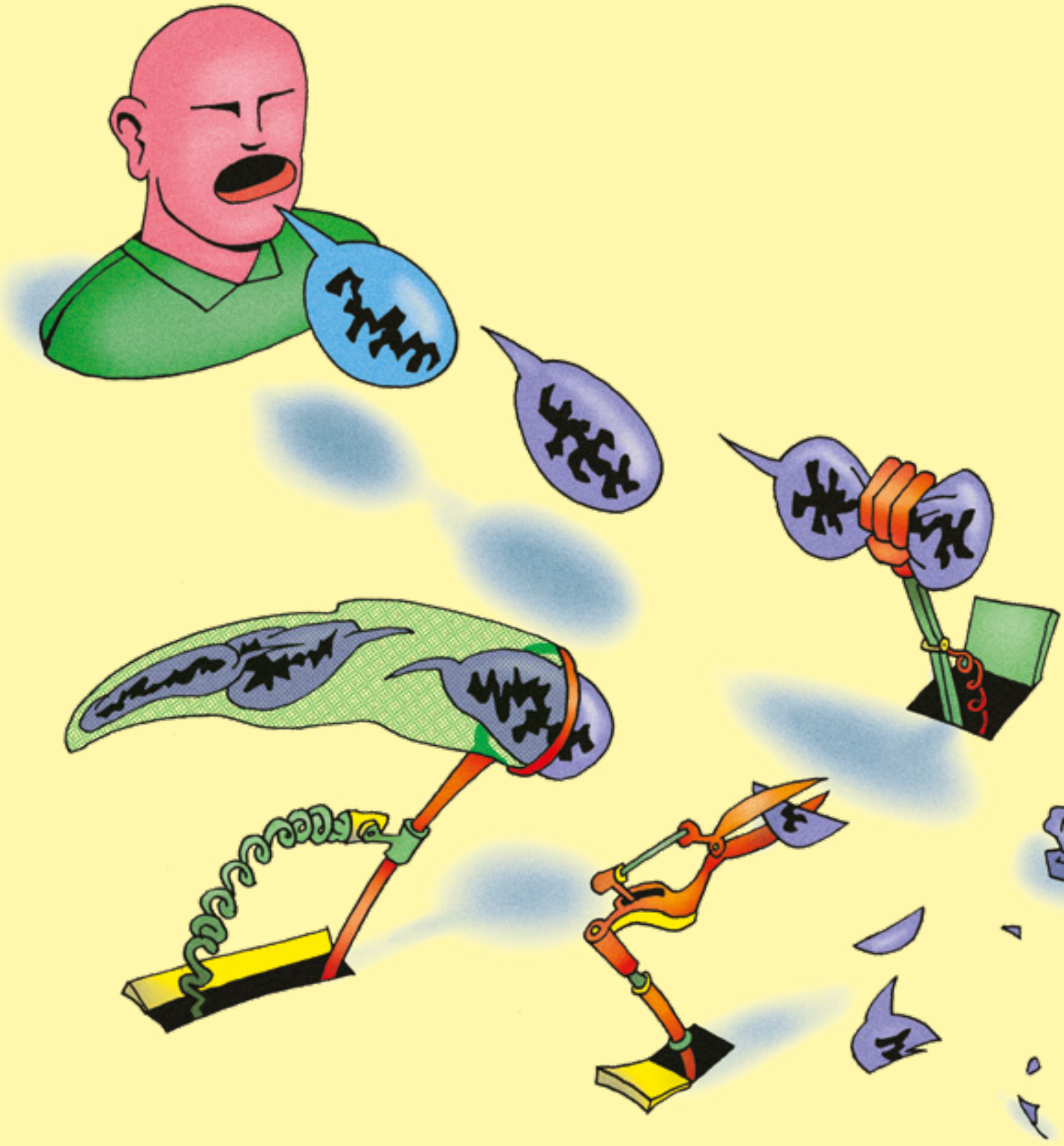
