lorsqu'une note est rédigée, seuls les contributeurs peuvent la voir dans une interface qui leur est dédiée, afin de pouvoir l'évaluer. Une fois qu'elle a obtenu un nombre suffisant d'évaluations, un algorithme décide de lui accorder le statut de note utile ou non. Lorsqu'une note obtient ce statut utile, elle devient visible par l'ensemble des utilisateurs du réseau social sous la publication trompeuse. La validation d'une note envoie également une notification à l'ensemble des usagers ayant interagi avec la publication annotée pour les en informer. L'intégralité des utilisateurs ayant aimé, repartagé, cité ou commenté la publication trompeuse est avertie à posteriori de la présence d'une note sur celle-ci, qu'ils sont invités à consulter.

Un nombre limité d'études a été réalisé sur le système des notes de la communauté. Une part importante d'entre elles vise à établir leur efficacité dans la réduction de la propagation de fausses informations sur le réseau X. Ces études révèlent une relativement bonne efficacité du système, avec notamment une réduction des partages des publications post-annotation estimée à 49%[30] ou à 61%[31]. Quand une publication est corrigée, la probabilité de suppression de la publication par son auteur augmente de 80% et le volume de réponses et de partages est également diminué de 30% environ[32]. Ce mécanisme de suppression volontaire de la publication par son auteur une fois celle-ci annotée semble plus être liée à l'influence observée sur les autres utilisateurs qu'à l'influence présumée[33]. C'est-à-dire que l'utilisateur qui retire sa propre publication le fait à la suite de réactions hostiles de certains usagers, plus que par anticipation des réactions des usagers qui seront plus tard confrontés à la note. Ce résultat est appuyé par la mesure d'une augmentation de sentiments négatifs dans les réponses d'une publication ayant été annotée. En moyenne, les éléments liés au champ lexical de la colère y augmentent de 13.2 % et ceux de l'indignation morale de 16 %[34]. Au-

[30] Renault, T., Restrepo-Amariles, D., et al., (2024). Collaboratively Adding Context to Social Media Posts Reduces the Sharing of False News.
[31] Chuai, Y., Pilarski, M., et al., (2024). Community notes reduce the spread of misleading posts on X.
[32] Renault, T., Restrepo-Amariles, D., et al., *op. cit.*
[33] Gao, Y., Zhang, M. M., et al., (2024). Can Crowdchecking Curb Misinformation? Evidence from Community Notes.
[34] Chuai, Y., Sergeeva, A., et al., (2024). Community Fact-Checks Trigger Moral Outrage in Replies to Misleading Posts on Social Media.

delà de l'effet positif à l'échelle d'une publication annotée, les avantages du système des notes de la communauté se retrouvent partiellement à l'échelle du réseau entier. Il a ainsi été mesuré dès le lancement du programme test *Birdwatch* une réduction en moyenne de la désinformation et de la présence de sentiments forts dans les publications à l'échelle du réseau [35]. De plus, une étude comparative des différentes formes de correction de désinformation montre que les systèmes participatifs réduisent beaucoup moins la diversité des sources partagées par les individus corrigés que les systèmes centralisés[36]. Il semble donc que les systèmes de vérifications collaboratifs parviennent à moins polariser les individus qui subissent les corrections que des systèmes traditionnels. Une majorité des études ayant évalués les performances correctives du système collaboratif conclue qu'elles sont similaires à celles d'un professionnel. Certaines publications font à la fois l'objet d'une annotation collaborative et d'un lien en commentaire vers un site de vérification d'information professionnel. Dans plus de 80% des cas les deux systèmes de vérification sont parfaitement en accord[37]. Le processus de filtrage et de sélection des notes à afficher semble être principalement à l'origine de ces bons résultats. Les notes citant des sources classées comme relativement neutres et factuelles reçoivent plus d'approbation et de notations positives que les autres[38]. De même, les notes contenant un langage émotionnel sont moins susceptibles d'être classées comme étant utiles[39]. Une analyse linguistique poussée de l'ensemble des notes de la communauté révèle que, dans l'ensemble, les notes ont presque toutes un haut score de lisibilité et de neutralité linguistique. Les notes jugées utiles contiennent cependant plus de pensée analytique que la moyenne et celles jugées comme non utiles ont, à l'inverse, plus de langage émotionnel, de négativité et d'injures[40]. Finalement, les utilisateurs du réseau X accordent plus de confiance aux notes de la communauté qu'à des signalements d'experts. Cette confiance semble toutefois davantage liée au contexte et aux sources fournies qu'à la personne rédigeant la note[41].

---

[35] Borwankar, S., et Zheng, J., (2022). Democratization of Misinformation Monitoring: The Impact of Twitter's Birdwatch Program. *SSRN Electronic Journal*.
[36] Kim, J., Wang, Z., et al., (2025). Differential impact from individual versus collective misinformation tagging on the diversity of Twitter (X) information engagement and mobility. *Nature Communications*. vol. 16, n° 973.
[37] Pilarski, M., Solovev, K. O., et al., (2024). Community Notes vs. Snoping: How the Crowd Selects Fact-Checking Targets on Social Media. *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*. vol. 18, p. 1262-1272.
[38] Kangur, U., Chakraborty, R., et al., (2024). Who Checks the Checkers? Exploring Source Credibility in Twitter's Community Notes.
[39] Phillips, S. C., Wang, S. Y. N., et al., (2024). Emotional language reduces belief in false claims.
[40] Yao, M., Tian, S., et al., (2024). Readable and neutral? Reliability of crowdsourced misinformation debunking through linguistic and psycholinguistic cues. *Frontiers in Psychology*. vol. 15, n° 1478176.
[41] Drolsbach, C. P., Solovev, K., et al., (2024). Community notes increase trust in fact-checking on social media. *PNAS Nexus*. vol. 3, n° 7.

Toutefois, ces bonnes performances apparentes des notes de la communauté dans la réduction des fausses informations restent à nuancer. Le principal problème du système à ce jour est le délai d'affichage. En moyenne, une note met 75 heures pour être affichée sur un message trompeur. Même si la réduction des repartages à l'issue de ce délai est en moyenne de 60 %, elle n'est que de 11 % sur l'ensemble de la durée de vie de la publication[42]. Ce délai d'affichage empêche ainsi de mesurer une réduction massive de l'exposition aux messages trompeurs sur l'ensemble du réseau[43] car les publications ont généralement atteint 80 % de leur viralité avant d'être annotées[44]. La demi-vie d'une publication, c'est-à-dire la durée à partir de laquelle la moitié des gens qui la verront un jour l'ont consultée, est très faible sur beaucoup de réseaux sociaux. Sur X, cette durée est généralement de moins d'une heure[45]. Il est ainsi très dur pour un système collaboratif, dont l'affichage nécessite les votes de plusieurs dizaines d'utilisateurs, d'être réellement performant. En plus de ce problème de délai d'affichage, on remarque également que le choix du contenu à annoter n'est pas neutre et repose en partie sur l'affiliation partisane. Un membre du système des notes de la communauté a ainsi trois fois plus de chances d'annoter la publication d'un adversaire politique que celle d'un co-partisan. On remarque également parmi les évaluateurs américains républicains un taux d'évaluations positives de 87,1 % pour le contenu de son camp politique, contre 25,9 % pour le contenu du camp opposé[46]. Ces limitations induisent que beaucoup considèrent que X ne respecte pas le DSA européen avec ce système des notes de la communauté. Le réseau X s'est séparé de ses équipes de modération des fausses informations en 2023 et ne fonctionne plus qu'avec ce système pour modérer la désinformation[47]. Une majorité d'utilisateurs est confrontée au contenu trompeur avant qu'une note soit mise sur la publication et, n'interagissant pas avec le plus souvent, ne reçoit pas de notification une fois la note affichée. Pour certains observateurs, un respect du DSA ne peut être obtenu qu'en combinant le système des notes de la communauté actuel avec des approches traditionnelles de vérification d'information professionnelle. Ces deux approches peuvent être complémentaires dans la mesure où les notes de la communauté se focalisent sur les fausses informations les plus virales, mais sont plus lentes, là où la vérification

[42] Chuai, Y., Pilarski, M., et al., *op. cit.*
[43] Chuai, Y., Tian, H., et al., *op. cit.*
[44] Renault, T., Restrepo-Amariles, D., et al., *op. cit.*
[45] Chuai, Y., Pilarski, M., et al., *op. cit.*
[46] Allen, J., et Martel, C., (2022). Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program. *CHI Conference on Human Factors in Computing Systems.*
[47] Bovermann, M., *op. cit.*

d'information professionnelle s'applique souvent à de plus petits comptes, mais est beaucoup plus rapide[48].

3. Les systèmes participatifs en ligne : Le modèle de Wikipédia

Le système des notes de la communauté étant très récent, il n'a pas encore fait l'objet d'un grand nombre d'études. Par exemple, très peu d'études ont été menées sur la sociologie des contributeurs ou les raisons du relatif bon fonctionnement du système. Dans l'objectif d'avoir des résultats sur les systèmes contributifs en ligne, nous pouvons étudier la littérature scientifique concernant Wikipédia, un dispositif qui a été abondamment analysé. En plus d'être un des plus importants systèmes contributifs en ligne, des revues de la littérature démontrent que le contenu de Wikipédia est d'une qualité comparable aux encyclopédies traditionnelles telles qu' *Encyclopaedia Britannica*[49]. Plusieurs études ont exploré les raisons de ce bon fonctionnement. Le premier élément mis en évidence est la diversité des rôles au sein du fonctionnement de la plateforme. Il y a en effet deux principales manières de contribuer à Wikipédia. La première est de rédiger ou de corriger du contenu et la seconde est de contribuer aux activités et à la structuration de la communauté[50]. Cette dernière consiste par exemple à voter pour promouvoir ou déclasser certains contributeurs à des postes clefs au sein de la communauté, comme celui de patrouilleur ou d'administrateur. Les patrouilleurs sont chargés de vérifier les modifications effectuées sur les pages et de les annuler si elles n'ont pas lieu d'être. Les administrateurs possèdent des droits étendus, comme celui de modifier les articles particulièrement sensibles, verrouillés par défaut[51]. La qualité du contenu repose donc en grande partie sur un système de double évaluation collaborative. Une évaluation de la qualité des contenus rédigés et une évaluation de la qualité des contributions des membres de la communauté[52]. Les standards de qualité et règles partagées entre contributeurs constituent également un ensemble de bonnes pratiques qui ont

[48] Pilarski, M., Solovev, K. O., et al., *op. cit.*
[49] Fallis, D., (2008). Toward an Epistemology of Wikipedia. *Journal of the American Society for Information Science and Technology.* vol. 59, n° 10, p. 1662-1674.
[50] Forte, A., Larco, V., et al., (2009). Decentralization in Wikipedia Governance. *Journal of Management Information Systems.* vol. 26, n° 1, p. 49-72.
[51] Spinellis, D., Louridas, P., et al., (2008). The collaborative organization of knowledge. *Communications of the ACM*. vol. 51, n° 8, p. 68-74.
[52] Stvilia, B., Twidale, M., et al., (2008). Information Quality Work Organization in Wikipedia. *Journal of the American Society for Information Science and Technology*. vol. 59, n° 6, p. 983-1001.

contribué à garantir de hauts standards de qualité sur la plateforme. Ces règles sont en constante évolution par des échanges entre membres et des arbitrages, et il a été démontré que la qualité d'une page Wikipédia est fortement corrélée avec la taille des discussions entre contributeurs[53]. Il est fréquent qu'émergent des désaccords au sein de ces discussions, particulièrement pour des pages relatives à des sujets clivants. Cependant, la structure hiérarchique de la plateforme a été pensée pour que ces conflits puissent être résolus. Les rédacteurs apportent des expertises thématiques et stimulent la réflexion critique par leurs échanges, et les administrateurs encadrent les débats et tranchent en cas de désaccord profond[54]. Notons également qu'en plus de ces mécanismes de gouvernance formels, les mécanismes informels comme la réputation, le mentorat ou les canaux de discussion parallèles jouent un rôle fondamental dans le fonctionnement de Wikipédia. Ces éléments informels semblent même jouer un rôle plus important que la structure formelle, avec notamment un rôle central d'individus très expérimentés, qui servent souvent de médiateur ou de mentor[55]. L'efficacité de cette structure formelle et informelle repose également sur la capacité de la plateforme à motiver des utilisateurs à y contribuer. De nombreuses études ont examiné les raisons de ces motivations. Il en ressort que la motivation principale est le plaisir de contribuer[56]. La participation à la rédaction du contenu de Wikipédia est également motivée dans une moindre mesure par des motivations extrinsèques comme le développement personnel et la réciprocité, c'est-à-dire le fait de donner en retour ce qu'on a reçu[57]. D'autres études montrent que des facteurs sociaux et relationnels comme le sentiment d'appartenance à une communauté jouent un rôle important dans la motivation des utilisateurs[58]. Il est cependant assez consensuel dans la littérature que des motivations comme la réputation, l'altruisme ou l'idéologie du savoir libre n'ont pas un effet significatif sur la motivation à contribuer[59].

[53] Stvilia, B., Twidale, M., et al., *op. cit.*
[54] Arazy, O., et Nov, O., (2011). Information Quality in Wikipedia: The Effects of Group Composition and Task Conflict. *Journal of Management Information Systems*. vol. 27, n° 4, p. 71-98.
[55] Schroeder, A., Wagner, C., et al., (2012). Governance of Open Content Creation: A Conceptualization and Analysis of Control and Guiding Mechanisms in the Open Content Domain. *Journal of the American Society for Information Science and Technology*. vol. 63, n° 10, p. 1947-1959.
[56] Nov, O., (2007). What Motivates Wikipedians? *Communications of the ACM*. vol. 50, n° 11, p. 60-64.
[57] Li, D., Xu, B., et al., (2014). An Empirical Study of Motivations for Content Contribution and Community Participation in Wikipedia. *Information and Management*.
[58] Cho, H., Chen, M., et al., (2010). Testing an Integrative Theoretical Model of Knowledge-Sharing Behavior in the Context of Wikipedia. *Journal of the American Society for Information Science and Technology*. vol. 61, n° 6, p. 1198-1212.
[59] Li, D., Xu, B., et al., *op. cit.*

Le relatif bon fonctionnement de Wikipédia semble donc reposer sur des mécanismes d'évaluation démocratiques des contenus et des contributeurs. Le dialogue, les débats et le compromis semblent également jouer un rôle central, le tout encadré par des mécanismes informels d'échange et de formation entre les membres.

Cadre théorique

Les conclusions tirées de l'analyse de la littérature scientifique concernant le système Wikipédia nous conduisent à placer notre étude dans le cadre théorique de la délibération habermassienne. Au fondement de la théorie d'Habermas se trouve la distinction entre l'agir stratégique et l'agir communicationnel[60]. Le premier se caractérise par une orientation de ses actions vers le succès et la manipulation d'autrui, là où le second est défini comme l'entente mutuelle d'acteurs dans le but de se mettre d'accord sur des raisons valides pour coordonner leurs actions. Habermas précise que ces raisons valides doivent être explicitées par les individus au travers du langage et les sépare en trois catégories. La vérité, c'est-à-dire l'adéquation avec la réalité observée, la rectitude, définie comme le respect des normes morales, et la véracité, correspondant à la sincérité subjective de l'individu. Habermas affirme que la rationalité dans un système démocratique doit découler d'une discussion explicite de ces éléments, qui doivent être justifiés, défendus et débattus jusqu'à l'obtention d'un positionnement accepté de tous. Dans *Morale et communication*[61], Habermas détaille les conditions que doivent satisfaire les débats pour s'inscrire dans un agir communicationnel. Il explique qu'une éthique de la discussion doit émerger par un échange entre les participants pour définir les normes et valeurs qui encadreront les débats. Il oppose ainsi une morale transcendante ou individuelle à une morale approuvée rationnellement par tous les participants à une discussion. Pour que ces normes et valeurs soient acceptées de tous, il faut que des conditions d'échange permettent à chacun d'exprimer, de contredire et d'interroger les éléments proposés. L'auteur identifie trois conditions principales : la liberté d'expression, l'égalité de participation et l'absence de contrainte ou de toute forme de pouvoir contraignant. Ce modèle développé par Habermas s'inscrit dans le principe de la démocratie délibérative, où la décision résulte d'une délibération publique et rationnelle

[60] Habermas, J. (1981). *Théorie de l'agir communicationnel*. Paris : Fayard, 1987.
[61] Habermas, J. (1983). *Morale et communication* : Conscience morale et activité communicationnelle. Paris : Éditions du Cerf.

plutôt que d'une simple agrégation des votes. L'auteur a complété son modèle dans *Droit et démocratie. Entre faits et normes*[62] par la conceptualisation de deux sphères structurant la formation de la volonté politique dans un système de démocratie délibérative. Habermas distingue la sphère informelle, constituée de l'ensemble des espaces publics sur lesquels ont lieu les échanges et débats, de la sphère formelle, regroupant les institutions administratives et juridiques. L'opinion se forme donc dans la sphère informelle par une construction collective, puis se voit inscrite institutionnellement dans la sphère formelle. Le droit occupe selon Habermas la double fonction de garantir les conditions d'un débat démocratique et d'être le produit de celui-ci. Les règles et le droit ne sont ni l'intersection des volontés individuelles ni l'expression de la volonté d'une majorité, mais le produit d'un consensus dont ils garantissent la formation libre. La prise de décision habermassienne repose donc sur une construction collective des règles par des échanges libres et égalitaires dont le cadre moral et normatif est lui-même le produit de ce type d'échange. Notre objet d'étude sociotechnique ayant pour objectif de produire des corrections par des prises de décision collectives, nous pourrons analyser nos résultats au prisme de la délibération habermassienne.

Question et hypothèses

La revue de la littérature scientifique montre de relativement bonnes performances du système des notes de la communauté. Il parvient à cibler efficacement la désinformation et une grande partie des notes comporte du contenu neutre et bien structuré. Ces bons résultats peuvent être surprenants dans la mesure où le réseau social X est réputé pour être particulièrement polarisé et militant. Cette réputation a été confirmée par de nombreuses études qui placent généralement la plateforme en tête des réseaux les plus polarisés[63]. Il semble donc surprenant qu'une construction collective au sein de cet environnement puisse produire des résultats perçus comme majoritairement neutres, d'autant que le contenu annoté aborde souvent des éléments d'actualité particulièrement clivants. Notre hypothèse est donc que la notion d'utilité qui émerge du système communautaire ne se fonde pas sur des

[62] Habermas, J. (1992). *Droit et démocratie* : Entre faits et normes. Paris: Gallimard, 1997

[63] Yarchi M., Badeb C. et al. (2020), Political Polarization on the Digital Sphere: A Cross-platform, Over-time Analysis of Interactional, Positional, and Affective Polarization on Social Media, *Political Communication*, DOI: 10.1080/10584609.2020.1785067.

principes délibératifs libres, mais sur des normes implicites partiales partagées par les contributeurs et sur des choix algorithmiques non démocratiques. Notre étude vise à répondre à la question suivante :

> Comment les notes de la communauté parviennent-elles à coordonner efficacement une correction de la désinformation, dans un espace numérique polarisé par des logiques militantes et idéologiques ?

Terrain et méthodologie

Notre terrain de recherche sera donc le système des notes de la communauté de X. Ce choix résulte du fait qu'il s'agit du premier système de vérification d'information collaboratif déployé sur un réseau social d'envergure mondiale. Ce choix s'explique aussi par l'accès en source ouverte de l'ensemble des données du dispositif.

Notre étude sera structurée en deux temps. Une première partie sera dédiée à l'étude de l'émergence d'une notion d'utilité par les choix individuels des contributeurs. Une seconde partie traitera du système sociotechnique dans son ensemble et de la structuration de la communauté. L'objectif est ainsi de répondre à la problématique par l'étude des normes explicites et implicites à l'échelle individuelle du contributeur, puis de passer à l'échelle du système entier pour étudier la portée des choix algorithmiques et de la structuration des contributeurs qui en découle.

L'étude des normes implicites de contribution sera menée par l'analyse d'un corpus que nous avons constitué. Ce corpus se compose de 150 notes de la communauté et de leurs publications associées. Ces notes ont été sélectionnées de telle sorte que la moitié soit des notes ayant obtenu le statut « utile » et que la moitié restante ait le statut « inutile ». L'ensemble de ces notes a été prélevé dans l'espace de correction francophone pour faciliter son analyse. Les notes « utiles » ont été sélectionnées via une interface du système dédiée aux contributeurs, qui répertorie par ordre chronologique les dernières notes en français ayant obtenu le statut utile. Ces groupes de notes ont été prélevés par quatre séries de vingt

notes au début de l'année 2025. En ce qui concerne les notes « inutile », la constitution du corpus a nécessité l'emploi de méthodes plus complexes, car aucune interface ne les répertorie. Nous avons donc dû programmer un script sur la base de données du système (base de données dont nous détaillerons plus tard la nature). Ce script a sélectionné l'ensemble des notes ayant obtenu le statut « inutile » et ayant été rédigées entre les mois de juin et de décembre 2024, car cette période correspond à la période la plus récente disponible dans la base de données. Ces notes ont ensuite été filtrées à l'aide d'une identification de mots clefs pour ne conserver que celles rédigées en français. Finalement, 75 notes ont été sélectionnées aléatoirement au sein de cet ensemble. Les données recueillies ne comportant que la note sans la publication associée, une recherche manuelle sur le réseau X a été réalisée pour chacune d'entre elles afin de retrouver le tweet associé. Le corpus constitué a ensuite été analysé selon 41 critères, certains concernant la note elle-même et d'autres la publication annotée. Ces résultats ont finalement été traités par des scripts en langage Python pour en extraire des statistiques et des visualisations.

Pour conduire une partie du second chapitre de notre étude, nous avons mené une analyse de rétro-ingénierie du code source des notes de la communauté. Ce code est en source ouverte et a donc pu être téléchargé dans son intégralité. Les éléments clefs de son fonctionnement, comme les mécanismes de seuil ou de sélection des notes utiles, ont été analysés en détail. Leur fonctionnement sera vulgarisé dans certaines sections de notre étude lorsqu'il apportera un éclairage sur les éléments abordés.

L'essentiel du second chapitre de notre étude se basera cependant sur des résultats issus d'une base de données. Cette base de données a été publiée en source ouverte par les développeurs du système des notes de la communauté dans une optique de transparence. Elle contient les données système associées à l'ensemble des contributeurs et de leurs contributions. Nous nous sommes essentiellement basés sur un fichier contenant les données de chacune des 1,6 million de notes proposées par la communauté et sur un second fichier contenant les 123 millions d'évaluations réalisées par les contributeurs sur les notes de leurs pairs. L'intégralité de la base de données a été téléchargée le 19 janvier 2025, mais nous n'utiliserons que les données antérieures à décembre 2024 pour nous assurer que toutes les notes aient eu le temps d'obtenir leur statut définitif. Les programmes réalisés utiliseront le plus souvent l'ensemble des données allant du début du système et cette date. Les analyses ne se limiteront donc pas à l'espace francophone dans la mesure où filtrer un volume aussi important de données par langue représente un temps de traitement beaucoup trop long.

Certains programmes utiliseront également des échantillons aléatoires de ces données lorsque la complexité du programme ne permettra pas l'emploi de l'ensemble des données. L'objectif et les grandes lignes du fonctionnement des scripts Python réalisés pour traiter ces données seront détaillés dans le développement de notre étude lorsque leurs résultats seront évoqués.

Pour compléter ces éléments de terrain, nous avons réalisé un sondage à destination des utilisateurs de X. Ce sondage interroge les sondés sur un ensemble de questions relatives à leur perception du système des notes de la communauté. L'objectif est de compléter la notion d'utilité construite par les contributeurs par une notion d'utilité perçue par les utilisateurs. Seules 61 personnes ont répondu au sondage, ces résultats comportent donc des intervalles de confiance importants. De plus, les sondés sont majoritairement des hommes, très majoritairement des étudiants et exclusivement des jeunes. Les résultats sont donc à interpréter avec prudence et ne se généralisent pas nécessairement à l'ensemble de la population.