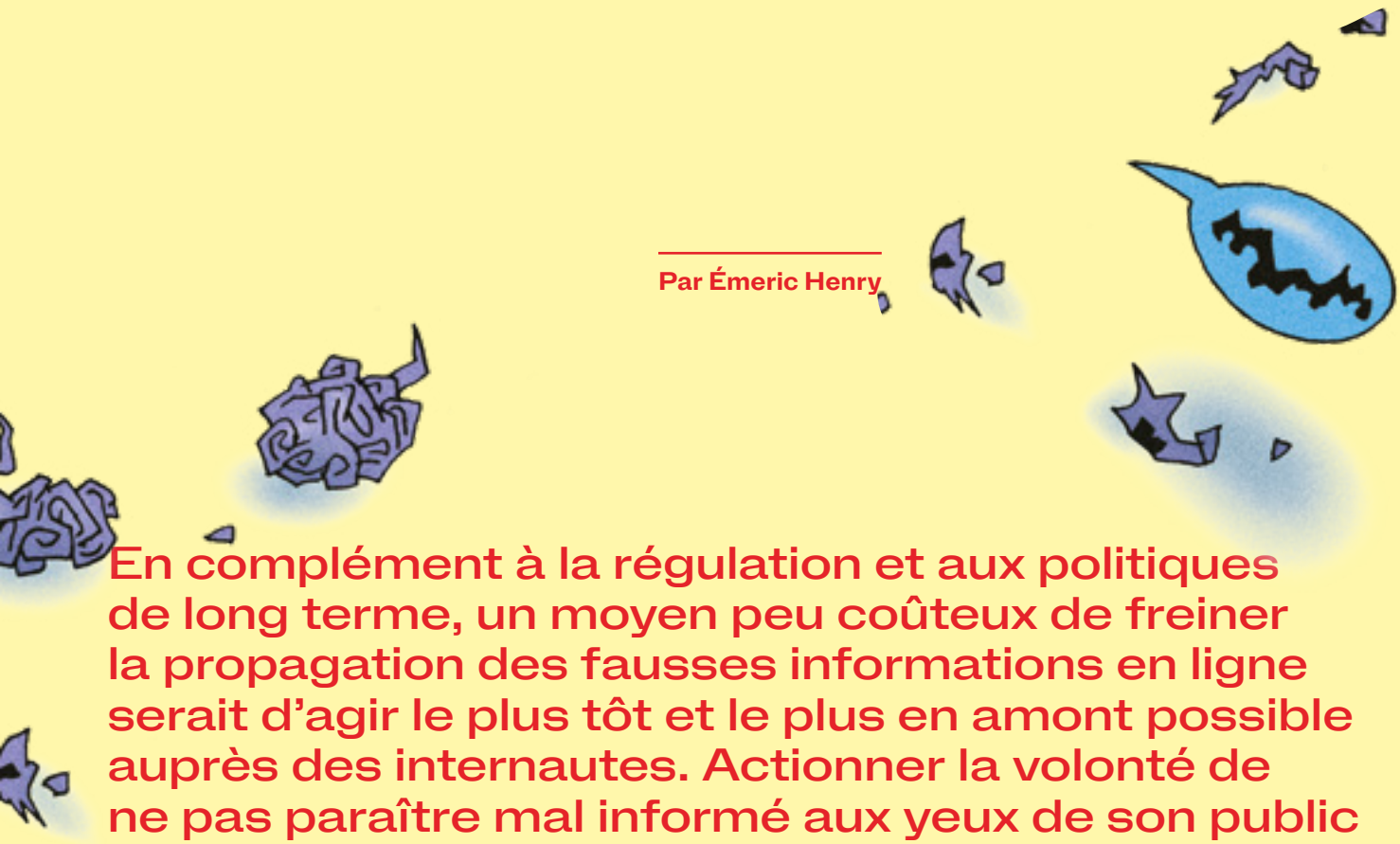