Le choix de ces données terrain a été réalisé dans l'objectif de combiner une analyse qualitative, issue de l'analyse de corpus, avec une analyse quantitative, issue de la base de données. L'analyse du code source du programme permet quant à elle de lever le voile sur les ressorts algorithmiques qui encadrent la structuration sociale de la communauté et la validation des notes. Le sondage apporte quant à lui un éclairage sur la perception du système par son cœur de cible, à savoir les utilisateurs de X, dans le but d'enrichir notre étude. La combinaison de ces éléments nous offrira une vision d'ensemble des mécaniques techniques et sociales à l'œuvre au sein du système.

Annonce du plan

Dans un premier chapitre, nous montrerons que l'utilité des corrections se façonne à travers les pratiques individuelles des contributeurs, qui sélectionnent et évaluent les notes selon des critères implicites liés au statut de l'émetteur, à la nature du contenu et aux normes de rigueur et de neutralité.

Dans un second chapitre, nous montrerons que l'orientation globale du dispositif vers des corrections consensuelles et stabilisées résulte de mécanismes algorithmiques et d'une structuration communautaire inégalitaire, qui favorisent l'efficacité du système tout en limitant l'expression de points de vue minoritaires.

Chapitre I. L'émergence d'une notion d'utilité dans une collaboration encadrée d'acteurs dictée par des choix individuels

Introduction du chapitre

Dans ce premier chapitre, nous allons analyser la construction des dynamiques correctives au prisme des choix individuels effectués par les contributeurs. La mise en commun des choix contributifs pris à l'échelle individuelle crée une notion commune d'utilité. Nous définirons cette notion d'utilité comme l'évaluation collective de la nécessité d'afficher publiquement une proposition de note ou d'annoter une publication. Cette notion s'applique donc à deux objets : les propositions de notes rédigées par les contributeurs et les publications du réseau social X. Ces deux aspects de l'utilité collective découlent des deux pratiques contributives du système des notes de la communauté : évaluer les notes de ses pairs et rédiger des propositions de notes. Chaque pratique construit donc sa propre notion d'utilité commune à partir des contributions individuelles. L'objectif de cette partie est de déterminer si la construction sociale de ces notions d'utilité peut expliquer les performances globales du système observées et de caractériser les normes implicites dans lesquelles elles s'inscrivent. Nous commencerons par détailler le fonctionnement des pratiques contributives individuelles et des normes explicites définies par le dispositif. Nous nous intéresserons ensuite aux caractéristiques des annotations jugées utiles dans l'objectif d'identifier et d'analyser les normes implicites qui rendent une note plus susceptible d'être validée. Enfin, nous étudierons le choix des publications à corriger et l'influence de facteurs idéologiques, sociaux et normatifs. Ce premier chapitre sera donc l'occasion de comprendre l'influence des normes implicites et explicites du système dans ses dynamiques globales. Nous croiserons des tendances observées à l'échelle du système dans son ensemble avec des analyses des choix individuels pour interpréter les résultats obtenus. Nous évaluerons donc dans ce chapitre si la coordination corrective est alignée avec les objectifs de neutralité et de rigueur affichés par la plateforme et si les normes collectives qui émergent peuvent à elles seules expliquer les performances correctives du dispositif.

Section 1 : Un dispositif ouvert de contribution mais contraint par des normes et les attentes des utilisateurs

1. Un système collaboratif de rédactions et de votes au service d'une correction collective

Avant d'aborder les normes de rédaction dictées par la plateforme, nous allons détailler le processus contributif, en commençant par l'action de rédaction, puis en évoquant le système d'évaluation par les pairs.

a. Une initiative individuelle de rédaction ouverte aux contributeurs habilités

N'importe quel utilisateur du réseau social X peut faire la demande de rejoindre le système des notes de la communauté. Les seules conditions d'accès à ce système sont le fait d'avoir rejoint le réseau social depuis plus de six mois, d'avoir renseigné un numéro de téléphone valide et de ne jamais avoir violé les règles d'utilisation de X. Une interface dédiée aux membres de la communauté est accessible depuis n'importe quelle publication pour proposer d'y ajouter une note. Cette interface demande tout d'abord à l'utilisateur d'indiquer pourquoi il considère que la publication est potentiellement trompeuse, en cochant une ou plusieurs cases relatives à des critères prédéfinis (annexe n°1). Le contributeur doit ensuite rédiger le contenu de la note qui sera potentiellement affichée sous la publication si elle est approuvée par la communauté. Le champ de rédaction impose une limite de 280 caractères, similaire à la taille maximale d'une publication sur X, ce qui impose au correcteur d'être concis. Ce processus de rédaction de note s'applique lorsque le contributeur considère que la publication est trompeuse. Dans le cas contraire où le contributeur considère que la publication n'est pas trompeuse et comporte injustement une note affichée ou en attente de validation, il peut proposer une nouvelle note. Cette note, uniquement visible par les autres contributeurs, vise à justifier le point de vue de son auteur auprès des autres membres. Une note contestant l'utilité d'une autre ne peut être rédigée que s'il existe au moins une note proposée sur la publication visée. Cette fonctionnalité est la seule permettant de créer ce qui peut s'apparenter à un débat interne entre membres. L'analyse de la base de données révèle que seuls 24 % du volume total des notes proposées sont constitués de notes de ce type. Contrairement à d'autres systèmes contributifs, aucune interface n'existe pour discuter des détails ou de la pertinence d'une note. La rédaction d'une note ne résulte que de l'initiative

individuelle d'un unique contributeur et, à part le principe précédent, aucun échange d'une quelconque nature n'est possible entre membres. Le système contributif est également entièrement anonymisé. Lorsqu'un contributeur rejoint le système collaboratif, un pseudonyme lui est affecté en générant aléatoirement trois mots du dictionnaire anglais. Il est donc impossible de retrouver le compte X d'un contributeur à partir de son compte « notes de la communauté ». Ce principe implique qu'aucune communication directe n'est possible entre les membres du système. Là où ces échanges informels jouent un rôle prépondérant dans le bon fonctionnement d'autres systèmes contributifs, comme Wikipédia, ils sont ici structurellement impossibles.

b. Un vote communautaire qui décide de l'affichage de la note

Une fois une note rédigée et proposée par un contributeur, s'amorce une phase de vote pour déterminer si elle doit être affichée ou non à tous les utilisateurs du réseau. N'importe quel contributeur peut évaluer une note, soit en découvrant la publication annotée, soit par le biais d'une interface qui sélectionne des notes nécessitant plus d'évaluations. L'évaluation d'une note se fait en deux étapes très rapides. Premièrement, le contributeur évalue la note comme utile, partiellement utile ou inutile. Il sélectionne ensuite un ou plusieurs critères prédéterminés pour préciser son choix, puis soumet son vote.

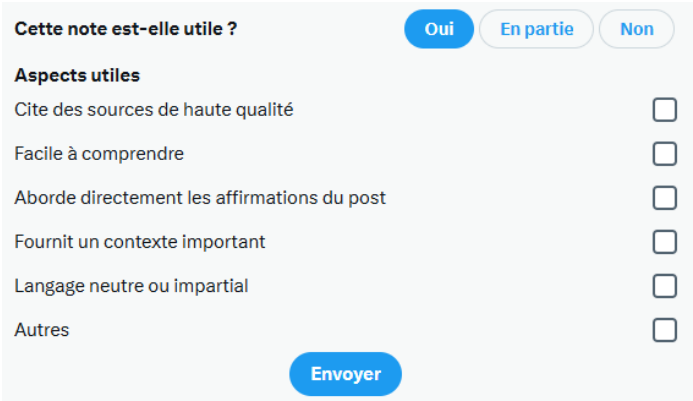

Figure n°2 : Interface de notation (Source : X.com)

Une fois un nombre nécessaire d'évaluations atteint, un algorithme complexe décide d'attribuer ou non le statut « utile » à la note. Les détails de cet algorithme seront évoqués dans une partie ultérieure de l'étude. L'accès à la base de données de l'ensemble des votes

soumis depuis la création du système nous permet de tracer l'évolution temporelle des votes sur n'importe quelle note. Voici la représentation de l'évolution temporelle des votes favorables et défavorables sur la note que nous avons fait figurer en introduction, relative au conflit israélo-palestinien.

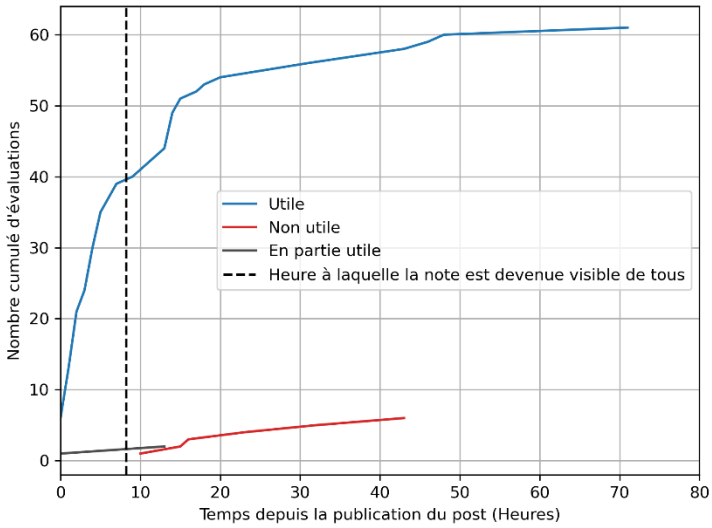

Figure n°3 : Evolution chronologique des votes d'une note de la communauté (Source : établi par l'auteur à partir des données téléchargées sur X.com)

Nous observons que la majorité des votes intervient quelques heures après la publication de la note. Près de la moitié du nombre de votes total a lieu en moins de cinq heures et il ne faut que huit heures pour que la note obtienne le statut « utile » et soit visible de tous. Le nombre de votes total nécessaire pour que la note devienne visible de tous reste limité, ici moins de quarante. Le processus de vote se caractérise donc par sa simplicité et sa rapidité. Les choix techniques et d'interface réalisés encouragent les contributeurs à évaluer rapidement de nombreuses notes. L'exemple précédent illustre la difficulté pour le système d'attribuer rapidement un statut « utile » ou « inutile » à une nouvelle note. Même si la plupart des utilisateurs qui ont consulté la publication l'ont fait durant sa première heure de publication, il faut plusieurs heures pour atteindre une masse critique d'évaluateurs et afficher la note publiquement. La volonté de minimiser ce délai est probablement la raison de la simplification à l'extrême du processus de vote à des choix simples, quantifiés et binaires. Notons que le nombre de votes et leur répartition ne sont pas visibles depuis l'interface des notes de la communauté. La représentation que nous avons réalisée se base sur des données système qui ne sont pas mises à jour en temps réel et dont l'accès et l'analyse sont complexes. Un contributeur évaluant une note n'a aucun élément lui indiquant l'état des

suffrages et n'est donc pas influencé par cette information. Le système, par ses choix techniques d'anonymat complet, d'absence de mécanismes d'échange et de visibilité de l'état des suffrages, fait en sorte que la notion d'utilité qui émergera des choix communautaires ne se base que sur le contenu de la note et de la publication. Le système technique a été conçu pour éliminer l'ensemble des éléments secondaires susceptibles d'influencer les évaluateurs.

2. Des attentes explicites qui orientent les pratiques contributives

a. Un processus libre de rédaction mais encadré par des attentes explicites

Bien que la rédaction d'une proposition de note de la communauté repose sur l'initiative des contributeurs, la plateforme X donne cependant un ensemble de recommandations et de bonnes pratiques à respecter. Ainsi, avant d'accéder à l'interface de rédaction d'une note, une fenêtre rappelle les valeurs du système.

Figure n°4 : Capture d'écran de l'interface affichée avant chaque rédaction de note (Source : X.com)

Ces valeurs sont développées sur le site des notes de la communauté et sont centrées sur trois grands axes : Mieux informer les utilisateurs de X, agir avec neutralité et aider les individus avec lesquels on n'est pas d'accord, et agir en bonne foi et de manière respectueuse [64]. Ces principes reprennent donc des éléments courants dans les règles d'utilisation de plateformes en ligne, comme l'interdiction de propos haineux ou provocateur, mais ajoutent des principes nouveaux de pédagogie et de neutralité idéologique.

[64] Community Notes. (n.d.). Valeurs fondamentales des contributeurs. X. https://communitynotes.x.com/guide/fr/contributing/values

Rejoindre le système des notes de la communauté est donc présenté par la plateforme comme une adhésion aux valeurs qu'elle porte. Les contributeurs ne sont pas simplement invités à corriger des informations, mais également à se conformer à une éthique de correction définie explicitement. Au-delà d'une simple régulation des pratiques contributives, ces principes, rappelés à chaque étape de la contribution, orientent les contributeurs dans leur construction de la notion d'utilité d'une note. Expliciter les valeurs qui doivent orienter leurs contributions et leurs jugements contribue à façonner la perception de ce qui constitue une contribution légitime ou non. Ces normes influencent également les contributeurs par la conviction que les autres les appliqueront également pour juger leur contribution. Dans un espace correctif où la validation d'une correction passe par son approbation communautaire, notre travail contributif n'est pas seulement dicté par nos valeurs, mais également par la représentation que nous nous faisons de celles des autres contributeurs. Expliciter des valeurs que l'ensemble des contributeurs est supposé suivre permet à la plateforme d'orienter leur perception.

b. L'inscription de ces normes dans un objectif de légitimation du système correctif

Ces normes explicites s'inscrivent dans une stratégie de légitimation institutionnelle du système des notes de la communauté. En affichant des principes de pluralisme et de neutralité, la plateforme cherche à positionner son système de correction comme une alternative crédible. La place centrale qu'occupe le principe de neutralité dans les orientations explicites définies par le dispositif révèle son utilisation comme norme de crédibilité. L'affirmation d'une absence de position idéologique dans la production commune devient une garantie de la qualité de ces productions. Ces valeurs rappellent les principes déontologiques des métiers de l'information, comme ceux de la profession de journaliste. Il semble y avoir une volonté de la plateforme d'inscrire son système de vérification d'information dans le prolongement des méthodes pratiquées par des professionnels. En adoptant cette éthique, le système se positionne comme un arbitre des débats qui ont lieu au sein du réseau social. Revendiquer des principes de neutralité et de rigueur permet de se placer au-dessus du contenu fortement idéologique du réseau. La forte polarisation du réseau social X, l'importance de la viralité des contenus et leur forte subjectivité imposent au système des notes de la communauté de s'en démarquer clairement. La correction neutre vient constituer une alternative à ces échanges par la réduction des clivages et des affrontements qu'elle permet. L'affichage de la place centrale de la valeur de

neutralité permet également au système de se distinguer des formes de vérifications d'information militantes du réseau. Notons que l'ensemble de ces normes de neutralité et de rigueur rappelle celles structurant le champ de la recherche scientifique. Dans son article intitulé « Le champ scientifique », Pierre Bourdieu a étudié les mécanismes conduisant les chercheurs à adopter des pratiques valorisées par leurs pairs[65]. Des normes communes présentées comme universelles, telles que la rigueur, l'objectivité et la neutralité, sont présentées comme le simple produit de l'histoire sociale de ce champ et de stratégies sociales. Bourdieu souligne le fait que le respect de ces normes et exigences communes répond principalement à une recherche d'accumulation de capital symbolique au sein du champ scientifique. Nous pouvons donc supposer que les normes définies par le dispositif des notes de la communauté ne sont mises en œuvre que dans la mesure où elles permettent un accroissement du capital symbolique des contributeurs dans le champ du système contributif.

3. Une norme d'utilité construite sans les utilisateurs, reçue de manière ambivalente

Après avoir analysé les normes explicites du système influençant la création d'une notion d'utilité par les contributeurs, nous allons maintenant nous intéresser à la réception de cette notion par les utilisateurs de X. Même si ce ne sont pas les utilisateurs qui construisent eux-mêmes cette notion d'utilité, l'efficacité du système des notes de la communauté dépend de la confiance des utilisateurs du réseau social.

a. Une forte confiance des utilisateurs dans l'efficacité globale des notes de la communauté, mais qui reste relative par rapport à d'autres dispositifs de vérification

Une note de la communauté ayant pour rôle de corriger des publications mensongères ou de contextualiser des publications trompeuses, il est important qu'une majorité d'utilisateurs aient confiance dans la véracité de ces corrections. Notre sondage réalisé auprès de 61 utilisateurs réguliers de X révèle qu'ils ont majoritairement confiance dans la

[65] Bourdieu, P. (1976). *Le champ scientifique.* Actes de la recherche en sciences sociales, 2(2-3), 88-104.

véracité des notes de la communauté, avec plus de la moitié des répondants qui affirment avoir un degré de confiance de 4 sur 5. Il est également intéressant de noter que cette confiance n'est pas totale, seuls quatre répondants affirment avoir un degré de confiance de 5 sur 5. De même, les répondants se prononcent majoritairement comme confiants dans la capacité du système à limiter la propagation des fausses informations.

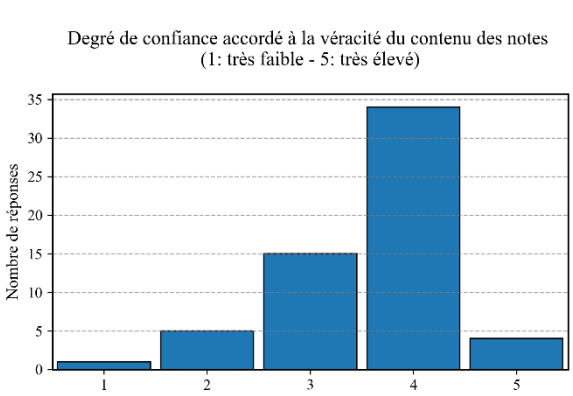

Figure n°5 : Résultats du sondage sur le degré de confiance accordé à la véracité du contenu des notes (Source : établi par l'auteur)

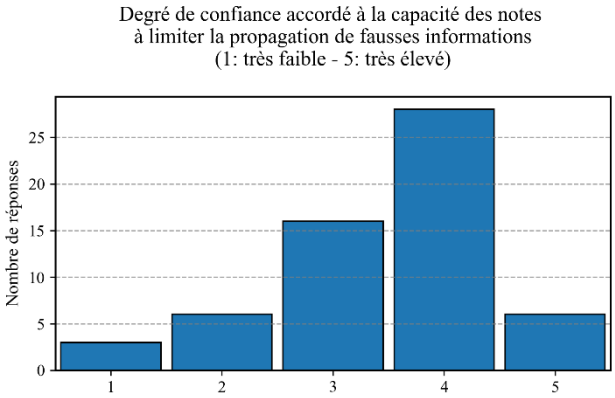

Figure n°6 : Résultats du sondage sur le degré de confiance dans la capacité des notes à limiter la propagation des fausses informations (Source : établi par l'auteur)

Malgré ces relativement hauts niveaux de confiance, lorsqu'on demande aux sondés de donner les méthodes de vérification d'information en lesquelles ils ont confiance, les systèmes collaboratifs ne recueillent que 57 % d'avis favorables, contre 72 % pour la vérification d'information par un professionnel. Même si les notes de la communauté ne génèrent pas autant de défiance que d'autres méthodes de vérification d'information comme le fact-checking par IA, le système ne semble pas convaincre autant que l'intervention de professionnels de la vérification d'information.

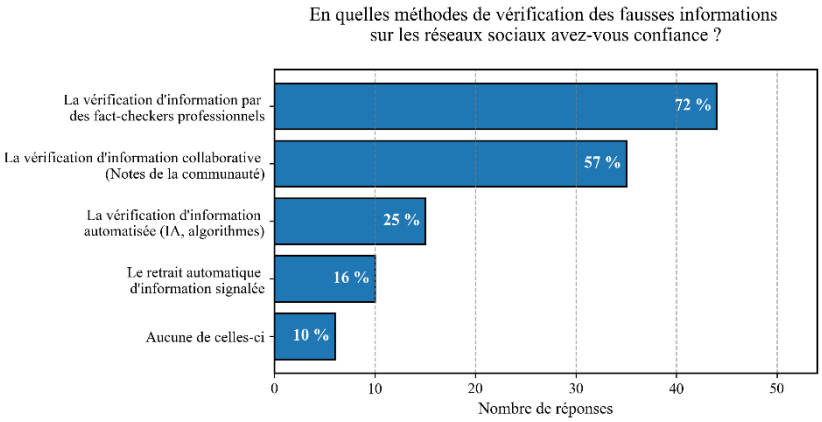

Figure n°7 : Résultats du sondage sur la confiance dans différentes méthodes de vérification d'information (Source : établi par l'auteur)

Nous rappelons que ces résultats doivent être interprétés avec prudence en raison de la taille réduite de l'échantillon et du profil des répondants, majoritairement des jeunes étudiants qui réalisent des études supérieures. Le profil sociodémographique de notre population sondée peut expliquer la relativement faible défiance observée à l'encontre des vérificateurs d'information professionnels. Leur appartenance à une catégorie sociale caractérisée par un fort capital culturel peut expliquer le crédit plus important qu'ils accordent à un système de vérification institutionnel, composé de professionnels qui incarnent une expertise légitime.

b. <u>Une défiance marquée vis-à-vis des rédacteurs des notes</u>

Malgré le fait qu'une proportion importante des utilisateurs sondés ait confiance dans la véracité des notes proposées et dans l'utilité globale du système, il apparait cependant qu'ils doutent de la compétence et de la neutralité du rédacteur des notes.

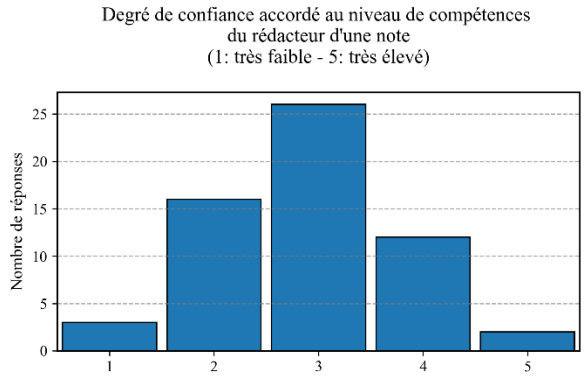

Figure n°8 : Résultats du sondage sur le degré de confiance accordé aux compétences du rédacteur d'une note (Source : établi par l'auteur)

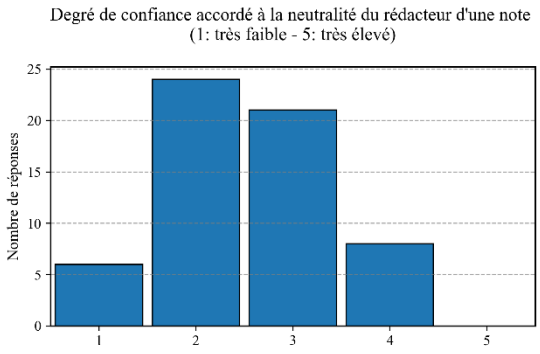

Figure n°9 : Résultats du sondage sur le degré de confiance accordé à la neutralité du rédacteur d'une note (Source : établi par l'auteur)

Alors que la vérification d'information professionnelle repose sur une confiance dans la neutralité et la compétence du professionnel, nous observons que, pour les sondés, cette confiance est relativement faible envers les contributeurs des notes de la communauté. La perception de qualité des notes repose donc bien plus sur une confiance dans le système collaboratif global que dans les individus le composant. C'est-à-dire que la confiance est donnée au dispositif en raison de sa structure et de ses règles, indépendamment des individus qui le composent. Ce type de confiance est appelé « confiance systémique » par Niklas Luhmann[66]. Elle caractérise la confiance dans des systèmes sociaux dont on ne peut pas, par nous-mêmes, vérifier le bien-fondé. L'auteur prend l'exemple d'un malade recevant un diagnostic. La confiance qu'il aura dans la qualité de ce diagnostic résultera moins de la confiance en son médecin que de sa confiance générale dans la médecine. Les utilisateurs de X semblent donc accorder leur confiance à la forme de régulation communautaire horizontale mise en place par le système correctif. Une majorité des sondés estime donc qu'un système correctif collaboratif peut produire des résultats fiables, malgré le fait qu'il soit composé d'individus qui individuellement ne le sont pas.

[66] Ouzilou, O. (2009). Niklas Luhmann, La confiance, un mécanisme de réduction de la complexité sociale. *Revue Interrogations.*

Encart n°1 : Une utilité détachée des enjeux démocratiques

Pour aller plus loin dans la compréhension de l'utilité perçue du système correctif par les utilisateurs, nous pouvons nous demander si, au-delà de la confiance en la véracité du contenu, le système dans son ensemble est perçu comme nécessaire. L'objectif explicite du système des notes de la communauté est de corriger les fausses informations sur le réseau social X. Comme évoqué dans l'état de l'art, la correction des fausses informations en ligne s'inscrit dans un respect de la loi. Ces lois reposent le plus souvent sur une conviction que les fausses informations menacent la démocratie. Le système des notes de la communauté en lui-même peut cependant s'inscrire dans un second objectif global, celui de préserver la liberté d'expression. Ce système se caractérise en effet par le fait de corriger sans supprimer le contenu trompeur et par sa structure décentralisée. Ces deux éléments sont perçus par certains comme une sécurité face à un risque de censure ou de manipulation des systèmes correctifs par des entités privées ou étatiques. Nous pouvons donc sonder notre population sur sa perception de ces deux risques : le risque posé par les fausses informations sur la société, et le risque de censure posé par leur modération.

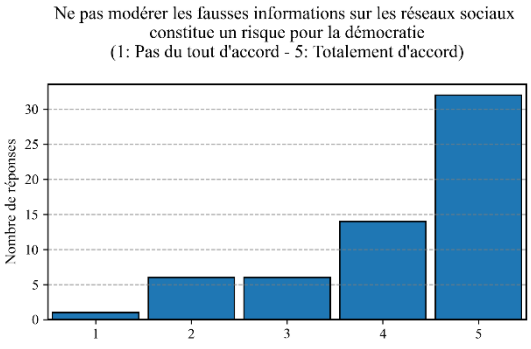

Figure n°10 : Résultats du sondage sur la perception du risque pour la démocratie de ne pas modérer la désinformation (Source : établi par l'auteur)

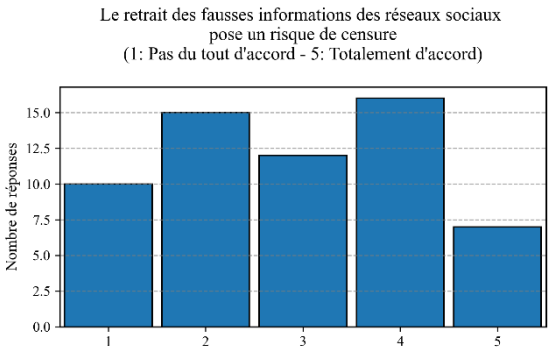

Figure n°11 : Résultats du sondage sur la perception du risque de censure du retrait des fausses informations (Source : établi par l'auteur)

Il apparait le résultat surprenant d'un relatif consensus sur la question du risque posé par les fausses informations sur la démocratie, mais d'une forte division des sondés sur la question du risque de censure. Pour analyser plus en détail ce second aspect, nous pouvons nous intéresser à un aspect de la correction d'information : le choix entre retrait ou contextualisation. Quelle que soit la méthode de vérification employée, une fois le contenu

identifié, deux solutions s'offrent aux concepteurs du dispositif : soit retirer le contenu trompeur pour le rendre invisible, soit le laisser en le signalant par le biais d'un marqueur ou d'une note. La seconde méthode est généralement privilégiée par ceux qui sont soucieux du risque de censure posé par le système correctif, là où les autres vont privilégier la première, plus efficace[67]. Lorsque nous interrogeons les sondés sur leur degré d'accord avec différentes méthodes de correction des fausses informations, il apparait qu'une majorité est étonnamment favorable à la fois à un retrait du contenu et à une contextualisation sans retrait. Les résultats des questions du sondage relatif à la perception des sondés sur le retrait ou non du contenu trompeur sont disponibles en annexe n°2. Ces résultats indiquent que la méthode de correction des informations n'a pas d'importance aux yeux d'une majorité des sondés. Ils ne semblent pas associer la méthode corrective employée et des enjeux démocratiques plus larges. Ce résultat vient surtout souligner le fait que les sondés ne se préoccupent pas particulièrement des considérations démocratiques entourant les systèmes de correction d'information. Même si l'homogénéité sociale des sondés invite à une certaine prudence dans l'interprétation des résultats, il semble que l'utilité perçue des notes de la communauté par les utilisateurs reste plus de l'ordre du confort d'utilisation du réseau que de la préservation de grands principes de démocratie ou de liberté.

Pour conclure cette section, l'utilité perçue d'une correction varie selon le type d'acteur. Les créateurs du dispositif des notes de la communauté définissent cette utilité par des normes explicites de pluralité et de neutralité. Les utilisateurs de X font davantage confiance au dispositif technique qu'aux contributeurs. Les contributeurs vont construire collectivement leur notion d'utilité par leurs choix contributifs. Ces choix contributifs, marqués par leur simplicité et leur anonymat complet, conduisent à ce que la notion d'utilité ne se construise qu'à partir du contenu des notes proposées et des publications, et non en fonction des votes des autres ou du statut du contributeur. La perception de l'utilité par les contributeurs, qui détermine la production corrective, sera le sujet de nos deux prochaines sections.

[67] Martel, C., Rand, D. G., et al., *op. cit.*

Section 2. Une utilité perçue des notes qui dépend de leur neutralité et de leur rigueur argumentative

Dans cette deuxième section, nous allons utiliser les résultats issus de l'analyse du corpus pour préciser les caractéristiques de la notion d'utilité de l'affichage d'une note. Cette section vise à identifier les critères implicites qui conduisent la communauté à décider de l'utilité d'une note ou de son inutilité. Nous rappelons que le corpus est constitué de 150 notes sélectionnées aléatoirement, dont la moitié a obtenu le statut « utile » et dont la seconde moitié a été jugée inutile. Les graphiques figurant dans cette section représenteront le taux de notes validées au sein du corpus parmi les notes qui ont une certaine caractéristique. C'est-à-dire que si ce taux est supérieur à 50 %, cela signifie que la caractéristique est surreprésentée parmi les notes utiles du corpus. À l'inverse, si ce taux est inférieur à 50 %, il indique une sous-représentation des notes ayant la caractéristique étudiée au sein du groupe des notes utiles, et donc que le critère réduit la probabilité de validation d'une note. Il convient cependant de garder une certaine prudence vis-à-vis des résultats de cette section et de la suivante, car nous ne mettons en évidence que des corrélations, qui ne sont pas nécessairement preuves de causalité. De plus, le système algorithmique, que nous analyserons au début du second chapitre, exerce également une forte influence dans les résultats qui seront exposés. Nous nous limiterons donc aux résultats suffisamment significatifs pour que l'hypothèse qu'ils soient partiellement causés par les normes implicites des contributeurs soit raisonnable.

1. Le ton et le registre de la note comme condition de recevabilité

a. La neutralité du ton comme norme centrale d'acceptation communautaire

La neutralité du ton est le premier critère que nous allons étudier. Parmi les critères d'analyse du corpus qui ont été sélectionnés figure la neutralité ou la partialité du ton employé, ainsi que la présence d'un parti pris idéologique flagrant du rédacteur.

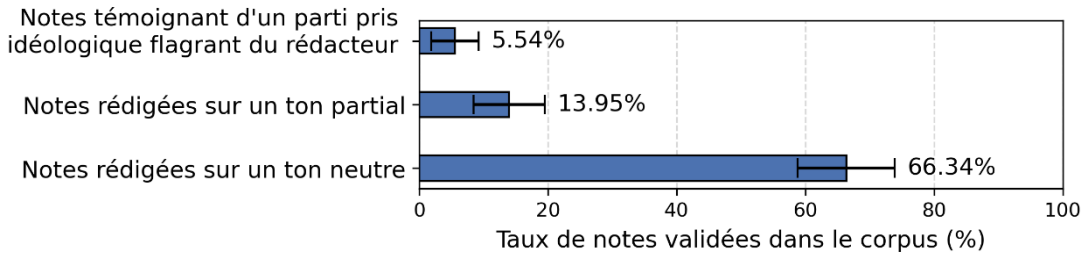

Figure n°12 : Taux de notes validées dans le corpus selon des caractéristiques de neutralité du ton (Source : établi par l'auteur à partir des données collectées sur X.com)

Les résultats obtenus témoignent d'une influence importante de la tonalité employée sur la validation d'une note. Parmi les notes du corpus qui emploient un ton neutre, 66,34 % figurent parmi les notes jugées utiles. À l'inverse, les notes employant un ton partial sont massivement surreprésentées parmi les notes jugées inutiles par la communauté, avec seulement 13,95 % d'approbation. Ces écarts statistiques importants soulignent l'importance de la norme de neutralité au sein de l'espace des contributeurs. Le collectif semble attendre de ses membres qu'ils jouent un rôle d'arbitre de la véracité des informations des utilisateurs du réseau. Parmi les critères de conformité qui émergent de ce rôle, le plus important semble être de donner l'impression de s'effacer derrière les faits énoncés et de renoncer à tout positionnement partial. Les notes du corpus contenant un parti pris idéologique évident sont massivement rejetées par la communauté d'évaluateurs, avec seulement 5,54 % qui font partie du groupe des notes utiles. Une note engagée semble donc être presque systématiquement disqualifiée, indépendamment de la véracité de la correction effectuée. La communauté semble donc valoriser des annotations donnant une impression de surplomb désaffilié du contenu corrigé. En réalité, aucune correction n'est parfaitement neutre, et un ton neutre peut cacher les intentions sous-jacentes idéologiques et partiales de son auteur. La neutralité perçue correspond plus à la parole jugée « acceptable » par une majorité et aux formes de rationalité dominantes. Ce cadre normatif ne valorise donc pas seulement une vérité factuelle, mais plutôt une certaine manière de l'énoncer, ancrée dans des codes partagés et dans un positionnement dominant. Adopter un ton qui sera perçu comme détaché et idéologiquement neutre apparait ainsi comme le critère premier adopté par les contributeurs pour statuer de l'utilité d'une note.

b. L’agressivité et la mise en cause personnelle rompent la légitimité

Certaines notes de notre corpus ne se limitent pas à une partialité idéologique, mais intègrent un jugement personnel de l’auteur de la publication, voire des insultes.

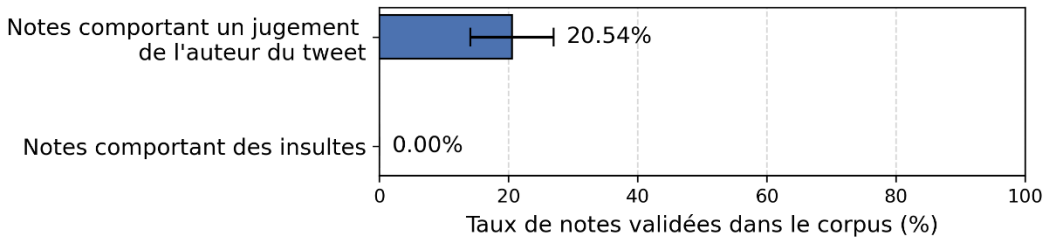

Figure n°13 : Taux de notes validées dans le corpus selon des caractéristiques d’agressivité et de mise en cause personnelle
(Source : établi par l’auteur à partir des données collectées sur X.com)

Les insultes sont présentes dans 7,33 % des notes du corpus, mais aucune d’entre elles n’a été jugée utile par la communauté, signe d’un rejet total de ce genre d’annotation. Les corrections contenant un jugement de l’auteur du tweet, comme une mise en doute de sa morale, de ses capacités ou de ses intentions, ne sont jugées utiles que dans 20,54 % des cas. Le rejet absolu des insultes constitue un résultat prévisible, mais qui souligne l’exigence de ton pacifié dans les annotations. La correction est pensée comme un procédé de clarification et de précision, et non comme un rapport de force ou de domination. La violence verbale semble constituer la transgression morale la plus importante au sein du système des notes de la communauté, mais n’est pas particulièrement rare, avec plus d’une note sur quatorze du corpus qui en contient. Cette présence notable d’insultes et d’agressivité dans les notes peut s’expliquer par un effet de porosité entre l’environnement du réseau et le système de correction. Le réseau social X se caractérise en effet par une présence importante de contenu agressif et de rapports de force idéologiques virulents. Même si le système des notes de la communauté obéit à des normes très différentes, il n’est pas exclu que les styles et usages du réseau influencent ceux de certains correcteurs ayant mal intériorisé les normes implicites du dispositif de correction. La présence d’insultes ou de notes totalement incohérentes, qui représentent 6,67 % du corpus, peut aussi révéler une instrumentalisation de la plateforme par certains contributeurs dans le but de tester ses limites ou d’exprimer une opposition au système.

Le faible taux de validation des notes comportant un jugement personnel de l'auteur confirme le fait que la communauté sanctionne également ce type d'annotation, perçue comme hors du cadre approprié. Il apparait donc que la volonté collective qui émerge du système ne perçoit pas les notes comme une mise en cause ou une accusation, mais simplement comme un complément d'information corrigeant le fond d'une publication trompeuse. L'espace de correction façonne donc la notion d'utilité de manière à privilégier le contenu mesuré et maîtrisé. Il semble y parvenir en suivant une logique d'exclusion de tout contenu qui contrevient à cet objectif, par des insultes ou des attaques personnelles par exemple.

c. Les fautes d'orthographe créent un filtre social de légitimité corrective

Près d'une note sur cinq de notre corpus contient une ou plusieurs fautes d'orthographe. Malgré leur représentation importante, elles ne sont jugées utiles que dans 9,31 % des cas.

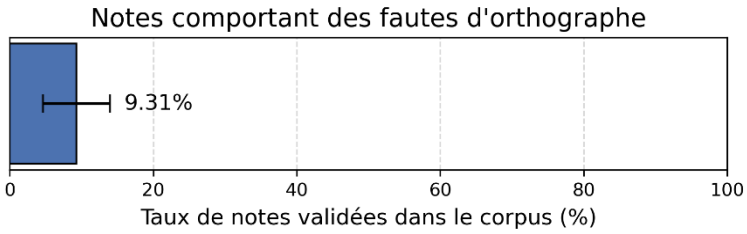

Figure n°14 : Taux de notes « utiles » dans le corpus parmi celles comportant des fautes d'orthographe
(Source : établi par l'auteur à partir des données collectées sur X.com)

La mauvaise orthographe semble donc constituer un élément disqualifiant de premier plan. Dans un espace où le rédacteur d'une note doit convaincre les autres contributeurs de l'utilité de son annotation, une bonne orthographe apparait presque indispensable pour obtenir une validation collective. Elle influe grandement sur la crédibilité de la note, et par extension, de son auteur. Un manque de rigueur sur la forme est donc perçu par le groupe comme équivalent à un manque de sérieux ou de compétences sur le fond. Reconnue comme un marqueur social important, la mauvaise orthographe est largement associée à l'origine sociale et au niveau d'éducation. Cet élément agit donc ici comme un filtre de distinction fort entre ceux qui maîtrisent les codes orthographiques et les autres. L'espace des notes de la communauté comporte la particularité d'un anonymat complet dans lequel la seule

information visible sur un contributeur est un pseudonyme généré aléatoirement par le système. L'orthographe, la qualité syntaxique et la richesse lexicale sont donc les seuls éléments à disposition des contributeurs pour situer socialement un membre de leur groupe. Dans un espace structuré par la notion d'évaluation, ces éléments deviennent prépondérants dans le choix d'évaluer une note comme utile ou non. Ils créent une forme d'exclusion des membres n'adoptant pas les normes et les attentes du groupe majoritaire. Si cette exclusion n'est pas systématique et explicite, elle participe néanmoins à la création d'un espace élitiste, réservé à ceux qui maîtrisent les normes linguistiques dominantes. L'influence de l'orthographe sur les hiérarchies sociales a été abondamment étudiée. Pierre Bourdieu en fait une composante du capital culturel dans son ouvrage *La Distinction,* dans lequel il explique qu'elle contribue à la reproduction des hiérarchies sociales[68]. L'orthographe agit donc ici comme un critère social d'acceptation dans la communauté des rédacteurs. L'utilité perçue d'une note dépend donc aussi de la capacité de son rédacteur à produire un contenu en adéquation avec les standards d'une écriture légitime.

2. Une hiérarchie des stratégies argumentatives qui valorise la réfutation explicite

Après avoir étudié l'influence du ton et du registre sur la probabilité de validation d'une note, nous allons maintenant nous intéresser aux stratégies argumentatives employées.

a. Dénoncer une manipulation est une forme de correction hautement valorisée

Parmi les stratégies argumentatives utilisées, celles qui sont les plus fréquemment reconnues comme utiles par la communauté sont celles qui dénoncent explicitement une forme de manipulation. Plus précisément, 83,37 % des notes qui dénoncent un plagiat ou une motivation cachée de l'auteur, comme de la publicité déguisée, sont jugées utiles dans notre corpus. De même, 76,18 % des notes révélant une manipulation d'image ou de vidéo sont jugées utiles.

[68] Bourdieu, P. (1979). *La distinction : Critique sociale du jugement*. Paris : Éditions de Minuit.

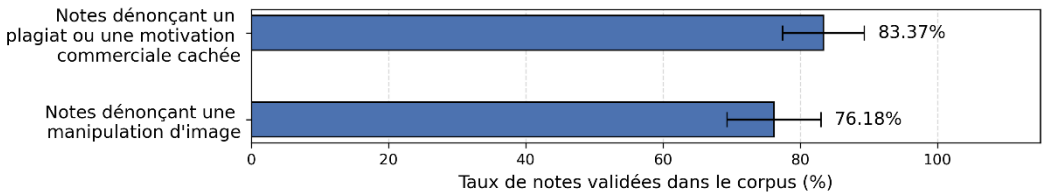

Figure n°15 : Taux de notes « utiles » dans le corpus parmi celles dénonçant des manipulations (Source : établi par l'auteur à partir des données collectées sur X.com)

Dans ces deux cas, les taux obtenus sont bien supérieurs à la moyenne. Pour rappel, 50 % du corpus est constitué de notes jugées utiles et les 50 % restants de notes jugées inutiles. Ce type de stratégie argumentative se distingue par le fait que le débat est déplacé d'une correction de fait vers le dévoilement d'une tromperie ou d'une intention cachée. En plus d'une fonction de correction, ces notes démontrent les intentions cachées de manipulation de l'auteur de la publication, ce qui constitue un poids argumentatif supplémentaire. Le fort taux de validation de ces notes dans le corpus peut se comprendre par le besoin des contributeurs de dévoiler au grand jour un élément de manipulation caché d'une publication. Au-delà de simplement dévoiler une erreur, il semble que révéler une manipulation volontaire soit perçu comme la mise au jour d'une transgression de l'ordre moral. Il est ainsi beaucoup plus consensuel dans la communauté des contributeurs de valider ce type de note, d'autant plus qu'elles échappent généralement à des logiques partisanes ou idéologiques. Dans le cas d'une manipulation d'image, la correction peut même être perçue comme purement technique et donc sujette à peu de débats. Les démonstrations employées sur ce genre de contenu sont généralement simples, se contentant de mobiliser des preuves claires et irréfutables, comme la source d'un contenu plagié. Ces corrections se caractérisent donc par leur clarté. Il ne s'agit pas de devoir nuancer, contextualiser ou interpréter. La contradiction est claire et indiscutable. En révélant objectivement une tromperie, ces notes s'alignent parfaitement avec les normes implicites du système en proposant aux utilisateurs un contexte important de manière parfaitement neutre et impartiale.

b. La contradiction frontale est plus valorisée que la contextualisation

Toutes les corrections de publications ne révèlent pas de manipulations ou d'intentions cachées. Beaucoup de notes ont pour objet de rectifier une information fausse

ou trompeuse. Pour ce faire, deux stratégies argumentatives se détachent dans les notes. La première est de contextualiser l'information présentée dans la publication sans la contredire directement. La seconde est de venir réfuter frontalement les éléments énoncés dans le tweet. Les notes adoptant une stratégie de contextualisation sont majoritaires dans le corpus, représentant 60,67 % des corrections. Les contradictions frontales sont moins représentées avec seulement 31,33 % des notes. Malgré cette sous-représentation, ce type de stratégie argumentative arrive plus fréquemment à convaincre la communauté de son utilité, puisque 65,96 % d'entre elles sont jugées positives dans le corpus.

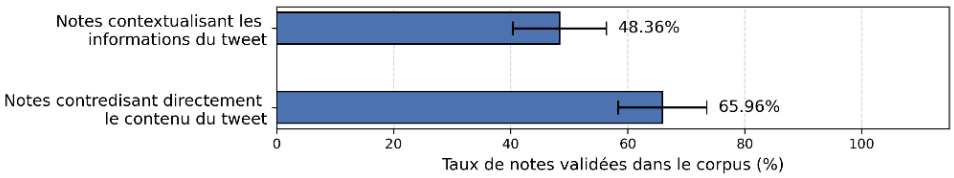

Figure n°16 : Taux de notes « utiles » dans le corpus selon leur méthode argumentative
(Source : établi par l'auteur à partir des données collectées sur X.com)

Les meilleurs taux de validation des notes contredisant clairement le contenu d'une note peuvent signifier que la communauté valorise davantage la clarté des intentions et les prises de positions nettes. Dans un espace où les contributeurs évaluent de nombreuses notes chaque jour, le fait de nommer explicitement une erreur apparait comme une norme implicite valorisée. À l'inverse, les contextualisations plus prudentes peuvent apparaitre comme plus ambiguës. Une part des évaluateurs peut sous-estimer la nécessité d'une correction si l'erreur n'est pas explicitement nommée. Cette stratégie argumentative a cependant l'avantage d'apparaitre comme plus neutre et d'éviter les conflits. Sa surreprésentation dans le corpus peut indiquer un décalage entre la perception par les contributeurs des normes implicites d'évaluation des autres contributeurs et celles qu'ils appliquent réellement. Alors qu'un rédacteur va privilégier la mesure et la prudence dans ses corrections, par peur d'apparaitre comme partial par ses pairs, les évaluateurs vont au contraire plutôt rejeter ce type de correction. Le compromis entre prudence et efficacité de correction apparait délicat à trouver pour de nombreux rédacteurs. La stratégie argumentative semble donc constituer une zone de flou normatif au sein des contributeurs, même si la confrontation directe reste plus valorisée par la communauté.

3. Les ressources extérieures légitiment les corrections

Pour terminer ce chapitre sur la notion d'utilité d'une note qui émerge du système, nous pouvons étudier les ressources extérieures que les rédacteurs des notes mobilisent. Le système encourage fortement les rédacteurs à inclure un lien vers une source à chaque note, ce que les rédacteurs font en très grande majorité. Nous allons donc plutôt étudier ici les références mobilisées explicitement dans l'argumentaire. Cette section abordera les références à une autorité, l'emploi d'un vocabulaire expert et l'évocation explicite de textes de loi.