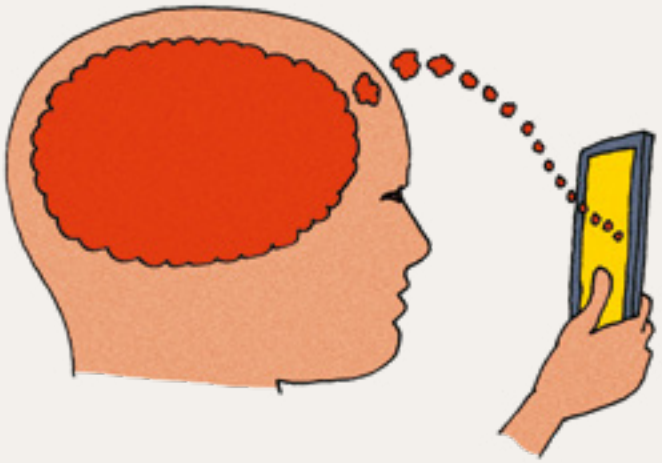
Microéconomiste, Émeric Henry est directeur du Département d'économie. Ses recherches portent essentiellement sur l'économie de l'innovation, l'économie digitale et l'économie politique, pour lesquelles il combine théorie et méthodes expérimentales et empiriques.

Combattre la désinformation sur les réseaux sociaux

Par Émeric Henry



En complément à la régulation et aux politiques de long terme, un moyen peu coûteux de freiner la propagation des fausses informations en ligne serait d'agir le plus tôt et le plus en amont possible auprès des internautes. Actionner la volonté de ne pas paraître mal informé aux yeux de son public et donc de ne pas nuire à sa réputation pourrait être un levier efficace, comme le montrent les différents traitements testés auprès d'un groupe d'internautes lors d'une enquête empirique récente à laquelle a contribué Émeric Henry.



L'équilibre entre la lutte contre la désinformation et la protection de la liberté d'expression est délicat à trouver.

Les réseaux sociaux ont fondamentalement modifié la manière dont nous interagissons, communiquons et accédons à l'information. Leur potentiel de propagation de la désinformation est une préoccupation majeure, tant pour les citoyens que pour les personnels politiques. Les fausses informations politiques sont fréquentes sur des plateformes comme Facebook, X/Twitter et Reddit, une réalité d'autant plus inquiétante qu'une part substantielle des utilisateurs s'appuie sur elles pour s'informer.

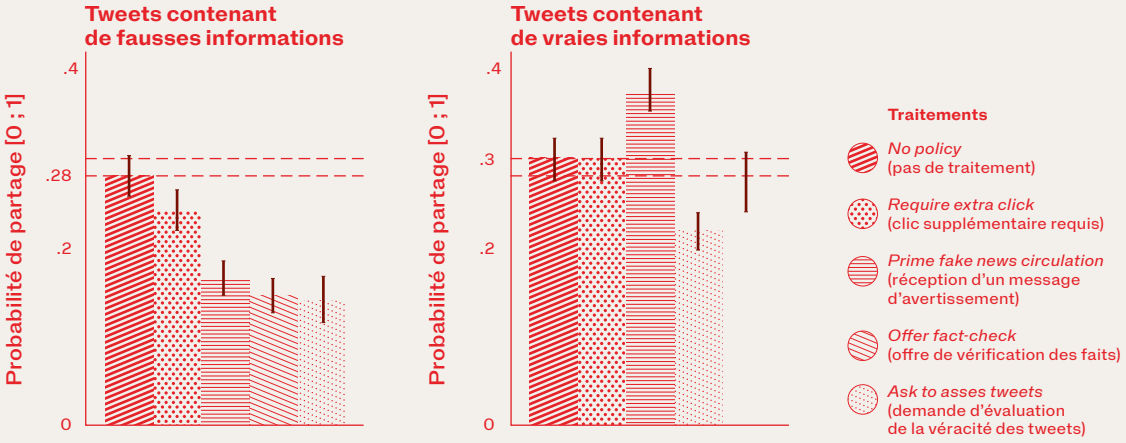
De manière générale, l'équilibre est délicat à trouver entre la lutte contre les fausses informations et la protection de la liberté d'expression. Aux États-Unis, les limites constitutionnelles entravent la régulation de la modération des contenus. L'Union européenne prévoit bien, quant à elle, une régulation des plateformes via le *Digital Services Act* (DSA), toutefois la lutte se concentre pour l'instant sur les contenus illégaux et exclut une part importante de la désinformation politique. Certains chercheurs prônent la mise en place de programmes d'éducation numérique pour apprendre aux citoyens à distinguer les informations exactes des fausses nouvelles et ainsi combattre le phénomène sur le long terme.