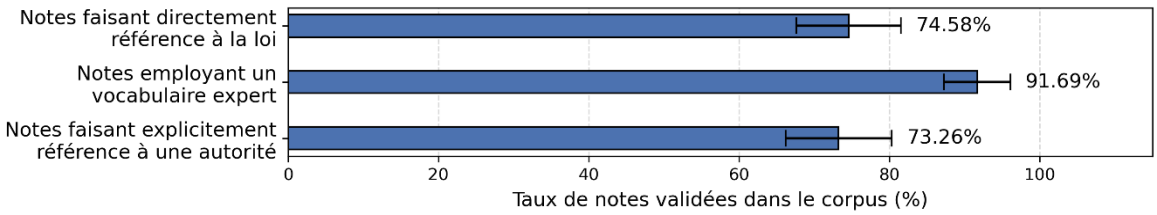

Figure n°17 : Taux de notes « utiles » dans le corpus selon le type d'autorité explicitement évoqué (Source : établi par l'auteur à partir des données collectées sur X.com)

a. Les références à des autorités constituent des marqueurs de crédibilité

Mobiliser explicitement une autorité reconnue dans sa note, comme une institution, un scientifique ou un expert, semble grandement favoriser l'approbation d'une note dans l'espace correctif. Près de 73 % des notes qui en comportent sont jugées utiles dans le corpus. Ce résultat peut s'expliquer par l'anonymat complet du système des notes de la communauté. Il ne permet pas aux contributeurs d'acquérir une légitimité aux yeux de leurs pairs, donc s'appuyer sur une source déjà connue dans l'espace public permet de s'approprier celle que cette source représente. Citer une information ou un point de vue déjà validé dans l'espace public semble faciliter sa validation dans l'espace des notes de la communauté. Cela permet de s'extraire du registre de l'opinion pour simplement restituer une position déjà admise dans des champs soumis à la méthode scientifique ou à l'éthique journalistique par exemple. L'espace des notes de la communauté étant anonyme et ouvert à tous, un contributeur ne peut pas, par lui-même, constituer la même crédibilité qu'un scientifique ou un expert qui engage sa crédibilité dans ses prises de position. Les références à des autorités externes viennent ainsi compenser la limite principale du système : l'anonymat. Cette limite implique qu'un contributeur ne peut jamais finir par être reconnu par ses pairs pour la qualité de son

travail et ne peut pas non plus engager sa crédibilité publique dans ses propos. Les références à des organisations ou à des individus provenant de champs dans lesquels il n'y a pas d'anonymat structurel permettent donc de compenser ces limites. De plus, citer une autorité externe constitue un gain de temps pour les évaluateurs de la note. Une référence à une autorité reconnue déplace l'acte d'évaluation d'une estimation de la véracité d'un propos vers une estimation du degré de crédibilité d'un acteur tiers. Citer une autorité reconnue permet également de s'extraire d'un espace conflictuel et polarisé pour faire appel au savoir commun. C'est donc faire appel à une forme d'universalisme de la correction, qui s'inscrit dans le savoir partagé plutôt que dans des logiques partisanes. Les références à des autorités extérieures constituent donc un marqueur important de crédibilité compensant le manque de légitimité des contributeurs anonymes et facilitant le travail des évaluateurs.

b. L'emploi de vocabulaire technique pour signaler la compétence

Parmi les notes analysées, celles qui emploient un vocabulaire expert, c'est-à-dire des termes et expressions spécialisées issues par exemple des domaines scientifiques, économiques, juridiques ou médicaux, affichent des taux de validation particulièrement élevés, avec 91,69 % de ces notes jugées utiles dans notre corpus. L'emploi de ce type de termes et d'expressions semble agir ici comme une preuve de maitrise du sujet. Le fait de connaitre et d'employer spontanément des termes experts semble apparaître aux yeux des évaluateurs comme un indice de compétence du rédacteur. Dans un espace caractérisé par l'anonymat complet des contributions, l'emploi de termes techniques peut permettre d'apparaitre comme un spécialiste du sujet, avant même que le contenu de la note ne soit examiné. L'emploi de ces termes constitue un mécanisme de distinction dans un espace caractérisé par des publications partiales, informelles et polémiques. Les contenus visés par une note de la communauté sont particulièrement sujets à des prises de positions approximatives sur des sujets complexes. L'emploi d'un lexique très soutenu permet de créer une distance symbolique entre la publication et la note corrective. L'auteur de la note se place dans une position dominante par rapport à l'auteur du tweet en démontrant sa maîtrise de termes techniques. Le lexique prend ici le rôle d'un marqueur visible de crédibilité d'une annotation. Dans un environnement où les contributeurs évaluent des dizaines de notes par jour, ce raccourci de crédibilité semble particulièrement valorisé. Utiliser des termes particulièrement techniques dans une correction sans que ce soit forcément nécessaire n'est pas absolument dévalorisé par une communauté qui doit rapidement décider de l'utilité ou

non d'une note. En résumé, la communauté valorise l'emploi d'un lexique expert par les rédacteurs, perçu comme témoignant de l'utilité d'une note et non comme une forme de distinction superficielle.

c. Le droit comme levier de validation communautaire

Le dernier type de ressource extérieure mobilisé dans les notes que nous allons étudier est le droit. Il est en effet fréquent qu'une note comporte l'évocation d'un article de loi, d'un règlement ou d'une norme. Les résultats de l'étude de corpus montrent que, comme pour les deux ressources précédentes, le droit semble grandement augmenter la probabilité de validité d'une note. Ainsi, 74,58 % des notes du corpus faisant référence explicitement au droit font partie des notes jugées utiles, ce qui suggère que le droit agit comme une ressource de légitimation importante. Ce résultat peut s'expliquer par le fait que le droit apparait comme parfaitement neutre et extérieur au débat, loin de tout clivage idéologique ou parti pris. Il est également un outil de clôture argumentative dans la mesure où l'appel à une loi ou une norme peut donner l'impression d'un argument final et irréfutable. L'objectif de l'invocation d'un texte de loi n'est pas de discuter ou d'ouvrir le débat, mais bien de trancher définitivement. Dans ce cadre, annoter une publication ne revient pas seulement à contredire les propos tenus, mais également à les qualifier juridiquement. Le rédacteur transpose ainsi le débat dans le champ juridique, dans lequel chaque élément est soit licite soit illicite. Le droit semble être perçu par la communauté comme un méta-cadre dans lequel peuvent s'inscrire les corrections pour caractériser sans ambiguïté une publication trompeuse. Il semble agir comme le plus petit dénominateur commun entre tous les contributeurs, et la communauté semble partager la norme implicite de ne jamais le remettre en question. Cet emploi du droit peut sembler paradoxal dans un espace au sein duquel les notes de la communauté jouent le rôle de régulateur en l'absence d'une autorité pour imposer des règles et des pénalités. L'évocation du droit ne sert donc pas à contraindre ou sanctionner directement par celui-ci, mais est simplement utilisée pour justifier une correction. En résumé, la forte validation des notes évoquant le droit révèle que la communauté valorise la mobilisation de cadres normatifs extérieurs perçus comme indiscutables, transformant la correction en simple qualification juridique.

Pour conclure cette section, l'analyse des critères de validation des notes de la communauté révèle un ensemble de normes implicites partagées par une majorité de contributeurs. Adopter un ton détaché et neutre semble être la norme dominante chez les contributeurs. L'espace correctif valorise également le contenu mesuré et maîtrisé, en excluant toute note inappropriée. Les contributeurs n'échappent pas aux normes implicites les plus courantes des champs contributifs, comme le standard d'une écriture légitime ou la valorisation de l'emploi de vocabulaire technique. La structure du dispositif, par ses interfaces de vote rapide et son anonymat structurel, semble aussi valoriser des normes de dénonciation frontale des tromperies et d'utilisation d'autorités externes. Les normes implicites des contributeurs émergent donc d'une combinaison complexe de facteurs. Elles résultent de l'appropriation individuelle des consignes du dispositif, de dynamiques sociales typiques des environnements collaboratifs, de l'effet des choix techniques du système correctif et des opinions individuelles. Les normes qui en résultent semblent avoir des effets ambivalents sur les performances du système. Certaines renforcent les principes de neutralité et de rigueur demandés par la plateforme, là où d'autres les remettent en question ou les redéfinissent. Si ces normes participent à l'efficacité globale du dispositif, elles ne semblent pas en être la cause principale. L'architecture globale du système, combinant choix individuels et mécanismes complexes de validation collective, semble constituer le principal moteur de l'efficacité mesurée.

Section 3. Une annotation guidée d'abord par le statut de l'émetteur de la publication et la sensibilité de son contenu

L'ensemble des résultats de la section précédente ne peut s'appliquer à la correction d'un tweet trompeur que si cette correction a été proposée par un contributeur. Dans un espace sur lequel près de 6000 tweets sont publiés chaque seconde, la principale condition pour qu'une publication soit corrigée est qu'un contributeur prenne l'initiative d'en rédiger une. Les volumes de publications du réseau imposent donc aux contributeurs de cibler une part infime du contenu dans leur rédaction de notes. Nous allons analyser dans cette section les dynamiques à l'œuvre dans ce choix.

1. L'incitation à corriger dépend de la légitimité perçue de l'auteur

La première caractéristique d'une publication pouvant influencer le choix d'un contributeur d'y apposer une annotation est son auteur. À partir de notre corpus constitué de 150 notes, dont la moitié a obtenu le statut « utile » et l'autre moitié le statut « non utile », nous avons analysé la répartition des publications annotées par type de compte. Nous avons également, pour chaque type de compte, mesuré le taux de notes ayant obtenu le statut utile au sein du corpus.

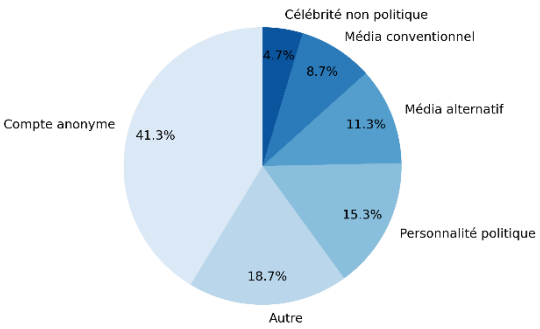

Figure n°18 : Type de compte visé par une note de la communauté dans le corpus (Source : établi par l'auteur à partir des données collectées sur X.com)

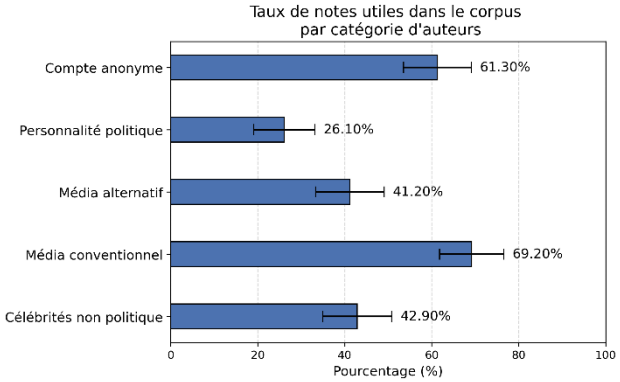

Figure n°19 : Taux de notes validées dans le corpus suivant le type de compte (Source : établi par l'auteur à partir des données collectées sur X.com)

a. Un anonymat qui expose davantage à la correction

Le premier élément qui ressort de ces graphiques est la présence massive de publications postées par des comptes anonymes dans le corpus, avec près de 41,3 % des tweets présents. Ce résultat indique une exposition importante de ces comptes au dispositif de correction, qui apparaissent comme une cible privilégiée des membres de la communauté. Cela peut s'expliquer par le fait que l'anonymat renforce la perception d'un risque informationnel dans la mesure où l'auteur de la publication engage moins sa responsabilité dans ses propos qu'une personne agissant sous sa vraie identité. La communauté considère donc probablement qu'il est plus acceptable, voire nécessaire, de corriger le contenu émanant de ce type de compte. De plus, près de 61,3 % des annotations de comptes anonymes présentes dans le corpus ont obtenu le statut utile dans le corpus, ce qui est beaucoup plus élevé que les taux de validation des annotations de tweets de personnalités publiques par exemple. Il semble donc plus facile d'obtenir des consensus de la communauté sur l'utilité d'une note quand elle porte sur un compte anonyme. Cela peut s'expliquer par le fait que corriger ce type de compte implique un coût symbolique et relationnel moindre pour le contributeur. Les corrections de comptes sans affiliation professionnelle, politique ou médiatique visible génèrent moins de controverses ou de confrontations symboliques. Un compte qui ne renvoie pas non plus à un collectif reconnu ou à un groupe d'individus pourra également moins exposer le contributeur à un soutien communautaire organisé. De plus, l'absence de capital symbolique et de légitimité sociale de l'auteur d'une publication rend la correction de ses propos plus facilement acceptée par les autres membres de la communauté. La correction est dans ce cas davantage perçue comme neutre et technique, ce qui peut expliquer que les taux de notes jugées « utiles » sur ce type de compte soient plus élevés que sur des comptes de personnalités. L'anonymat est ainsi à la fois un facteur d'exposition plus fort au dispositif de correction et un facteur de réception plus favorable de l'intervention du contributeur, car il diminue la charge symbolique du fait de corriger.

b. Des personnalités publiques corrigées avec retenue

Dans le corpus étudié, les personnalités politiques ne représentent que 15,3 % de l'ensemble des tweets annotés. Il est difficile d'interpréter ce résultat dans la mesure où, même si les comptes de personnalités politiques ne représentent qu'une infime proportion de l'ensemble des comptes, ils bénéficient d'une très grande exposition et se retrouvent très fréquemment au cœur de polémiques sur le réseau. Cependant, le fait que ces comptes soient trois fois plus souvent annotés que les comptes de personnalités publiques non politiques semble indiquer leur exposition importante à la correction communautaire. Malgré cette plus forte exposition, la reconnaissance des corrections est bien plus faible pour les personnalités politiques que pour les autres personnalités publiques, avec seulement 26,1 % de notes du corpus validées. Les figures politiques sont donc plus souvent corrigées, mais les corrections sont moins souvent validées, alors que les personnalités du monde du divertissement sont moins ciblées, mais les corrections sont plus reconnues par la communauté. Ce résultat peut s'expliquer par la différence d'autorité symbolique entre ces deux types de personnalités publiques. Les célébrités non politiques sont des artistes, influenceurs et célébrités médiatiques qui appartiennent majoritairement au monde du divertissement. Elles incarnent donc beaucoup moins une autorité institutionnelle et décisionnelle que les personnalités politiques. Leurs publications sont donc certainement perçues par la communauté comme moins engageantes idéologiquement et leurs mensonges ou erreurs comme ayant moins de conséquences. De plus, la nature d'autorité plus diffuse des personnalités médiatiques rend une correction moins conflictuelle au sein de la communauté de correcteurs. La parole politique incarne quant à elle des positions idéologiques précises et des appartenances partisanes claires. Corriger une personnalité politique engage donc fortement le correcteur dans une position qui peut être perçue par la communauté comme partisane. L'aspect politique d'une correction la rend donc plus nécessaire aux yeux de la communauté, mais aussi plus difficilement acceptée par les contributeurs. L'utilité d'une correction ne dépend donc pas que de la présence publique de l'auteur, mais également du type d'autorité qu'il incarne et de la charge idéologique de son discours.

c. Des médias alternatifs plus ciblés que les médias traditionnels, protégés par une légitimité institutionnelle

Une partie des notes de la communauté concerne des publications de médias. Environ 9 % des tweets présents dans le corpus concernent des médias institutionnels reconnus et 11,3 % se réfèrent à des médias que nous qualifions de « médias alternatifs ». Ces médias regroupent des comptes s'auto-qualifiant de « média », et des collectifs se revendiquant comme tels, sans journaliste accrédité. Nous remarquons que le nombre de notes proposées sur des publications de médias alternatifs est plus important que sur des médias reconnus. Ce résultat peut s'expliquer par le fait que ces comptes, souvent dénués de contrôle éditorial et généralement militants, relaient plus d'informations perçues comme trompeuses par la communauté. Leur statut non institutionnel et le fait que leurs publications soient souvent massivement relayées dans les sphères militantes peuvent également expliquer l'attention particulière portée à ces comptes par les contributeurs. Malgré cette plus forte exposition aux corrections communautaires, nous observons que les annotations des médias alternatifs sont bien moins souvent jugées utiles que celles des médias traditionnels. Il y a donc plus de désaccord et d'incertitude au sein de la communauté d'évaluateurs sur ces notes. Ce résultat peut s'expliquer par la plus grande polarisation du public des comptes de médias alternatifs et la teneur idéologique plus marquée du contenu des publications de ces comptes. Corriger un média alternatif peut être plus souvent perçu par la communauté comme une mise en cause controversée et une attaque idéologique de la part du correcteur. De plus, corriger un média reconnu correspond à s'inscrire dans un cadre de confiance éditoriale déjà régulé et à venir apporter une précision ou un ajustement ponctuel de l'information. À l'inverse, une correction d'un média alternatif peut s'apparenter à une disqualification de l'ensemble de la source, perçue comme non professionnelle ou fondamentalement trompeuse. La perception de la correction par une partie de la communauté comme une mise en cause de la légitimité du compte peut expliquer que les annotations des médias alternatifs soient plus conflictuelles et donc moins fréquemment validées. Encore une fois, le statut perçu de la source influe significativement sur la validation des corrections, ce qui révèle une hiérarchie implicite de la légitimité. Le statut perçu des médias traditionnels les protège donc partiellement de la correction, mais rend la correction plus acceptée par la communauté.

2. Un arbitrage délicat des affrontements idéologiques dans un espace polarisé

Après avoir étudié l'effet du statut perçu de l'émetteur du tweet sur la présence d'une note de la communauté, nous allons désormais nous intéresser au contenu de la publication annotée. L'analyse du contenu des tweets nous permet de classifier ces publications en grandes catégories.

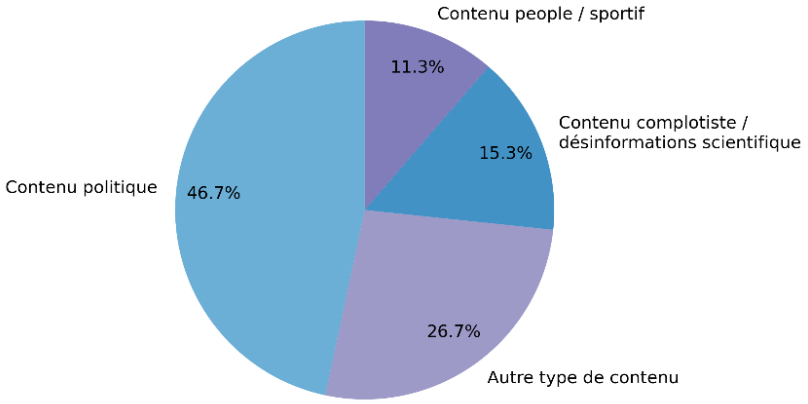

Figure n°20 : Type de contenu des publications annotées dans le corpus
(Source : établi par l'auteur à partir des données collectées sur X.com)

a. Le contenu politique comme foyer principal de mobilisation corrective

Il ressort de cette analyse que le contenu politique est de loin le premier registre thématique du corpus, avec près de 46,7 % des publications. Il apparait clair que les contributeurs mobilisent leurs efforts de correction prioritairement dans le champ politique, loin devant d'autres domaines, comme la désinformation scientifique ou le divertissement. Cette mobilisation massive peut s'expliquer par la perception que les contenus politiques ont un impact social important, nécessitant impérativement une correction. Cela peut également s'expliquer par l'engagement idéologique de certains contributeurs. Même si le système de notes de la communauté encourage des principes de neutralité et de pluralisme, il est probable qu'une part significative des contributeurs soit motivée dans ses corrections par des convictions personnelles fortes. Dans ce cadre, corriger une publication dont le contenu défend une position d'un camp opposé peut répondre à un besoin de rétablir une forme de

vérité, et donc motiver le contributeur à proposer une correction. La correction devient alors un instrument de confrontation idéologique, en plus d'un instrument de rectification.

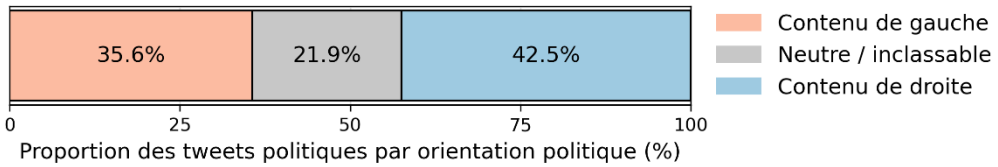

Figure n°21 : Orientation idéologique du contenu des publications annotées du corpus parmi celles qui évoquent un sujet politique (Source : établi par l'auteur à partir des données collectées sur X.com)

La répartition par affiliation idéologique des tweets du corpus comportant un contenu politique montre un certain équilibre entre les publications classées à droite et celles identifiées comme de gauche, avec respectivement 42,5 % et 35,6 % des tweets politiques annotés. Ce relatif équilibre signifie qu'aucun camp politique n'est victime d'un ciblage massif de la part des notes de la communauté. Il peut s'interpréter par une norme de pluralisme au sein d'une partie des contributeurs, qui corrigent les informations trompeuses provenant de tous les bords politiques. Il peut également s'interpréter par la présence d'une diversité idéologique au sein des membres des notes de la communauté. La présence d'une telle diversité pourrait expliquer à elle seule une correction d'information des deux bords politiques. Finalement, cette dynamique peut aussi traduire une volonté collective de maintenir la crédibilité du système correctif, qui serait mise à mal par un ciblage plus fréquent d'un bord politique.

b. <u>Une correction du contenu de droite perçu par la communauté comme plus légitime que corriger la gauche</u>

Malgré cet apparent pluralisme dans la répartition des annotations, une analyse du taux de validation au sein de notre corpus nous montre un contraste net entre les notes corrigeant des tweets de droite et celles annotant des publications de gauche.

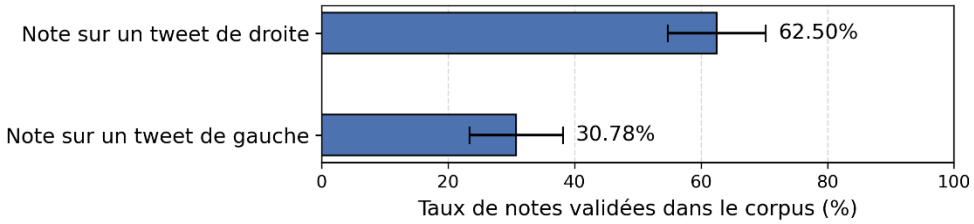

Figure n°22 : Taux de validation des notes selon la classification idéologique du contenu de la publication qu'elles corrigent
(Source : établi par l'auteur à partir des données collectées sur X.com)

Cette asymétrie suggère que la légitimité perçue d'une correction par la communauté peut dépendre de la position idéologique de la publication, en valorisant les corrections produites sur des publications classées à droite. Plusieurs hypothèses peuvent éclairer ces résultats, comme le fait que la communauté adopte un cadrage implicite du contenu de droite comme plus propice à la désinformation. Il peut être également perçu comme plus légitime socialement par les membres de la communauté d'intervenir sur un contenu d'extrême droite si ce contenu peut être interprété comme constituant une menace à l'encontre de certains groupes sociaux. À l'inverse, les corrections visant des publications de gauche rencontrent plus de difficultés à être approuvées par la communauté. Cela peut révéler une perception plus polémique de ce type d'annotation, qui est plus facilement perçue comme révélatrice d'une prise de position. Il est également possible que les notes visant des publications orientées à gauche comportent moins de prudence argumentative et adoptent plus fréquemment des tons polémiques et militants, ce qui nuirait à leur approbation par une communauté qui vise la neutralité idéologique. Finalement, il reste probable qu'un biais idéologique structurel des contributeurs les conduise à adopter une plus grande tolérance dans leurs évaluations des annotations de contenu de droite. L'utilité corrective ne repose donc pas uniquement sur le simple contenu d'un tweet, mais est aussi influencée par son orientation idéologique et par les rapports de forces implicites au sein des membres des notes de la communauté.

3. Corriger pour distinguer la désinformation sérieuse des contenus ludiques ou secondaires

Le contenu politique, même s'il constitue près de la moitié du contenu corrigé par les notes dans notre corpus, n'est pas la seule catégorie de publication sur laquelle la communauté juge utile de proposer une correction. Dans cette partie, nous allons nous intéresser aux deux autres catégories que nous avons identifiées au début de cette partie, à savoir le contenu complotiste et les publications liées au monde du divertissement et de l'humour. L'intérêt de l'étude de ces deux catégories réside dans le fait qu'elles sont respectivement le cœur de cible du système et une zone grise du dispositif.

a. La désinformation et le complotisme mobilisent massivement

Les notes proposées sur des publications complotistes ou de désinformation scientifique représentent près de 15 % de notre corpus. Ce niveau de présence montre que ce type de contenu est source d'une mobilisation importante, sans être la catégorie la plus représentée dans le corpus. L'attention qui lui est portée par la communauté peut s'expliquer par le danger perçu qu'il représente pour les utilisateurs de la plateforme. Les notes réalisées sur cette catégorie de contenu obtiennent un taux de validation dans le corpus très élevé de près de 74 %, ce qui en fait la catégorie avec le plus haut taux de notes affichées.

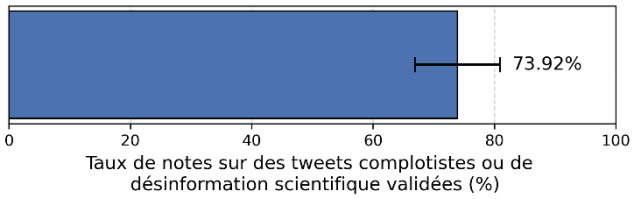

Figure n°22 : Taux de validation des notes du corpus corrigeant une publication complotiste
(Source : établi par l'auteur à partir des données collectées sur X.com)

Ce taux montre un fort consensus de la communauté dans la correction de la désinformation scientifique et du complotisme. Cela peut s'expliquer par le fait que la désinformation scientifique constitue le cœur de cible du système des notes de la communauté, favorisant la consensualité des contributeurs dans la nécessité d'annoter ce contenu et dans la validation des annotations proposées. Le cadre normatif du dispositif de

correction rend difficilement défendable un contenu factuellement erroné ou ouvertement mensonger. La désinformation et le complotisme représentent ainsi un point de convergence normatif des membres de la communauté sur ce qui mérite d'être corrigé. Les différences idéologiques des membres s'effacent dans le cadre de leur correction et les contributeurs se rangent naturellement derrière une norme commune de véracité et de neutralité. La notion d'utilité se crée ici dans un cadre peu conflictuel où la communauté converge vers une définition partagée de ce qui doit être corrigé.

b. Le contenu de divertissement et humoristique neutralise l'efficacité corrective

La dernière grande catégorie de publication sur laquelle des notes sont proposées est celle de l'humour ou du divertissement. Le divertissement représente, par exemple, les tweets évoquant des sujets sportifs ou ceux en lien avec des célébrités. Ces publications constituent une zone grise pour le système des notes de la communauté. Le cadrage normatif explicite prévoit que le contenu clairement humoristique ne doit pas faire l'objet d'une note. Il est cependant difficile d'établir une frontière claire entre un contenu clairement humoristique et un autre utilisant l'humour pour faire passer des fausses informations problématiques.

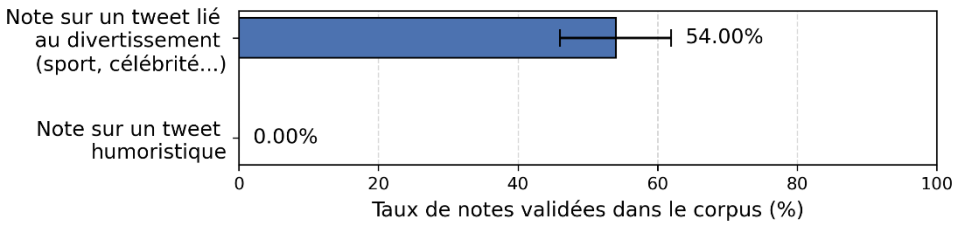

Figure n°23 : Taux de validation des notes du corpus corrigeant une publication humoristique ou liée au divertissement
(Source : établi par l'auteur à partir des données collectées sur X.com)

Une analyse des taux de validation pour ces deux catégories montre que près de la moitié des annotations de publications liées au monde du divertissement sont jugées utiles. Aucune note du corpus proposée sur un tweet humoristique n'a cependant obtenu le statut d'utile. Ces catégories de publication constituent donc la frontière de la notion d'utilité d'une annotation construite par la communauté. Cette utilité se heurte parfois à l'ambiguïté de

l'intention de l'auteur d'une publication. Différencier l'humour d'une volonté de désinformer ou de tromper les utilisateurs dépend du cadre interprétatif qu'adopte le contributeur. Cela implique également que les annotations sur ce type de contenu peinent à être consensuelles au sein de la communauté, et donc à obtenir le statut de note utile. Ces catégories de publication révèlent donc les limites d'un système qui impose une lecture littérale des publications pour identifier les tromperies. Le mécanisme correctif échoue donc à produire des annotations pertinentes et consensuelles quand le registre de la publication sort d'une interprétation littérale et utilise le second degré ou la parodie.

Pour conclure cette section, nous avons démontré que la probabilité d'une correction varie fortement selon l'auteur de la publication associée et le registre qui y est employé. Les comptes anonymes sont largement ciblés et leurs corrections majoritairement approuvées grâce à la faible charge symbolique de l'acte de correction. Les personnalités politiques sont aussi largement ciblées, mais la charge idéologique de leurs publications réduit grandement la consensualité de leurs corrections. Les médias traditionnels sont plus protégés que les médias alternatifs par leur légitimité institutionnelle, mais leurs corrections entrainent moins de débats et de désaccords au sein des contributeurs. Le contenu politique est massivement privilégié par les contributeurs. Des publications provenant de tout le spectre politique sont annotées, mais les annotations sur des tweets de droite sont beaucoup plus fréquemment jugées utiles que celles classées à gauche, signe de biais idéologiques implicites parmi les contributeurs. La perception de l'utilité d'une annotation sur une publication peut être consensuelle au sein de la communauté, comme dans le cas de la désinformation scientifique, mais aussi très différenciée, comme dans le cas des contenus humoristiques. Ce contenu humoristique et celui lié au divertissement semblent constituer une zone grise, échappant partiellement aux mécanismes correctifs. Les corrections ne sont donc pas évaluées indépendamment de leur contexte d'énonciation, mais sont au contraire influencées par des hiérarchies implicites de légitimité et des biais idéologiques. Le type d'émetteur et la nature de la publication influencent donc la représentation de ce qui constitue une correction nécessaire et acceptable.

Chapitre II. La fabrication d'un consensus correctif par la structuration algorithmique au détriment du pluralisme

Introduction du chapitre

Après avoir montré que l'utilité des annotations se construit à partir des choix individuels des contributeurs, nous allons élargir notre analyse à l'étude des structures algorithmiques et sociales d'ensemble du dispositif. Si l'utilité d'une note apparait comme résultant d'une coordination spontanée d'acteurs, nous allons montrer que cette perception masque un ensemble de mécanismes algorithmiques et collectifs qui encadrent ces contributions. Là où, dans le premier chapitre, nous avons étudié en détail cette notion d'utilité collective émergeant des choix individuels, dans ce chapitre nous nous concentrerons sur la notion de consensualité. En effet, les choix individuels des contributeurs, même s'ils s'inscrivent dans des normes implicites qu'ils peuvent être un grand nombre à partager, doivent être agrégés pour décider de l'utilité d'une note. Le bon fonctionnement du système repose ainsi tout autant sur les normes contributives individuelles de neutralité et de rigueur que sur la capacité du système à convertir ces contributions individuelles en un résultat qui sera perçu comme neutre et rigoureux par un plus grand nombre. Ce sera donc l'analyse détaillée des choix algorithmiques d'agrégation des contributions individuelles qui nous permettra de comprendre pleinement la manière dont le dispositif transforme les clivages en consensus. En plus d'influer sur les mécaniques de validation des notes, nous verrons que ces choix techniques influent fortement sur la structuration sociale des contributeurs. Derrière une apparence de stricte égalité de contribution se cachent de très fortes disparités de pratiques contributives et de statuts. Ce chapitre sera donc l'occasion de compléter notre compréhension de l'apparent bon fonctionnement du système correctif, mais également de comprendre ce qui est sacrifié pour l'obtenir.

Section 1. Un consensus transversal construit par l'algorithme au détriment de la délibération

1. Une classification des notes qui produit un filtrage massif

a. Une majorité de notes peinent à atteindre un statut définitif

Comme évoqué en introduction et dans le premier chapitre, ce sont les votes des membres de la communauté qui décident de l'affichage ou non d'une note. Si les votes de la communauté sont favorables, l'algorithme va attribuer dans sa base de données à la note le statut de « note utile » et elle sera affichée à tous les utilisateurs du réseau. Dans le cas contraire, si l'algorithme détermine que la communauté rejette la note, le statut de « note inutile » lui sera affecté, et elle ne sera plus recommandée aux évaluateurs dans l'interface des notes en attente de notations. Dans le cas où l'algorithme n'a pas assez d'éléments pour décider de l'utilité ou de l'inutilité d'une note, il lui attribue le statut « en attente de plus d'évaluations ». Comme nous l'avons vu dans l'état de l'art, la majorité des notes obtiennent leurs évaluations durant les heures qui suivent leur publication. Cependant, il arrive souvent que plusieurs jours après avoir été soumise à la communauté, une note ne recueille pas assez de votes pour que l'algorithme puisse décider de lui affecter le statut de note utile ou de note inutile. Cette note conserve donc le statut « en attente de plus d'évaluations » et ne sera jamais affichée. L'accès à la base de données nous permet d'évaluer la répartition du nombre de votes que reçoivent les notes pour chacun des trois statuts définitifs : utile, non utile et en attente. Pour cela, nous avons sélectionné aléatoirement cent notes de chaque catégorie ayant été proposées aux votes au moins un mois avant le téléchargement de la base de données. Nous avons ensuite récupéré, pour chacune, le nombre d'évaluations qu'elles ont recueillies. Nous pouvons représenter la répartition du nombre d'évaluations par un diagramme en boîte, qui permet de visualiser les valeurs extrêmes, les premiers et troisièmes quartiles et la médiane du jeu de données obtenu :

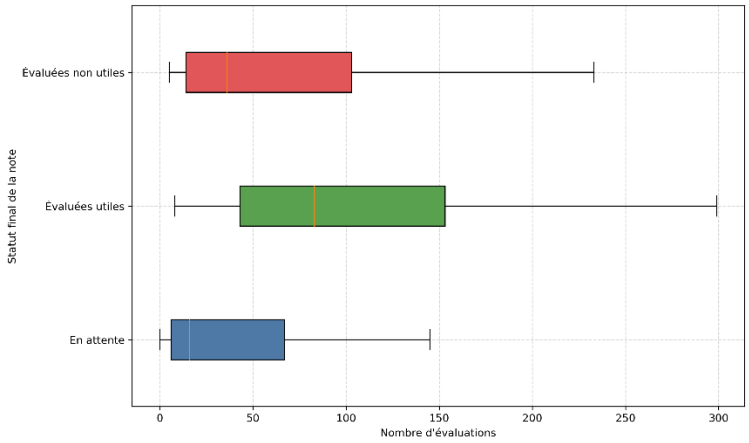

Figure n°24 : Diagrammes en boîtes de la répartition du nombre de votes pour chaque statut définitif
(Source : établi par l'auteur à partir des données téléchargées sur X.com)

Nous remarquons une différence notable entre la répartition du nombre de votes des notes « en attente » et celle des notes « utiles » et « non utiles ». Cependant, ce critère ne semble pas expliquer à lui seul ce statut « en attente », dans la mesure où une proportion importante de ces notes a un nombre de votes identique ou supérieur à des notes auxquelles a été attribué un statut définitif « utile » ou « non utile ». Il semble donc techniquement possible de configurer le système de manière à attribuer un statut définitif à une part importante des notes « en attente » en abaissant les seuils de choix de l'algorithme. Le choix technique de n'affecter un statut définitif à une note que quand des niveaux de certitude très élevés dans la pertinence de ce statut ont été atteints a donc été retenu par les concepteurs du système.

b. Une rareté structurelle des notes affichées

Pour étudier les conséquences de ce choix technique sur la répartition du statut final des notes, nous pouvons visualiser cette répartition grâce à la base de données. Pour cela, nous ne considérerons que les notes qui ont été soumises aux évaluations de la communauté avant le 1er septembre 2024, de manière à garantir qu'elles aient eu le temps d'obtenir un statut définitif.

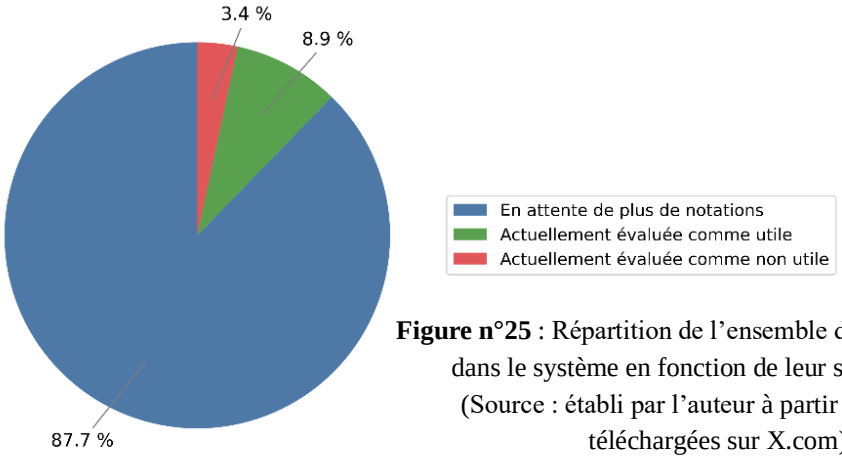

Figure n°25 : Répartition de l'ensemble des notes présentes dans le système en fonction de leur statut définitif (Source : établi par l'auteur à partir des données téléchargées sur X.com)

Cette visualisation révèle que près de 88 % des notes proposées depuis la création du système des notes de la communauté n'obtiennent aucun statut définitif et conservent le statut « en attente ». Cette répartition est stable d'un mois à l'autre et ne provient donc pas du nombre de votes insuffisant dont souffrait probablement une grande partie des notes durant les premiers mois du système. Il apparait donc que seulement 8,9 % des notes proposées par la communauté finissent par être visibles de tous. Pour comprendre ce résultat, et surtout la décision technique qui en est à l'origine, nous pouvons analyser ce qu'implique le fait pour une note d'obtenir le statut « utile » ou « non utile ». Se voir attribuer le statut « non utile » ne provoque aucun changement au niveau de la visibilité de la note, puisqu'elle ne serait pas non plus visible par les utilisateurs si elle conservait le statut « en attente ». Ce statut est cependant utilisé pour mesurer l'efficacité corrective des rédacteurs et des évaluateurs, comme nous le détaillerons dans une partie ultérieure. Ce mécanisme rendrait la présence d'une part importante de *faux positifs* néfaste au mécanisme d'évaluation des contributeurs, ce qui peut expliquer le faible taux de notes « inutiles ». En ce qui concerne le statut « utile », son implication est évidemment l'affichage de la note à tous. Il y a donc une volonté, par des choix techniques, de rendre l'affichage des notes rare. Près de 11 notes sur 12 qui seront proposées par des rédacteurs ne seront jamais affichées aux utilisateurs de X. Cet « élitisme » des contributions différencie le système des notes de la communauté d'autres systèmes contributifs en ligne comme Wikipédia, sur lesquels les propositions de rédaction ou de modification sont beaucoup plus souvent acceptées par la communauté. Le dispositif technique des notes de la communauté est donc ici principalement un système de tri de masse qui vient exclure une part très importante des contributions des utilisateurs.

2. Un mécanisme d'identification des convergences pour créer une sagesse des foules

a. Un système d'évaluation croisée pour détecter des accords entre des profils différents

Pour lever le voile sur le fonctionnement réel de l'algorithme, nous avons téléchargé et étudié le code en libre accès des notes de la communauté, en nous focalisant sur le mécanisme régissant l'affichage d'une note. Le mécanisme implémenté reprend les travaux d'Aviv Ovadya publiés en 2022 dans un rapport[69]. Dans cet article, l'auteur dénonce les algorithmes de recommandation *engagement-based,* c'est-à-dire basés sur le nombre de clics, de *j'aime* ou sur le temps de visionnage. Ce type d'algorithme est accusé de récompenser le contenu sensationnel et polarisant et donc indirectement d'affaiblir les systèmes démocratiques. L'analogie d'un système « centrifuge » est employée par l'auteur pour décrire ces algorithmes qui polarisent structurellement leurs utilisateurs. La proposition d'Aviv Odaya est d'utiliser de nouveaux types d'algorithme pour introduire une force « centripète », par analogie avec le principe physique qui vient compenser la force centrifuge. Ces nouveaux algorithmes, appelés *bridging-based algorithm,* que nous traduirons par « algorithmes de consensus transverses », se caractérisent par l'objectif de récompenser le contenu et les comportements qui rapprochent des groupes idéologiquement opposés. Pour y parvenir, le principe est de mesurer le degré de consensualité d'un contenu parmi des groupes idéologiquement diversifiés, puis, de valoriser ceux qui obtiennent les scores les plus élevés, au lieu de récompenser ceux qui génèrent le plus de réactions. Dans le cas des notes de la communauté, ce type d'algorithme semble particulièrement approprié, dans la mesure où les contenus que le système doit approuver ou désapprouver sont souvent hautement idéologiques et clivants. Employer un algorithme standard utilisant un simple seuil de taux de votes positifs pour déterminer l'utilité d'une note ne pourrait à l'évidence pas produire des bons résultats. En effet, ce type de système serait fortement biaisé en direction de l'idéologie de la majorité et il suffirait à des groupes militants de se coordonner dans leurs votes pour afficher ou retirer la note de leur souhait. N'afficher que des notes consensuelles parmi des groupes de gens idéologiquement différents apparait comme la

[69] Ovadya, A. (2022). Bridging-Based Ranking: How Platform Recommendation Systems Might Reduce Division and Strengthen Democracy [En ligne]. https://www.belfercenter.org/project/technology-and-public-purpose

meilleure option pour garantir la fiabilité du système dans un environnement polarisé. C'est ce que réalise l'algorithme régissant l'affichage des notes de la communauté. Le fonctionnement concret du système de validation de note fait intervenir des concepts d'algèbre linéaire et d'apprentissage automatique avancés. Une vulgarisation plus détaillée du fonctionnement réel de l'algorithme est disponible en annexe n° 3. Les éléments clefs de son fonctionnement sont les suivants : Le système va attribuer un score d'utilité aux notes, en cherchant à valoriser les notes dont les votants sont modélisés comme idéologiquement éloignés de la note. L'algorithme va ensuite répartir les contributeurs dans des groupes, de manière à ce qu'ils soient regroupés par habitude de notation. Avant de valider une note, le système va s'assurer que les utilisateurs l'ayant jugée utile appartiennent à plusieurs groupes différents. Ce procédé est la manière choisie par X pour implémenter l'algorithme de consensus transverse. Le choix technique de placer la consensualité des notes au cœur du système a donc été retenu par les concepteurs du dispositif, et c'est ce choix qui explique la rareté des notes utiles.

Encart n°2 : Un écart entre la consensualité perçue par les utilisateurs et celle mise en œuvre

Le sondage réalisé auprès d'utilisateurs réguliers du réseau social X révèle qu'ils portent un regard globalement négatif sur l'algorithme régissant le système d'affichage des notes de la communauté.

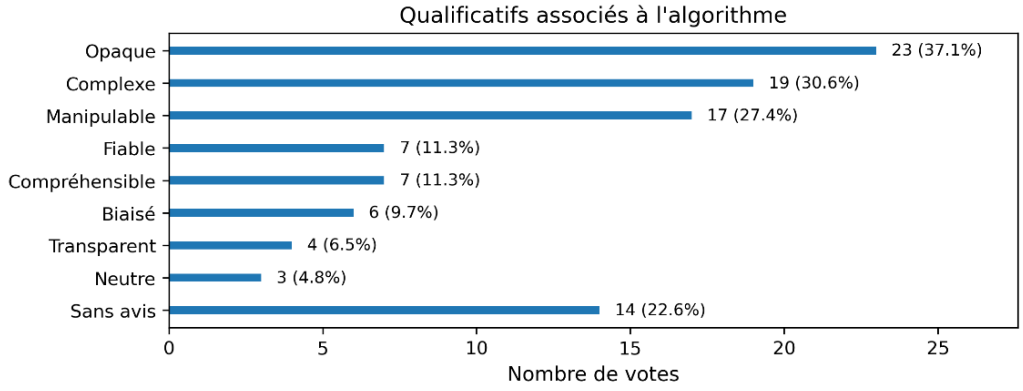

Figure n°26 : Résultats du sondage, question sur les qualificatifs associés à l'algorithme (Source : établi par l'auteur)

Le traitement algorithmique réalisé semble être perçu comme une boîte noire opaque par une majorité de sondés. Le résultat le plus étonnant est sans doute le fait que seuls 6,5 %

d'entre eux ont sélectionné le qualificatif « transparent », bien que l'algorithme soit en source ouverte. Ces résultats peuvent donc s'expliquer par la méconnaissance généralisée du fonctionnement de l'algorithme. Quand on interroge les utilisateurs sur leur connaissance de son fonctionnement, une écrasante majorité ne le connaît pas du tout ou peu (annexe n° 4). Le système apparait donc incompris pour une part très importante des sondés. Ce sentiment est probablement dû à l'absence de rétroaction et de transparence dans l'interface des notes de la communauté. Un utilisateur face à une note n'a aucune information sur le nombre de personnes qui l'ont évaluée ou encore sur le taux de votes positifs qu'elle a reçu. Le système réel mis en œuvre rend impossible une explication claire des critères de validation d'une note. Le choix retenu par les concepteurs du système a été de faire reposer l'évaluation de la consensualité d'une note sur des principes algorithmiques solides mais extrêmement complexes. Ce choix rend totalement inaccessible sa compréhension et son appropriation par les utilisateurs. Il apparait donc une distinction entre la consensualité technique mise en œuvre et la consensualité perçue des décisions issues de l'algorithme. La légitimité du consensus qui émerge du système ne repose pas sur la pertinence des choix techniques mis en œuvre, mais sur la perception que les utilisateurs en ont. Établir cette consensualité par des principes inaccessibles à une très grande majorité d'utilisateurs, c'est risquer qu'elle soit perçue comme une forme de manipulation.

b. La formation d'une sagesse des foules dans un environnement polarisé

L'architecture technique retenue par le système des notes de la communauté fait l'hypothèse qu'un groupe hétérogène peut, en combinant ses jugements, aboutir à une forme de correction pertinente et fiable. Ce principe s'inscrit dans le prolongement de la théorie de la sagesse des foules, développée par James Surowiecki[70]. L'auteur postule que, sous certaines conditions, l'agrégation des contributions individuelles d'une foule peut produire des résultats similaires, voire supérieurs, à ceux d'experts. James Surowiecki identifie quatre conditions que doit satisfaire une foule pour pouvoir déclencher le phénomène de sagesse des foules. La première est la diversité des opinions des individus. Les contributeurs doivent avoir des idéologies et des perspectives différentes pour garantir une hétérogénéité de leurs apports. La deuxième est l'indépendance des jugements. Chaque personne doit pouvoir se

[70] Surowiecki, J. (2004). *The Wisdom of Crowds.* New York : Anchor Books.

faire son propre avis indépendamment de celui des autres. La troisième condition est la décentralisation. Les contributeurs doivent appuyer leurs décisions sur leurs connaissances individuelles et non sur un ensemble partagé. La dernière condition identifiée par l'auteur est l'agrégation. Les résultats individuels doivent être transformés en décision collective par un mécanisme fiable. Nous pouvons appliquer chacune de ces conditions au système des notes de la communauté pour déterminer s'il s'inscrit bien dans ce modèle de sagesse des foules.

La diversité des opinions est clairement présente dans le système contributif. Le réseau X est fortement polarisé et diversifié idéologiquement, et cette diversité se traduit bien dans le système des notes de la communauté. Nous avons démontré cette diversité idéologique dans la troisième section du premier chapitre, dans laquelle nous avons remarqué une relative équipartition entre les publications politiques annotées classées à gauche et celles classées à droite.

L'indépendance des votes des contributeurs est structurellement présente dans le système des notes de la communauté, dans la mesure où les votes réalisés sur une note ne sont pas visibles de tous. L'absence d'espace de discussion entre contributeurs garantit également l'indépendance du jugement. Cependant, la possibilité de rédiger des notes uniquement à destination des contributeurs dans le but de justifier qu'il n'est pas utile d'afficher une note sur une publication peut constituer un manquement à ce principe d'indépendance, même si ce type de note reste marginal.

La décentralisation des votes est assurée par le système, dans la mesure où les évaluateurs fondent leur jugement uniquement sur leurs croyances, opinions et expertise. Les éléments évoqués dans les deuxièmes et troisièmes sections du premier chapitre viennent cependant nuancer ce constat. En effet, de nombreuses normes implicites partagées par une part importante des membres de la communauté viennent influencer les votes individuels.

Finalement, l'agrégation réalisée permet bien de transformer les votes individuels en un score global d'utilité, nommé *intercept* dans le système (cf. annexe n° 3).

À première vue, les notes de la communauté s'inscrivent bien dans le cadre du modèle de Surowiecki. Cependant, comme nous l'avons évoqué dans la sous-partie précédente, le modèle ne se contente pas d'agréger les résultats, mais vérifie également que les votes utiles proviennent d'utilisateurs qui appartiennent à des groupes idéologiques

différents. Cette dernière condition semble indispensable pour créer une sagesse des foules au sein d'une population biaisée et polarisée. En effet, les conditions évoquées par Surowiecki permettent de déclencher ce phénomène uniquement dans l'évaluation d'éléments pour lesquels les votants n'ont pas de biais structurels majeurs. Demander à une foule d'évaluer la température d'une pièce puis agréger les propositions en considérant leur médiane produit de très bons résultats dans la mesure où la population n'est pas composée de groupes d'individus qui vont structurellement surévaluer ou sous-évaluer cette température. Dans le cas de la correction de contenu politique ou clivant, des sous-ensembles de votants vont avoir des biais très importants dans leurs contributions. Une simple agrégation des résultats ne peut donc pas à elle seule produire des résultats satisfaisants sur du contenu particulièrement clivant, sauf à supposer que les biais structurels des groupes idéologiques se compensent parfaitement. Une étude a ainsi montré qu'en matière de vérification d'information, une foule n'était aussi performante qu'un expert que si les méthodes d'agrégation de leurs contributions utilisaient du *machine learning* pour compenser les biais des participants[71].

Nous proposons donc d'ajouter une cinquième condition au modèle développé par Surowiecki pour qu'une sagesse des foules puisse émerger en présence de forts clivages structurels dans une population. Cette condition est qu'une mesure du degré de consensualité de la décision collective parmi des groupes idéologiquement différenciés doit être réalisée et dépasser un seuil fixé. Ce cinquième critère permet de garantir que le résultat final n'est pas simplement la position d'un groupe homogène dominant dans la population, mais qu'il est issu de contributions positives provenant de sous-groupes idéologiquement variés. Dans le cas des notes de la communauté, le choix qui a été retenu est de n'approuver une note que si les contributeurs l'ayant jugée utile proviennent de plus de trois *clusters* différents (cf. annexe 3). Déclencher un phénomène de sagesse des foules pour évaluer du contenu clivant dans un environnement polarisé semble donc nécessiter de regrouper idéologiquement les participants et de s'assurer de la consensualité de chaque résultat.

[71] Godel, W., Sanderson, Z., et al., (2022). Moderating with the Mob: Evaluating the Efficacy of Real-Time Crowdsourced Fact-Checking. *Journal of Online Trust and Safety*. vol. 1, n° 1, p. 1-36.

3. Un consensus sans débat construit sur une modélisation réductrice de l'espace délibératif

a. L'impossibilité d'un découpage idéologique fonctionnel sans arbitraire

Nous avons vu que le modèle de sagesse des foules complété du critère de consensualité permet de s'assurer que les résultats des décisions collectives ne seront pas manipulés par un sous-groupe biaisé. Cependant, ce modèle suppose qu'une mise en œuvre technique de ce principe puisse être réalisée. L'élément clef à mettre en œuvre est le regroupement des utilisateurs dans des groupes idéologiques relativement homogènes. En réalité, nous avons toujours pris l'exemple de biais idéologiques pour illustrer le fonctionnement de l'algorithme, mais l'idéologie politique n'est pas le seul élément qui peut rendre les évaluations biaisées. En effet, comme évoqué dans le premier chapitre, le contenu politique est loin de constituer le seul thème abordé par les notes de la communauté. Les éléments pouvant orienter structurellement les évaluations des groupes de contributeurs ne sont pas non plus tous liés à la politique. Il peut s'agir, par exemple, de sensibilités particulières à certains registres ou à certaines structures argumentatives. Se pose donc la question de déterminer la manière la plus neutre, juste et pertinente de regrouper les individus de notre foule biaisée.

Le système des notes de la communauté a fait le choix de le faire uniquement selon leurs habitudes de notation. Concrètement, cela signifie qu'à l'issue du processus de regroupement, deux contributeurs vont se retrouver dans le même groupe s'ils tendent statistiquement à prendre les mêmes décisions d'utilité ou d'inutilité sur un contenu similaire. Ce regroupement est effectué par des procédés d'algèbre linéaire et des méthodes de *clustering*, se basant sur l'historique des évaluations des contributeurs. Cette méthode a le double avantage de ne pas se baser sur une analyse partiale et subjective de ce que la plateforme considérerait comme « l'idéologie », et d'englober structurellement tous les types de biais. En effet, le regroupement automatique par proximité d'habitude de notation permet d'établir des groupes de contributeurs ayant des biais proches, mais sans avoir à définir et catégoriser explicitement ces biais.

Encart n° 3 : À quels critères correspond le regroupement des utilisateurs effectué automatiquement par le système ?

L'accès au numéro des groupes auxquels appartiennent l'ensemble des contributeurs dans la base de données nous permet d'étudier le regroupement final plus en détail. Une observation des groupes de provenances des utilisateurs ayant voté pour des notes prises au hasard semble indiquer qu'ils tendent systématiquement à provenir d'un même groupe. Ce groupe diffère d'une note à l'autre, mais semble dépendre de la langue dans laquelle est écrite la publication. Pour vérifier cette hypothèse, nous avons réalisé un programme qui sélectionne aléatoirement des notes ayant reçu plus de 300 évaluations et qui détermine pour chacune le cluster majoritaire d'appartenance des contributeurs l'ayant évaluée. Nous avons ensuite sélectionné aléatoirement, pour chaque cluster parmi les onze les plus représentés, dix notes qui lui sont associées, et avons déterminé pour chaque note la langue dans laquelle elle est rédigée.

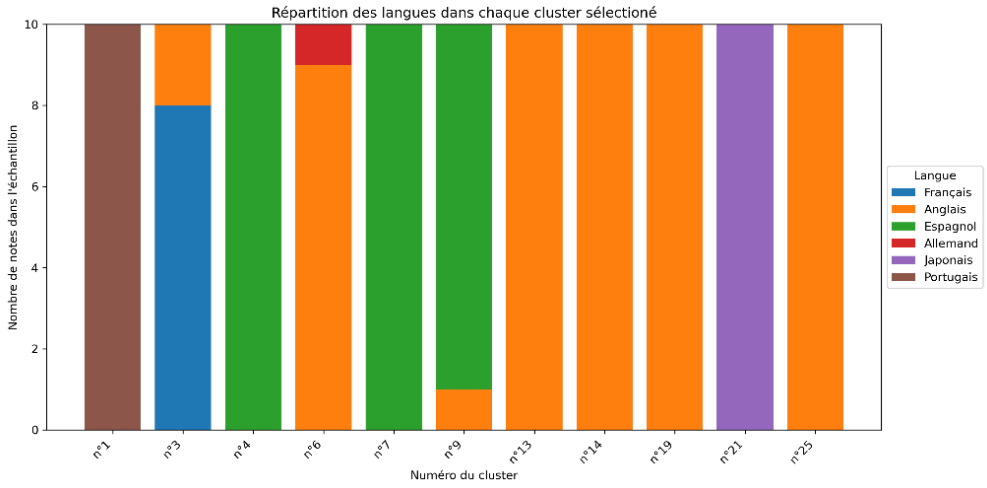

Figure n°27 : Répartition des langues dans les échantillons de notes des différents groupes d'utilisateurs du système
(Source : établi par l'auteur à partir des données téléchargées sur X.com)

Il apparait clairement sur ce graphique que le regroupement réalisé suit davantage des critères linguistiques et géographiques qu'idéologiques. Le problème est certainement dû au fait qu'un réseau social mondial comme X comporte de nombreux segments mono-linguistiques ou nationaux relativement isolés les uns des autres. Fonder le regroupement des utilisateurs uniquement sur les habitudes de notation induit donc un regroupement selon des critères qui ne sont pas du tout ceux que l'on s'attendrait à avoir dans le but d'évaluer la consensualité des notes.

Au-delà d'un simple problème technique, le mauvais regroupement décrit dans l'encart précédent révèle une limite épistémologique de la modélisation algorithmique du consensus. Les séparations idéologiques des individus ne peuvent pas être mises en évidence par un simple traitement algorithmique des habitudes de notation. Elles doivent être façonnées artificiellement à partir des contributions ou des données qui leur sont associées. Cela implique de définir dans la conception de l'algorithme ce qui constitue un écart idéologique. Un algorithme qui regroupe simplement les individus par habitude de notation ne fait aucune différence entre un écart dû à des espaces linguistiques différents et ceux imputables à un écart idéologique. Pour être pleinement fonctionnel, un algorithme de consensus transverse doit forcément être orienté par ses concepteurs pour découper l'espace social en segments qui constitueront selon eux une diversité. Ce type de système ne pourra donc structurellement jamais garantir de pluralisme, mais seulement le consensus parmi des groupes issus de la représentation que ses concepteurs se font de ce qui constitue la diversité.

b. Des choix techniques qui appauvrissent la diversité contributive

En plus de devoir décider des catégories artificielles à créer pour regrouper les utilisateurs et évaluer la consensualité des notes, le système doit également définir une manière d'attribuer un groupe à chaque utilisateur. Le procédé retenu par les notes de la communauté est celui d'attribuer à chaque contributeur une valeur à l'issue du processus d'apprentissage automatique, nommée dans le système *facteur du contributeur* (cf. annexe 3). Cette valeur peut être interprétée comme son profil de jugement, qui contient tous ses biais structurels. Le choix a été fait par X de n'attribuer qu'une seule valeur à chaque contributeur au lieu d'un ensemble de valeurs. Cela signifie que les évaluateurs sont répartis sur un simple axe, et non pas sur un plan, ou un espace de plus grande dimension. La diversité des profils des contributeurs est ainsi simplement vue comme modélisable par une seule dimension, comme si une classification idéologique universelle existait sur un axe gauche – droite. En réduisant la diversité des profils des contributeurs à un seul axe, le système ne rend pas efficacement compte de la diversité réelle des sensibilités et des formes de désaccord. De plus, les tensions liées à des formes marginales de désaccord ou de représentation idéologiques sont invisibilisées par un système qui réduit les habitudes de notation des centaines de milliers de contributeurs à une seule valeur. Ce choix technique repose également sur l'idée qu'il est possible d'ordonner les habitudes de notation des contributeurs de manière cohérente et stable dans le temps. Définir qu'un contributeur A est

deux fois plus proche de B que de C, en se basant uniquement sur leurs historiques de notation, n'a pas forcément de sens dans un espace où les différences d'évaluation se fondent sur des écarts de perception, de compétence ou d'intérêts de contenus extrêmement diversifiés, allant du tweet humoristique au contenu politique clivant. Ce procédé fait donc également l'hypothèse que la distance idéologique est mesurable et quantifiable. Cette hypothèse soulève de nombreux problèmes dans la mesure où la différence idéologique est réduite à un calcul qui ne prend en compte que les éléments observables du comportement en ligne des individus. Le résultat des opérations de classification idéologique ne peut donc pas, par exemple, prendre en compte la signification sociale des divergences observées ou la nature du contenu corrigé. Quantifier un degré d'accord entre deux contributeurs ne peut se faire sans assimiler ces concepts à une valeur mesurable détachée de tout fondement culturel et social. De plus, les éléments sur lesquels se base le système des notes de la communauté pour déterminer ces distances idéologiques sont de simples informations binaires. Il s'agit des votes des contributeurs, qui ne peuvent s'exprimer que par deux options : « utile » et « non utile ». Ce format, même s'il rend les évaluations plus rapides et fluides, induit une forte restriction dans les modalités d'évaluation d'une note. Les formes d'accord et de désaccord sont elles aussi simplifiées à l'extrême et réduites à un critère binaire. Les éléments ambivalents, les positions nuancées ou les jugements prudents sont effacés face à un choix forcé d'accord ou de désaccord complet. La notion d'argumentation ou de débat disparait donc totalement du processus de sélection d'une note, qui est réduit à un vote privé sans échange.

c. La création d'un consensus vidée de sa dimension délibérative

L'algorithme mis en œuvre pour décider de l'utilité ou non d'afficher une note révèle la manière dont la notion de consensualité est perçue par ses concepteurs. Là où, au sein d'un système délibératif au sens d'Habermas, le consensus émerge d'un compromis issu de délibération, il ne résulte ici que d'une simple sélection. La notion de consensualité semble donc perçue comme étant seulement l'ensemble des positions individuelles qui ne sont pas rejetées par un trop grand nombre de sensibilités idéologiques. Le dispositif n'est absolument pas pensé ou conçu pour que ce consensus émerge de débats, de compromis ou de l'évolution de la position de certains contributeurs. L'absence structurelle de toute forme d'échanges entre les contributeurs rend impossible la mise en place d'un débat, même rudimentaire. Le

système sociotechnique mène donc au résultat surprenant de produire des corrections acceptées par un large spectre idéologique, mais sans induire aucune modification de l'opinion de qui que ce soit. Nous pouvons supposer que, face au constat d'un réseau hautement militant, polarisé et violent, les créateurs du système correctif n'ont pas cherché à mettre en place un système délibératif, même rudimentaire. Exclure tout contenu rejeté par n'importe quel groupe idéologique permet d'obtenir des résultats globalement neutres dans un cadre dans lequel toute délibération constructive apparait illusoire. Cependant, cette méthode comporte plusieurs désavantages majeurs. Le premier, qui a été développé au début de cette section, est le très faible taux de notes validées qu'elle induit structurellement. Le système nécessite donc une masse critique importante de contributions pour produire assez de notes affichées. Un autre problème posé par cette méthode est qu'elle ne permet en réalité aucune création collective. Des annotations sur des sujets complexes et clivants pourraient bénéficier d'une construction collective issue d'échanges entre contributeurs, mais le système impose que la note soit le fruit d'un travail d'un seul homme. Le système ne met en réalité en place qu'une collecte des opinions figées et individuelles de chacun pour identifier les productions individuelles au croisement des avis de chacun. Encore une fois, l'absence de débat et le choix de n'afficher que le contenu constituant le plus petit dénominateur commun de groupes hétérogènes conduisent à une invisibilisation structurelle de toute position marginale.

Pour conclure cette section, le choix a été fait de définir algorithmiquement la pertinence d'une note par son absence de refus de la part de groupes idéologiquement diversifiés. La foule de contributeurs, utilisée pour identifier les points de convergence, crée une forme de sagesse qui est plus sélective que contributive. La constitution de ces groupes et l'ensemble des mécanismes algorithmiques qui mettent en œuvre ce processus ne sont pas neutres. Ils traduisent la conception de la diversité qu'ont leurs concepteurs et constituent immanquablement une réduction de l'espace idéologique qui empêche la prise en compte des formes périphériques d'idéologie. Le rôle du dispositif technique n'est pas ici d'organiser un débat, mais bien de créer une coordination silencieuse au sein de laquelle le conflit est neutralisé par les algorithmes. La consensualité apparente ne résulte d'aucune délibération et n'est principalement qu'un leurre, produit d'une invisibilisation massive du contenu conflictuel.

Section 2. Une structuration communautaire qui favorise l'efficacité du système au détriment du pluralisme démocratique

Après avoir étudié les mécanismes de sélection des contenus et démontré qu'ils ne résultent que d'une invisibilisation massive de tout élément clivant, nous allons étudier la structuration sociale de la population des contributeurs. Ce que nous appelons structuration sociale est la répartition des rôles qui émerge par des mécanismes techniques ou sociaux au sein de notre population. L'objectif de cette section est de déterminer quel rôle joue cette structuration dans l'émergence de la production collaborative et ce qu'elle révèle du pluralisme du système contributif.

1. Une habilitation de rédaction définie par des seuils de performance

a. Un droit à la contribution qui n'est pas universel et une sociologie de contributeurs hétérogène

Une des caractéristiques premières du système des notes de la communauté est qu'il est ouvert à n'importe quel utilisateur. Il suffit à un internaute de posséder un compte X et de satisfaire certaines règles de sécurité, comme le fait de posséder le compte depuis plus de six mois et d'avoir renseigné un numéro de téléphone valide, pour pouvoir demander à rejoindre le dispositif. Cette apparente ouverture cache un système plus complexe et hiérarchisé dans lequel tous les contributeurs ne partagent pas les mêmes droits.

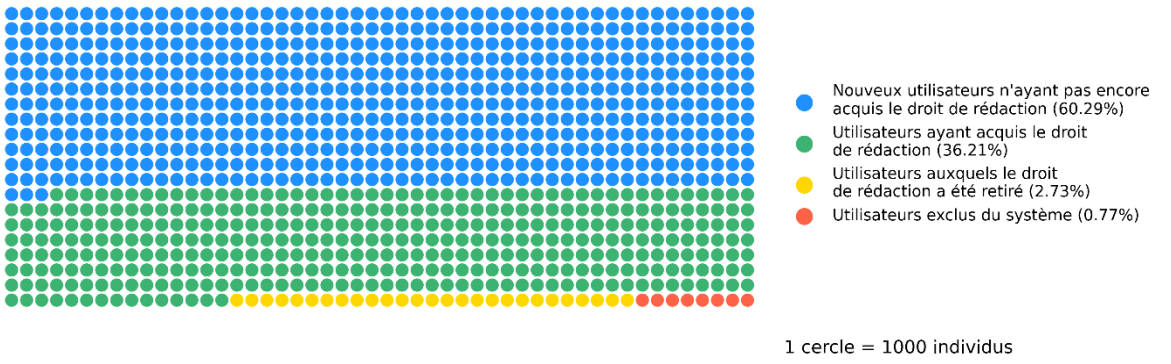

Figure n°28 : Visualisation de la répartition des contributeurs selon leurs droits de rédaction
(Source : établi par l'auteur à partir des données téléchargées sur X.com)

Cette visualisation représente la répartition des droits de rédaction de l'ensemble des contributeurs inscrits, c'est-à-dire ceux qui ont fait la demande de rejoindre le système et dont la demande a été acceptée. Elle révèle que ce droit n'est étonnamment détenu que par une minorité de contributeurs. La majorité des utilisateurs inscrits (60,29 %) est encore en attente de l'obtention de ce droit et reste limitée à évaluer le contenu produit par les autres membres. L'accès à la fonctionnalité de rédaction nécessite de franchir un certain score d'évaluation, c'est-à-dire d'aider un certain nombre de notes à atteindre le statut « utile » ou « inutile » par ses évaluations. Dans le cas où le statut final d'une note est différent du vote réalisé par l'évaluateur, sa note est dégradée. Une fois que ce score dépasse un seuil défini, le contributeur débloque la possibilité de rédiger ses propres notes. Ce droit n'est cependant jamais acquis définitivement et un contributeur peut le perdre dans plusieurs situations. Dans le cas où un trop grand nombre de ses évaluations s'avèrent contraires au statut final pris par la note, le contributeur se voit retirer son droit. C'est également le cas si un taux trop important de notes qu'il a rédigées obtient le statut « inutile ». D'autres traitements algorithmiques peuvent conduire à la perte de ce droit. Par exemple, s'il est estimé que le contributeur a des comportements suspects, comme le fait de voter trop rapidement, de changer subitement d'habitude de notation ou de considérer utiles des notes massivement signalées comme inappropriées.

Ces mêmes critères sont également utilisés par le système pour retirer aux contributeurs leur droit de vote. Ce retrait n'est pas explicite dans la mesure où il s'agit de ne plus prendre en compte le vote du contributeur dans l'évaluation finale de l'utilité d'une note. Il est difficile d'évaluer la proportion d'utilisateurs qui ont perdu leur droit de vote, car il ne s'agit pas d'un statut explicite visible dans la base de données comme celui du droit de rédaction. La prise en compte ou non des votes d'un utilisateur semble constamment recalculée par le système. Il est intéressant de noter que le contributeur est averti lorsqu'on lui retire son droit de rédaction, mais pas quand ses votes ne sont plus pris en compte par le système. Ce retrait du droit de vote s'apparente donc à un mode de régulation silencieux et invisible.

b. Un droit de participation décidé par un algorithme

La structuration sociale des contributeurs est hiérarchisée par l'obtention de droits : le droit de rédiger une note et celui d'avoir son vote pris en compte par le système. Ils ne sont pas vus comme des droits fondamentaux et peuvent être perdus à tout moment par un utilisateur ne se conformant pas aux attentes du système. Contribuer à la production collective n'est donc pas un droit, mais un privilège, détenu par une minorité, évalué par les pairs et constamment remis en question. Cette décision technique peut paraitre surprenante dans la mesure où ce sont des calculs issus des votes de la communauté elle-même qui aboutissent à décider qui doit en faire partie ou non. Si tout individu dont les votes sont trop éloignés de ceux de la majorité des contributeurs ne peut pas intégrer cette communauté, émerge le risque d'un renforcement de biais structurels. Nous pouvons faire l'hypothèse que l'ensemble des normes implicites qui émergent des choix des contributeurs, que nous avons évoqué dans les deuxièmes et troisièmes sections de la première partie, est dans une certaine mesure entretenu, voire amplifié, par ces mécanismes. Les principes mis en œuvre ont la particularité d'être une forme d'autorégulation de la communauté entièrement automatisée dans laquelle aucun contributeur n'évalue directement ses pairs. L'autorégulation des systèmes contributifs est généralement régie par des systèmes de votes ou de débats, comme c'est par exemple le cas sur Wikipédia pour la modération de contenu. Dans le cas des notes de la communauté, l'autorégulation ne fait donc intervenir aucune décision ni action des contributeurs. Seules les données issues de leurs votes sont utilisées pour en inférer le degré de marginalité du contributeur et l'exclure s'il dépasse un certain seuil. L'exclusion d'individus d'un système participatif n'est donc vue que comme une identification des comportements marginaux, qui peut donc être réalisée par un algorithme.

Le fait d'établir l'indésirabilité d'un contributeur à partir d'une mesure des divergences entre son comportement numérique et celui de la majorité des membres du dispositif rappelle des principes de gouvernementalité algorithmique développés par Antoinette Rouvroy et Thomas Berns [72]. Leur texte décrit la normativité établie par la gouvernementalité algorithmique comme dénuée « d'échelle, d'étalon, de hiérarchie ». Cette nouvelle forme de normativité s'adapte en temps réel aux comportements observés, sans

[72] Rouvroy, A., & Berns, T. (2010). Gouvernementalité algorithmique et perspectives d'émancipation : Le disparate comme condition d'individuation par la relation ?. *Revue européenne des sciences sociales, 177*, 163-176.

norme morale, sociale ou juridique fixe. Cette normativité, fondée uniquement sur des corrélations de données, correspond parfaitement à ce qui est réalisé dans le système des notes de la communauté où l'algorithme vient établir des corrélations statistiques entre les habitudes de notation des individus pour identifier les marginaux à exclure. Les données collectées sur les contributions des individus ont donc une portée prédictive. Elles servent à prédire et à modéliser les caractéristiques d'un contributeur qui sera accepté dans le système, et donc celles de celui qui en sera exclu. La gouvernance algorithmique se caractérise également par une réduction des individus à un ensemble de données, sans prise en compte de toute autre forme de réalité. Un contributeur n'existe aux yeux du dispositif que comme un nombre qui reflète ses biais idéologiques. En plus d'une absence de mécanismes de délibération, les choix du dispositif se font donc au sein d'une modélisation numérique qui déforme profondément la réalité. L'exclusion automatique des contributeurs sur la base d'irrégularité dans cette modélisation numérique empêche également toute forme de réflexion collective sur les conditions de participation au système.

Notons qu'une partie des principes d'exclusion est conçue en tant que mécanisme de sécurité, pour éviter que des robots ou des individus souhaitant saboter le système par des votes incohérents puissent fausser les résultats. Cependant, ces choix aboutissent à assimiler la marginalité à un caractère indésirable. Cela implique bien entendu une invisibilisation structurelle de tout rapport à l'information ou mode de pensée trop déviant par rapport aux formes dominantes. En plus de l'homogénéisation des contenus validés par l'algorithme de consensus transverse décrit dans la section précédente, le système va homogénéiser les contributeurs autorisés à rédiger et à voter. Ce principe de modération possède cependant l'avantage de pouvoir être mis en place dans des écosystèmes hautement militants et polarisés comme X. Employer des mécanismes de modération comme ceux de Wikipédia ne pourrait probablement pas se faire sans qu'ils soient détournés à des fins militantes. Encore une fois, la solidité et la sécurité du système contributif semblent avoir été privilégiées au pluralisme et à la liberté de contribuer.

2. Une autocensure qui renforce les asymétries rédactionnelles

a. Des écarts importants de contribution parmi les utilisateurs autorisés

Le droit de rédaction n'étant détenu que par 36,21 % des contributeurs, nous allons nous concentrer sur les habitudes de rédaction de cette population dans ce qui suit. Nous pouvons commencer par étudier la proportion de ces contributeurs qui font usage de leur droit de rédaction.

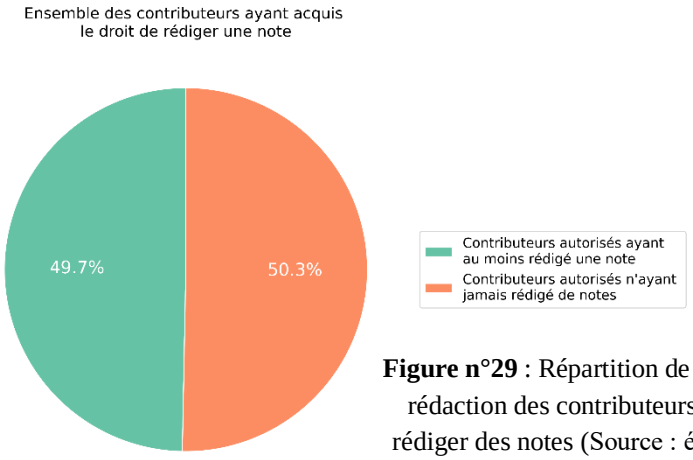

Figure n°29 : Répartition de la participation à la rédaction des contributeurs ayant le droit de rédiger des notes (Source : établi par l'auteur à partir des données téléchargées sur X.com)

Nous obtenons le résultat surprenant que près de la moitié des contributeurs qui ont obtenu le droit de rédaction n'ont jamais rédigé une seule note. Nous allons donc continuer de réduire notre population d'étude à cette moitié de rédacteurs ayant proposé aux suffrages de leurs pairs au moins une note. Nous appellerons cette population les rédacteurs actifs. L'élément que nous souhaitons évaluer dans cette partie est la répartition des contributions parmi les rédacteurs. Le fichier contenant l'ensemble des notes du système dans la base de données nous indique qu'au total 1 601 965 notes ont été rédigées par 226 493 contributeurs différents. Le nombre moyen de notes rédigées par contributeur actif est donc d'environ 7. Nous pouvons désormais représenter la répartition du nombre de notes total rédigées par les contributeurs actifs pour étudier sa distribution autour de cette valeur.

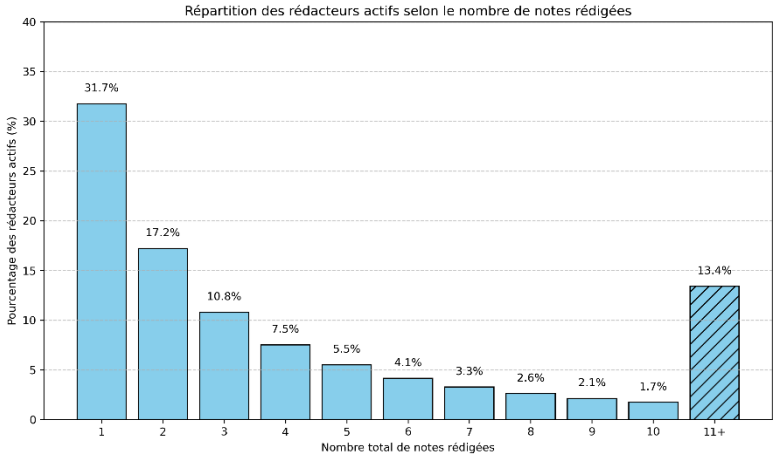

Figure n°30 : Répartition des rédacteurs actifs selon le nombre de notes rédigées
(Source : établi par l'auteur à partir des données téléchargées sur X.com)

Il apparait que près de 75 % des contributeurs actifs ont rédigé moins de 7 notes, c'est-à-dire moins de notes que la moyenne au sein de la population. Un calcul de la médiane du nombre de notes rédigées révèle qu'elle ne s'élève qu'à 3 notes. La valeur moyenne est donc tirée vers le haut par une minorité de contributeurs particulièrement actifs. Seuls 13,4 % des contributeurs actifs ont proposé aux votes de leurs pairs plus de 10 notes. Rapportés à l'ensemble des contributeurs, ils ne représentent que 2,41% de la population. Nous pouvons également mesurer que seuls 16 % des contributeurs sont à l'origine de la rédaction des deux tiers des notes. La rédaction des notes se caractérise donc par une profonde asymétrie dans l'espace des contributeurs. Une majorité des individus sont privés du droit de rédaction et une grande partie de ceux qui ont ce droit ne l'utilisent pas ou peu, laissant l'essentiel de la production participative dans les mains d'une petite minorité.

b. L'autocensure comme facteur majeur de non-participation effective à la correction communautaire

L'asymétrie contributive que nous observons ne peut pas uniquement s'interpréter par la présence de mécaniques algorithmiques de seuils et d'exclusion. Nous venons en effet de voir que cette asymétrie est également présente au sein de la population des membres ayant le droit de contribuer. Une moitié de ces individus ne participe jamais et près des trois quarts de la moitié restante ne participent que très rarement. Pour tenter de comprendre ce phénomène, nous pouvons étudier les raisons évoquées par les utilisateurs de X pour ne pas demander à rejoindre le système des notes de la communauté.

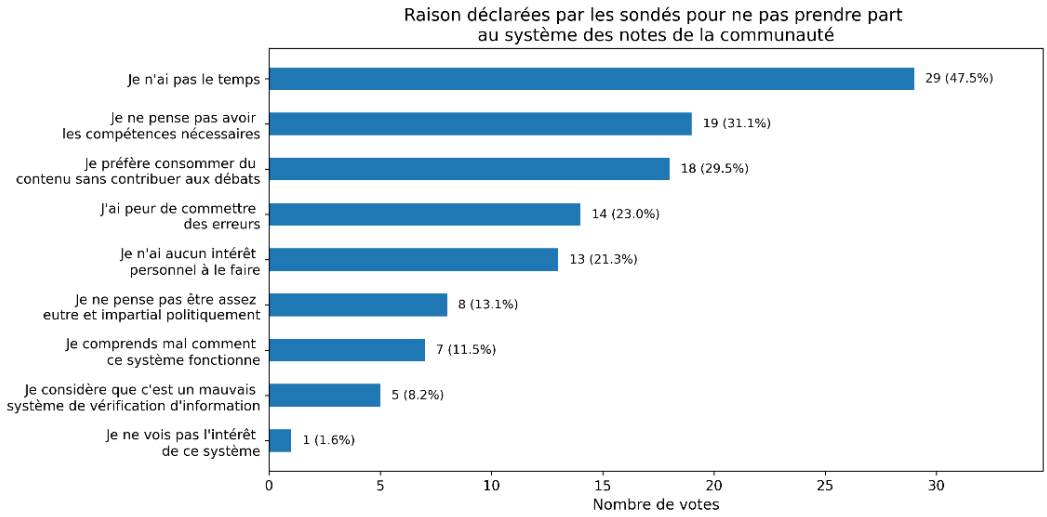

Figure n°31 : Résultats du sondage – Raisons de ne pas prendre part au système des notes de la communauté (Source : établi par l'auteur)

Le sondage réalisé révèle qu'après le manque de temps, la croyance dans le fait de ne pas avoir les compétences nécessaires arrive en seconde position des raisons évoquées, déclarée par près d'un tiers des sondés. En l'absence de données sur les contributeurs au système des notes de la communauté, nous pouvons faire l'hypothèse que des mécaniques d'autocensure peuvent partiellement expliquer cette asymétrie importante. La légitimité « légale » conférée par la plateforme aux contributeurs par le déblocage du droit de rédaction semble être dissociée de la légitimité implicite que les rédacteurs s'attribuent. Ce décalage peut être causé par la pression d'un système particulièrement normatif et centré sur des mécaniques d'évaluation. Il est également probable que la position sociale occupée par un contributeur dans la société exerce une influence importante sur sa légitimité perçue à contribuer au système.

3. Un système qui aboutit à la création d'une élite rédactionnelle

a. Les notes qui finissent par être jugées utiles sont rédigées par une très faible proportion des rédacteurs

Dans la partie précédente, nous avons étudié la répartition de la population en termes de contribution de rédaction. Dans cette partie, nous allons nous intéresser à la répartition des taux d'utilité des notes proposées. En effet, proposer des notes ne suffit pas à pleinement contribuer au mécanisme de correction. Encore faut-il que ces propositions soient acceptées par la communauté et affichées au public. Nous proposons donc d'étudier la répartition des contributeurs non plus seulement selon leur nombre de notes rédigées, mais selon leur nombre de notes jugées utiles.

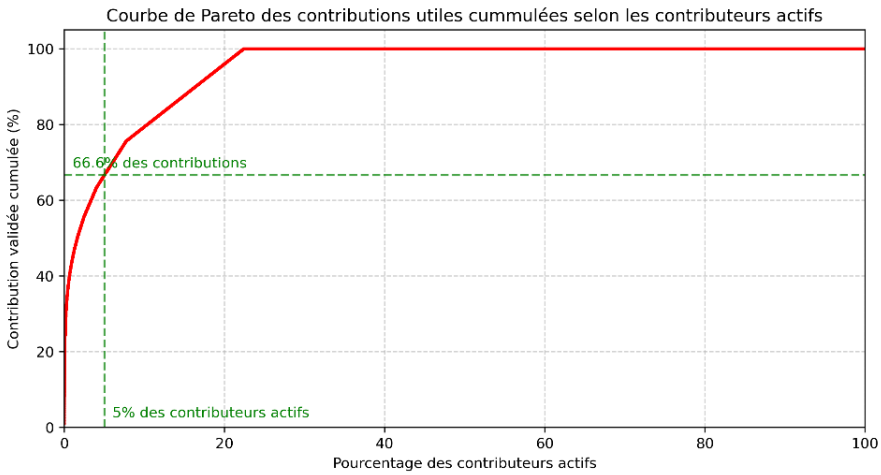

Figure n°32 : Courbe de Pareto des notes utiles selon le nombre de rédacteurs actifs qui en sont à l'origine
(Source : établi par l'auteur à partir des données téléchargées sur X.com)

Alors que près de 16 % des contributeurs actifs étaient à l'origine des deux tiers des notes rédigées, seule 5 % de cette population est à l'origine des deux tiers des notes jugées utiles. À une asymétrie de contribution s'ajoute une asymétrie de performance de rédaction. La communauté ne semble valoriser les rédactions que d'un nombre très restreint d'individus. Sur un ensemble de plus d'un million de contributeurs inscrits, moins de dix mille d'entre eux produisent les deux tiers des éléments correctifs affichés. Le système repose donc en grande partie sur l'émergence d'une élite rédactionnelle en son sein. Notons que la formation de cette élite se fait sans mécanisme de distinction. Les propositions de

notes sont entièrement anonymes, publiées par des comptes dont le système attribue un pseudonyme aléatoire. Même si certains scores de performance des contributeurs sont accessibles dans l'interface, peu de mécanismes peuvent expliquer la formation d'un tel groupe. Ce résultat rappelle ceux établis dans l'état de l'art relatifs aux motivations des contributeurs de Wikipédia à participer au système. L'élément de motivation premier qui ressortait de ces études était le plaisir de contribuer[73]. Nous pouvons supposer que cette motivation est également prépondérante parmi nos contributeurs.

b. Une élite qui contribue massivement et qui possède des caractéristiques innées, sources de ses bons scores

Pour étudier plus en détail « l'élite » que nous avons identifiée, c'est-à-dire le groupe d'environ dix mille contributeurs responsables des deux tiers des notes affichées, nous pouvons analyser le lien entre la fréquence de rédaction et le taux de notes validées.

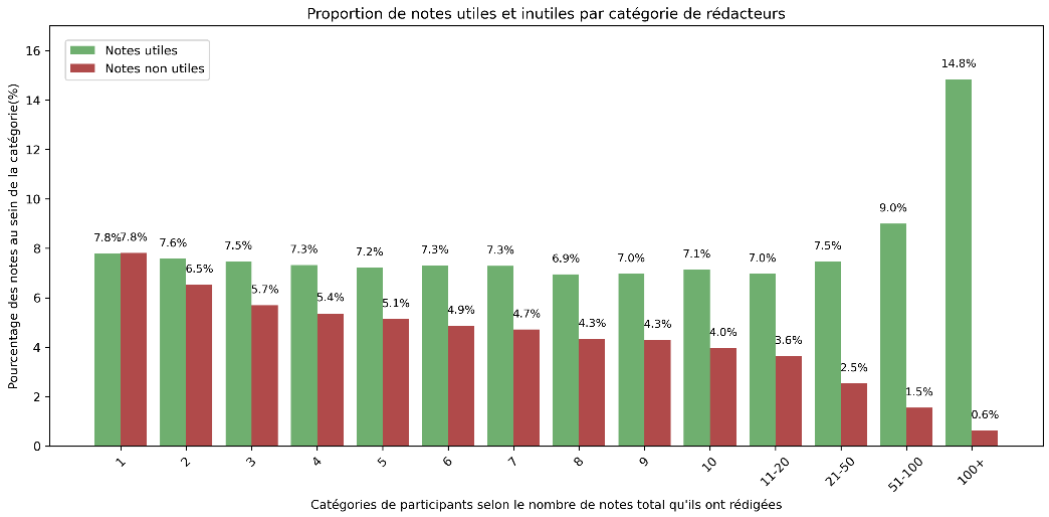

Figure n°33 : Proportion de notes utiles et non utiles selon le nombre de notes totales proposées par leur rédacteur
(Source : établi par l'auteur à partir des données téléchargées sur X.com)

Le graphique ci-dessus représente les taux de notes validées et invalidées en fonction du nombre total de notes rédigées par les contributeurs. Par exemple, les barres vertes et rouges situées au niveau de la graduation quatre signifient qu'au sein de l'ensemble des notes

[73] Nov, O., *op. cit.*

rédigées par les contributeurs qui n'ont rédigé que quatre notes depuis la création de leur compte, 7,3 % d'entre elles ont atteint le statut final « utile » et 5,4 % le statut « inutile », le reste étant resté en attente de plus d'évaluations. À la lecture de ce graphique, il apparait clairement que plus un rédacteur rédige de notes, plus la probabilité que ses propositions soient acceptées par la communauté est élevée. Se pose alors la question de savoir si ce résultat s'explique par un mécanisme d'apprentissage progressif ou par le fait que les contributeurs les plus performants sont aussi ceux qui contribuent le plus. Dans la première hypothèse, on suppose que la faculté de rédiger des notes qui seront validées par le système s'apprend, et que le taux de validation de ses notes tend à s'améliorer avec l'expérience. Pour tester cette hypothèse, nous pouvons réduire notre population aux individus ayant rédigé plus de cent notes et étudier l'évolution de leur taux de validation. Pour cela, nous avons récupéré les notes proposées par ces contributeurs et les avons ordonnées selon leur date de publication. Nous avons ensuite regroupé ces notes par groupes de dix et calculé le taux de validation de l'ensemble, que nous avons représenté par un graphique en barres pour chaque tranche de dix notes.

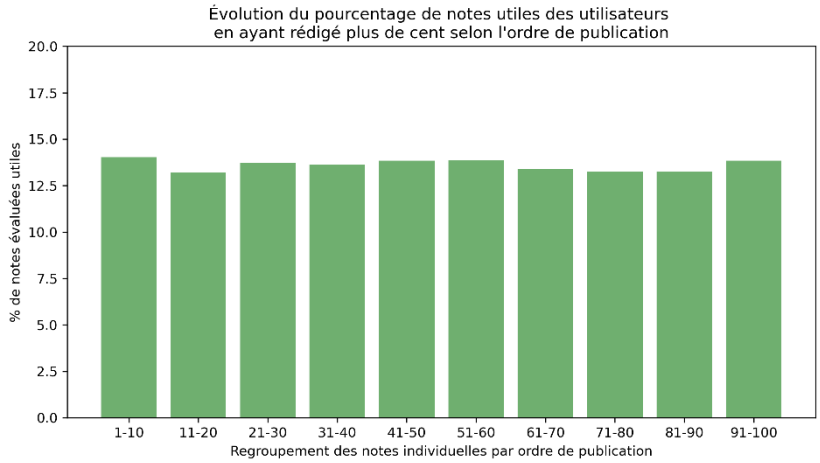

Figure n°34 : Évolution des taux de notes utiles des contributeurs en ayant proposées plus de cent selon leur ordre de publication
(Source : établi par l'auteur à partir des données téléchargées sur X.com)

La hauteur de la première barre de ce graphique indique que, parmi l'ensemble des notes proposées par des contributeurs en ayant rédigé plus de cent depuis la création de leur compte, près de 14 % des dix premières notes rédigées par chacun d'entre eux ont obtenu le statut de note utile. Nous remarquons que ce pourcentage est stable d'une dizaine à l'autre. Le graphique précédent nous avait indiqué que, parmi les notes des contributeurs n'en ayant

rédigé au total que dix, seules 7,1 % d'entre elles avaient atteint le statut utile. Ici, parmi les dix premières notes proposées par notre population chronologiquement, près de 14 % d'entre elles ont obtenu le statut « utile », soit près du double. Ce résultat surprenant indique que le fait de produire des notes performantes semble être inné et non acquis. Le système ne semble pas permettre une amélioration des compétences de rédaction et de conformité aux attentes implicites de ses pairs. Tout semble indiquer que les contributeurs les moins performants tendent à être exclus du système ou découragés de participer. Les choix techniques et la structuration sociale aboutissent à ce que seule une minorité alignée avec les attentes implicites du système contribue massivement. Les choix techniques ne permettent pas non plus une amélioration des performances de contribution, ils agissent simplement comme un filtre encourageant les contributeurs alignés avec les attentes du système.

c. Une structuration sociale calibrée à un écosystème clivant

Les résultats précédents montrent que la structuration communautaire du dispositif ne correspond pas du tout à un espace de délibération démocratique. Cette structuration est centrée sur une identification de la faible minorité d'utilisateurs qui va produire un contenu aligné avec les normes implicites de la communauté. Elle repose également sur des choix techniques d'exclusion, comme la perte automatique du droit de rédaction en dessous d'une certaine performance. Cette fonctionnalité, combinée au découragement de certains de ne jamais voir leurs propositions de notes validées, aboutit à démobiliser les rédacteurs qui ne sont pas alignés avec les attentes du système. En dehors de cette élite très minoritaire, qui ne constitue que près de 1 % de notre population, l'ensemble des autres contributeurs n'a qu'un rôle d'appoint et ne sert qu'à identifier le contenu relativement consensuel. La fonction dans le système contributif d'une majorité des membres actifs n'est que d'exprimer leurs biais individuels pour permettre à l'algorithme d'identifier le contenu consensuel. Le principe mis en œuvre suppose cependant qu'une intersection idéologique existe entre les votants pour que puisse être identifié le contenu à afficher. C'est ce qui explique l'exclusion du système de vote des utilisateurs dont les habitudes contributives les identifient comme marginaux. Le dispositif comporte donc la particularité intéressante de nécessiter des individus biaisés, mais d'être rendu inopérant s'ils le sont trop. La consensualité parmi des groupes ayant des biais différents fonctionne comme une garantie de pertinence et de neutralité d'une note, mais nécessite de contrôler le niveau d'hétérogénéité de ces groupes, au risque que les

contenus consensuels soient rarissimes. Notons que ces considérations théoriques sont à nuancer en pratique dans la mesure où les seuils et conditions de consensualité sont actuellement relativement bas dans le programme du système (cf. annexe 3). De plus, ce principe reste pertinent pour du contenu relativement neutre comme de la dénonciation de plagiat ou de désinformation flagrante. Il peut cependant devenir problématique sur des corrections de contenu politique, qui représentent, comme nous l'avons vu, près de la moitié des corrections dans l'espace francophone. La structuration sociale des contributeurs qui découle des choix techniques répond donc pleinement et uniquement à un objectif de performance du système. Elle repose sur les hypothèses que le débat est impossible, que les gens ne changeront pas d'avis, que la très grande majorité des individus ont des biais idéologiques importants et que seule une très faible minorité de contributeurs pourra rédiger du contenu relativement neutre. Dans le cadre de ces hypothèses, la solution technique mise en œuvre permet de structurer efficacement les acteurs pour obtenir des résultats relativement neutres et consensuels. Un calcul constant des performances des rédacteurs permet d'encourager la faible minorité de contributeurs performants à rédiger des notes. L'algorithme de consensus transverse utilise les biais idéologiques de l'ensemble des contributeurs pour analyser leurs votes et identifier les notes acceptées par des groupes idéologiques différents. Enfin, le système exclut les individus qui ne sont pas nécessaires à son bon fonctionnement, à savoir ceux dont les habitudes de votes sont trop différentes de celles de la majorité et qui ne sont donc pas utiles pour identifier les notes consensuelles. Cette structuration en « trois couches sociales », même si elle crée une production collective qui est aux antipodes des principes de la délibération habermassienne, semble être adaptée à un environnement très polarisé comme X.

Conclusion générale

L'analyse conduite dans ce mémoire nous a permis d'explorer en profondeur un dispositif de vérification d'information implémenté dans un environnement militant et polarisé. À travers l'exemple des notes de la communauté, nous avons pu étudier en détail les ressorts sociotechniques qui régissent le fonctionnement de ce type de système. Notre objectif était de comprendre comment il parvenait à produire des résultats considérés comme en très grande partie neutres et impartiaux alors qu'il est régi par un système contributif au sein d'un environnement particulièrement clivant.

Nous avons commencé par démontrer qu'une notion d'utilité est construite par les choix individuels des contributeurs. Ces choix sont guidés par les attentes explicites du système, qui simplifie au maximum les actions contributives. Nous avons ensuite démontré que ces choix individuels, qu'il s'agisse de l'évaluation d'une note ou de la décision d'annoter une publication, sont inscrits dans un ensemble de normes implicites partagées, issues d'une combinaison complexe de facteurs. Ces normes émergent de dynamiques sociales, de biais idéologiques ou encore du statut perçu des individus corrigés. Elles conduisent à ce que soient valorisés la neutralité, la dénonciation frontale des tromperies ou encore le standard d'une écriture légitime. Nous avons vu que ces normes s'appliquent également au contexte d'énonciation des informations corrigées, c'est-à-dire aux tweets jugés trompeurs. Le statut de l'émetteur ou sa légitimité institutionnelle ont un effet significatif sur les normes correctives. Le contenu politique constitue la majorité du contenu ciblé et révèle un biais idéologique au sein de la communauté francophone. Le degré de consensualité de l'utilité d'annoter certains types de contenu, comme la désinformation scientifique, est très élevé, là où d'autres catégories de publications, comme celles humoristiques ou liées au divertissement, constituent une *zone grise* corrective.

Après avoir analysé les normes dans lesquelles s'inscrivent les contributeurs à l'échelle individuelle, nous avons étudié le système dans son ensemble, par l'étude de sa structuration algorithmique et sociale. L'analyse détaillée du fonctionnement de l'algorithme a montré que le consensus est créé par une recherche des contenus qui vont être approuvés par des groupes de contributeurs idéologiquement différents. La mise en œuvre de ce principe implique un filtrage massif des contributions et des choix structurels partiaux qui

réduisent l'espace délibératif et invisibilisent les idées marginales. Nous avons fini par étudier le rôle de la structuration sociale des contributeurs dans les performances du système. Il est apparu que les contributeurs sont très hiérarchisés en droits et en production contributive. Le choix des individus autorisés à contribuer au système se fait sur des critères purement fonctionnels et non délibératifs. Les choix techniques aboutissent à une répartition des rôles très inégalitaire mais fonctionnelle dans un espace polarisé. Elle comprend une élite qui rédige les notes, une masse biaisée qui a pour rôle de permettre l'identification des consensus, et des marginaux qui sont exclus.

Le système des notes de la communauté doit donc son efficacité apparente à deux éléments. Le premier est l'ensemble de normes explicites, et surtout implicites, qui guident les actions individuelles des contributeurs vers des contributions qui vont favoriser les contenus neutres et pertinents. Le second, et plus important, est le traitement algorithmique qui mesure, classifie et valide les opinions des contributeurs, en cherchant les intersections idéologiques.

Très éloigné des principes de la délibération émancipatrice théorisés par Jürgen Habermas, le dispositif des notes de la communauté confie pleinement le rôle de la délibération à un algorithme. A l'opposé d'autres systèmes contributifs dans lesquels l'échange et les débats sont au cœur des productions collectives, les contributions des membres ne sont ici que la matière première d'un algorithme qui établit ce qui est consensuel ou ne l'est pas. L'efficacité apparente du système est obtenue en sacrifiant la présence d'échanges, de travail commun, d'égalité contributive et de prise en compte des courants marginaux.

Ces résultats soulèvent plusieurs pistes pour prolonger l'analyse. Tout d'abord, une étude plus détaillée du profil des contributeurs pourrait compléter le premier chapitre. Analyser l'influence de facteurs sociaux et idéologiques sur leurs contributions ou déterminer plus précisément leurs motivations pourrait être particulièrement éclairant. Étudier l'influence réelle des différents mécanismes techniques et algorithmiques mis en évidence serait également utile pour comprendre s'il est possible de conserver les bonnes performances du dispositif tout en rendant l'espace contributif plus démocratique et délibératif.

Bibliographie

Références non académiques :

Community Notes. (n.d.). Valeurs fondamentales des contributeurs. X. https://communitynotes.x.com/guide/fr/contributing/values

Congressional Research Service. Liability for Online Content: An Overview of Section 230 of the Communications Decency Act [En ligne]. https://sgp.fas.org/crs/misc/LSB10082.pdf

Croquet, P., & Szadkowski, M. (2025). Les « community notes », un outil de modération communautaire à double tranchant. *Le Monde*, 9 janvier 2025.

Pew Research Center (2019). Republicans Far More Likely Than Democrats To Say Fact Checkers Tend To Favor One Side [En ligne]. https://www.pewresearch.org/short-reads/2019/06/27/republicans-far-more-likely-than-democrats-to-say-fact-checkers-tend-to-favor-one-side/

U.S. Congress. S.2972 — Safe Tech Act [En ligne]. https://www.congress.gov/bill/117th-congress/senate-bill/2972

Références académiques :

Abramowitz A. I., Saunders K. L. et al. (2015), Is Polarization a Myth?, *The Journal of Politics*, DOI: 10.1017/S0022381608080493.

Allcott, H., et Gentzkow, M., (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*. vol. 31, n° 2, p. 211-236.

Allen, J., et Martel, C., (2022). Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program. *CHI Conference on Human Factors in Computing Systems.*

Arazy, O., et Nov, O., (2011). Information Quality in Wikipedia: The Effects of Group Composition and Task Conflict. *Journal of Management Information Systems*. vol. 27, n° 4, p. 71-98.

Borwankar, S., et Zheng, J., (2022). Democratization of Misinformation Monitoring: The Impact of Twitter’s Birdwatch Program. *SSRN Electronic Journal.*

Bourdieu, P. (1979). *La distinction : Critique sociale du jugement*. Paris : Éditions de Minuit.

Bourdieu, P. (1976). *Le champ scientifique*. Actes de la recherche en sciences sociales, 2(2-3), 88-104.

Bovermann, M., (2024). Putting X's Community Notes to the Test. *Verfassungsblog*. 08 janvier 2024.

Boyadjian, S., (2020). Désinformation, non-information ou sur-information ? Les usages différenciés des réseaux sociaux dans la jeunesse étudiante. *Réseaux*. vol. 38, n° 222, p. 25-56.

Cardon, D., (2019). Pourquoi avons-nous si peur des fake news ? *Médias.* n° 59-60, p. 96-108.

Cho, H., Chen, M., et al., (2010). Testing an Integrative Theoretical Model of Knowledge-Sharing Behavior in the Context of Wikipedia. *Journal of the American Society for Information Science and Technology*. vol. 61, n° 6, p. 1198-1212.

Chuai, Y., Pilarski, M., et al., (2024). Community notes reduce the spread of misleading posts on X.

Chuai, Y., Sergeeva, A., et al., (2024). Community Fact-Checks Trigger Moral Outrage in Replies to Misleading Posts on Social Media.

Chuai, Y., Tian, H., et al., (2024). Did the Roll-Out of Community Notes Reduce Engagement With Misinformation on X/Twitter? *Proceedings of the ACM on Human-Computer Interaction*. vol. 8.

Clarke, J., Chen, H., et al., (2020). Fake News, Investor Attention, and Market Reaction. Information Systems Research. p. 1-18.

Drolsbach, C. P., Solovev, K., et al., (2024). Community notes increase trust in fact-checking on social media. *PNAS Nexus*. vol. 3, n° 7.

Fallis, D., (2008). Toward an Epistemology of Wikipedia. *Journal of the American Society for Information Science and Technology.* vol. 59, n° 10, p. 1662-1674.

Fiorina M. P., Abrams S. A. et al. (2008), Polarization in the American Public: Misconceptions and Misreadings, *The Journal of Politics*, DOI: 10.1017/S002238160808050X.

Forte, A., Larco, V., et al., (2009). Decentralization in Wikipedia Governance. *Journal of Management Information Systems.* vol. 26, n° 1, p. 49-72.

Gao, Y., Zhang, M. M., et al., (2024). Can Crowdchecking Curb Misinformation? Evidence from Community Notes.

Godel, W., Sanderson, Z., et al., (2022). Moderating with the Mob: Evaluating the Efficacy of Real-Time Crowdsourced Fact-Checking. *Journal of Online Trust and Safety*. vol. 1, n° 1, p. 1-36.

Grinberg, N., Joseph, K., et al., (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science*. vol. 363, n° 6425, p. 374-378.

Habermas, J. (1992). *Droit et démocratie* : Entre faits et normes. Paris: Gallimard, 1997

Habermas, J. (1983). *Morale et communication* : Conscience morale et activité communicationnelle. Paris : Éditions du Cerf.

Habermas, J. (1981). *Théorie de l'agir communicationnel*. Paris : Fayard, 1987.

Harsin, J., (2018). Un guide critique des fake news : de la comédie à la tragédie. *Pouvoirs*. n° 164, p. 99-119.

Iyengar S., Sood G. et al. (2012), Affect, not ideology: A social identity perspective on polarization, *Public Opinion Quarterly*, DOI: 10.1093/poq/nfs038.

Kangur, U., Chakraborty, R., et al., (2024). Who Checks the Checkers? Exploring Source Credibility in Twitter's Community Notes.

Kim, J., Wang, Z., et al., (2025). Differential impact from individual versus collective misinformation tagging on the diversity of Twitter (X) information engagement and mobility. *Nature Communications*. vol. 16, n° 973.

Li, D., Xu, B., et al., (2014). An Empirical Study of Motivations for Content Contribution and Community Participation in Wikipedia. *Information and Management*.

Loomba, S., de Figueiredo, A., et al., (2021). Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature Human Behaviour*. vol. 5, n° 3, p. 337-348.

Martel, C., Rand, D. G., et al., (2024). Misinformation warning labels are widely effective: A review of warning effects and their moderating features. *Current Opinion in Psychology*.

Nov, O., (2007). What Motivates Wikipedians? *Communications of the ACM*. vol. 50, n° 11, p. 60-64.

Nyhan, B., Reifler, J., et al., (2010). When Corrections Fail: The Persistence of Political Misperceptions. *Political Behavior*. vol. 32, n° 2, p. 303-330.

Oh, O., Agrawal, M., et al., (2013). Community Intelligence and Social Media Services: A Rumor Theoretic Analysis of Tweets During Social Crises. *MIS Quarterly*. vol. 37, n° 2, p. 407-426.

Osmundsen, M., Bor, A., et al., (2021). Partisan polarization is the primary psychological motivation behind political fake news sharing on Twitter. *American Political Science Review*. vol. 115, n° 3, p. 999-1015.

Ouzilou, O. (2009). Niklas Luhmann, La confiance, un mécanisme de réduction de la complexité sociale. *Revue Interrogations*.

Ovadya, A. (2022). Bridging-Based Ranking: How Platform Recommendation Systems Might Reduce Division and Strengthen Democracy [En ligne]. https://www.belfercenter.org/project/technology-and-public-purpose

Pennycook, G., Rand, D. G., et al., (2019). Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*. vol. 116, n° 7, p. 2521-2526.

Phillips, S. C., Wang, S. Y. N., et al., (2024). Emotional language reduces belief in false claims.

Pilarski, M., Solovev, K. O., et al., (2024). Community Notes vs. Snoping: How the Crowd Selects Fact-Checking Targets on Social Media. *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*. vol. 18, p. 1262-1272.

Renault, T., Restrepo-Amariles, D., et al., (2024). Collaboratively Adding Context to Social Media Posts Reduces the Sharing of False News.

Schroeder, A., Wagner, C., et al., (2012). Governance of Open Content Creation: A Conceptualization and Analysis of Control and Guiding Mechanisms in the Open Content Domain. *Journal of the American Society for Information Science and Technology*. vol. 63, n° 10, p. 1947-1959.

Spinellis, D., Louridas, P., et al., (2008). The collaborative organization of knowledge. *Communications of the ACM*. vol. 51, n° 8, p. 68-74.

Stvilia, B., Twidale, M., et al., (2008). Information Quality Work Organization in Wikipedia. *Journal of the American Society for Information Science and Technology*. vol. 59, n° 6, p. 983-1001.

Surowiecki, J. (2004). *The Wisdom of Crowds*. New York : Anchor Books.

Swire-Thompson, B., DeGutis, J., et al., (2020). Searching for the Backfire Effect: Measurement and Design Considerations. *Journal of Applied Research in Memory and Cognition*. vol. 9, p. 286-299.

Vauchez, Y., (2022). La crédulité des crédules. Débat public et panique morale autour des fake news en France. *Émulations*. n° 41.

Warner, M. (2023). Legislation to Reform Section 230 Reintroduced in the Senate, House [En ligne]. https://www.warner.senate.gov/public/index.cfm/2023/2/legislation-to-reform-section-230-reintroduced-in-the-senate-house

Yao, M., Tian, S., et al., (2024). Readable and neutral? Reliability of crowdsourced misinformation debunking through linguistic and psycholinguistic cues. Frontiers in Psychology. vol. 15, n° 1478176.

Yarchi M., Badeb C. et al. (2020), Political Polarization on the Digital Sphere: A Cross-platform, Over-time Analysis of Interactional, Positional, and Affective Polarization on Social Media, *Political Communication*, DOI: 10.1080/10584609.2020.1785067.

Annexes

Table des annexes :

Annexe n°1 : Interface de rédaction des notes de la communauté

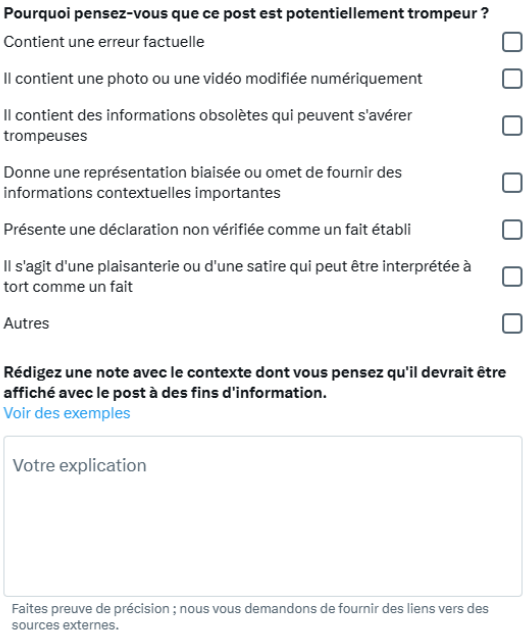

Figure n°35 : Capture d'écran de l'interface de rédaction des notes de la communauté (Source : X.com)

Annexe n°2 : Résultats du sondage relatif aux questions sur le retrait des fausses informations

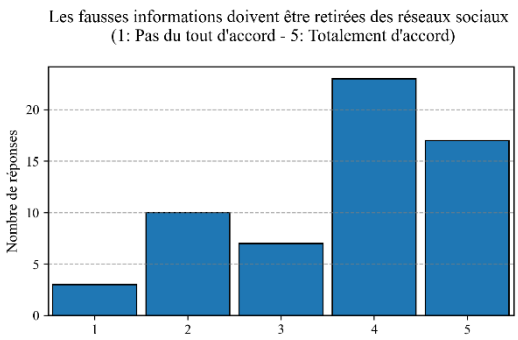

Figure n°36 : Résultat du sondage – Question du retrait des fausses informations des réseaux (Source : établi par l'auteur)

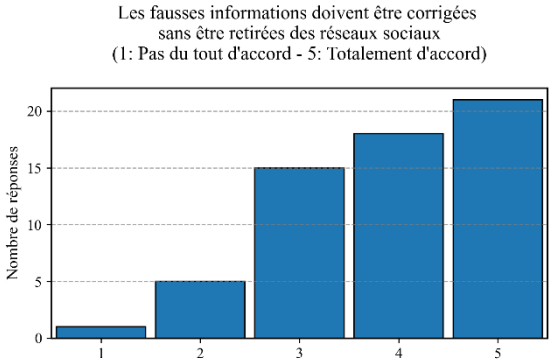

Figure n°37 : Résultat du sondage – Question de la correction sans retrait des fausses informations des réseaux (Source : établi par l'auteur)

Annexe n°3 : Vulgarisation de l'algorithme bridging-based

Initialement, le système considère une matrice contenant toutes les évaluations. Cette matrice peut être vue comme un très grand tableau dont les colonnes sont l'ensemble des contributeurs et les lignes l'ensemble des notes. Pour chaque vote réalisé par un contributeur, le tableau contient 1 pour « utile » ou 0 pour « non utile » à l'intersection de la ligne de la note et de la colonne du contributeur.

Ensuite, deux variables sont attribuées à toutes les notes et à tous les contributeurs : Un **intercept** et un **facteur.**

L'objectif est que ces valeurs puissent modéliser les votes dans le tableau, qui sont définis par l'équation suivante :

$$Utilité = Int_{note} + Int_{contrib} + Fact_{note} \times Fact_{contrib} + (terme\ correctif)$$

L'apprentissage va donc chercher à estimer les valeurs des intercepts et des facteurs pour l'ensemble des notes et des contributions, de manière à ce que, pour chaque case du tableau remplie, la formule ci-dessus parvienne à estimer la valeur réelle du vote (0 pour non utile et 1 pour utile).

L'apprentissage automatique peut se comprendre comme le fait d'attribuer initialement des valeurs aléatoires à ces valeurs, puis de les modifier progressivement dans une direction qui garantit de se rapprocher de la valeur qu'on cherche à prédire, jusqu'à obtenir un résultat satisfaisant.

Une fois l'apprentissage terminé, on obtient un couple de valeur pour chaque note et chaque utilisateur. La différence fondamentale entre les intercepts et les facteurs est que, dans l'équation, les valeurs d'intercept de la note et du contributeur sont indépendantes l'une de l'autre (car simplement additionnées). À l'inverse, les valeurs de facteurs de la note et de l'évaluateur sont multipliées entre elles, c'est donc leur combinaison qui influe sur l'estimation globale de l'utilité. Elles correspondent donc au degré de proximité de l'évaluateur et de la note évaluée. Ce degré de proximité peut se comprendre comme un type de contenu préféré par le contributeur, un style d'argumentation ou une sensibilité particulière pour le contenu de la note. On peut donc interpréter nos quatre variables de la manière suivante :

- Intercept de la note : la valeur moyenne de l'utilité de la note perçue par les contributeurs.
- Intercept du contributeur : le niveau de sévérité du contributeur dans ses évaluations.
- Le facteur de la note : son positionnement dans l'espace des styles ou des types de correction.
- Le facteur du contributeur : son profil de jugement implicite des notes.

L'utilité finale d'une note est définie par la valeur de son intercept. La force de l'équation présentée précédemment est que, pour prédire efficacement les notations, la valeur choisie d'intercept de la note sera d'autant plus grande que le produit entre les facteurs sera faible. Un produit entre facteurs faibles est le signe d'un écart idéologique important entre la note et le correcteur. La valeur de l'intercept de la note contient structurellement le degré de consensualité de sa prise de position. Plus précisément, elle est élevée si une majorité des évaluateurs en sont idéologiquement éloignés, mais ces évaluateurs peuvent toujours être homogènes entre eux.

Pour pallier cette limite et se rapprocher d'un système évaluant la consensualité, le système réalise un clustering des contributeurs, c'est-à-dire qu'un groupe est attribué à chaque

contributeur. Ce regroupement est réalisé sur la base des valeurs de facteurs associés à chaque contributeur qui, nous le rappelons, correspondent aux habitudes de notation de l'évaluateur. Les utilisateurs ayant des valeurs de facteur proches sont regroupés au sein du même cluster. Pour qu'une note soit définitivement acceptée, il faut premièrement qu'elle dépasse un certain score d'intercept, mais également que les contributeurs ayant voté en sa faveur proviennent d'au moins trois groupes différents. Le système comporte trois familles de groupes, c'est-à-dire trois modèles de regroupement différents comportant chacun entre 5 et 15 clusters. Les notes dont le score d'intercept a dépassé le seuil fixé par le système doivent comporter des votes positifs d'utilisateurs provenant d'au moins trois clusters différents pour être jugées valides. Cette condition doit être remplie pour au moins un des trois modèles pour qu'elle soit affichée. D'autres conditions annexes que nous ne détaillerons pas ici s'ajoutent finalement pour confirmer l'utilité de la note.

Annexe n°4 : Résultat du sondage des utilisateurs de X – Question sur la connaissance du fonctionnement de l'algorithme

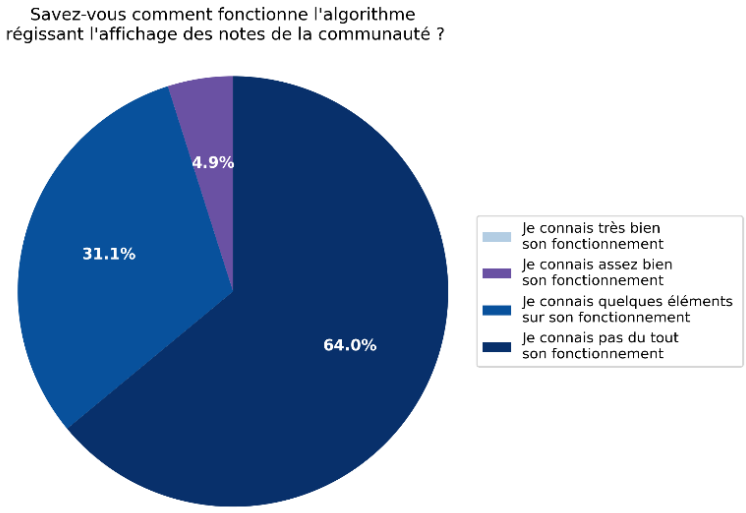

Figure n°38 : Résultat du sondage – Connaissance du fonctionnement de l'algorithme d'affichage des notes (Source : établi par l'auteur)

Table des matières

Table des tableaux, graphiques et illustrations

Résumé

Ce mémoire de recherche examine le système de vérification d'information collaboratif des « notes de la communauté », mis en place sur le réseau social X., Notre objectif est de comprendre comment émerge, dans un environnement fortement polarisé, un système correctif perçu comme relativement fiable. Pour ce faire, nous combinons une étude qualitative des corrections, une analyse approfondie de l'algorithme de validation des notes et des analyses des données du système. Ces analyses révèlent que le relatif bon fonctionnement du système repose sur l'émergence d'une norme d'utilité commune dans l'espace des contributeurs et sur des choix techniques créant les conditions d'un filtrage algorithmique qui détermine seul la consensualité. L'étude souligne également les limites de ce modèle participatif, notamment l'absence de véritable délibération et la marginalisation des points de vue minoritaires.

Mots clefs

Réseaux sociaux ; fact-checking ; fausses informations ; notes de la communauté ; X/Twitter ; système participatif ; polarisation.

Abstract

This research paper examines the collaborative fact-checking system known as "Community Notes," implemented on the social network X. Our objective is to understand how a corrective system perceived as relatively reliable emerges in a highly polarized environment. To achieve this, we combine a qualitative study of corrections, an in-depth analysis of the note validation algorithm, and data analyses of the system. These analyses reveal that the system's relative effectiveness relies on the emergence of a common utility norm among contributors and on technical choices that create conditions for algorithmic filtering, which determines consensus. The study also highlights the limitations of this participatory model, particularly the absence of genuine deliberation and the marginalization of minority viewpoints.

Keywords

Social media ; Fact-checking ; misinformation ; community notes ; X/Twitter ; crowdsourced system ; polarization.