Une tout autre approche pourrait être d'intervenir le plus tôt possible auprès des utilisateurs, c'est-à-dire avant même qu'ils décident de partager ou non des contenus sur les réseaux sociaux. Une telle politique serait moins coûteuse et certaines de ses composantes, faciles à mettre en œuvre. Il pourrait s'agir d'exiger des clics de confirmation au moment de la décision de partager, de pousser les utilisateurs à réfléchir aux conséquences de leur partage de fausses informations via un type d'intervention appelé *nudge* en anglais et dont Gordon Pennycook, psychologue, et David Rand, professeur de sciences de gestion, cérébrales et cognitives, ont récemment démontré l'efficacité. Ce pourrait être aussi de proposer de la vérification des faits (*fact-checking*), une politique déjà déployée par certaines plateformes.

Comment inciter à réfléchir avant de partager ?

Quelle efficacité auraient ces différentes interventions ? Quels mécanismes actionneraient-elles ? Une étude expérimentale récente, intitulée « Curtailing False News, Amplifying Truth » (Restreindre les fausses informations, amplifier la vérité), apporte des éléments de réponses. Menée par Sergeï Guriev, Émeric Henry, Théo Marquis et Ekaterina Zhuravskaya lors de la campagne des législatives de mi-mandat de 2022 aux États-Unis, elle a évalué l'impact de différents types de traitement sur la circulation tant des fausses informations que des vraies. L'étude a exposé à 3 501 participants étatsuniens, utilisateurs de X/Twitter, à quatre tweets d'information politique : deux comportant de la désinformation, deux autres comportant des faits véridiques.

EFFETS DES DIFFÉRENTS TRAITEMENTS DE LUTTE CONTRE LA DÉSINFORMATION



Source : S. Guriev, E. Henry, T. Marquis et E. Zhuravskaya, « Curtailing False News, Amplifying Truth », CEPR Discussion Paper No. 18650, 2023, https://cepr.org/publications/dp18650

Les participants, qui devaient décider de partager ou non l'un ou plusieurs de ces tweets sur leur compte, ont été aléatoirement répartis en groupes afin de recevoir différents traitements. Dans un premier groupe (groupe de contrôle appelé *No policy*), ils pouvaient faire ce qu'ils voulaient avec ces quatre tweets. Dans un deuxième groupe (*Require extra click*), ils devaient effectuer un clic de plus afin de confirmer leur décision de partage, soit une action un peu fastidieuse. Dans un troisième groupe (*Prime fake news circulation*), ils recevaient, avant de pouvoir partager, un message d'avertissement (*nudge*) inspiré des incitations proposées par Pennycook et Rand : « Veuillez réfléchir attentivement avant de retweeter. Rappelez-vous qu'une quantité significative de fausses nouvelles circulent sur les réseaux sociaux. » Dans un quatrième groupe (*Offer fact-check*), ils étaient informés que deux tweets contenaient des fausses informations détectées par PolitiFact.com, ONG réputée en vérification des faits, et recevaient le lien pour accéder au *fact-checking*. À la fin de l'enquête, tous les participants ont été invités à évaluer la véracité et l'inclinaison partisane de chaque tweet.

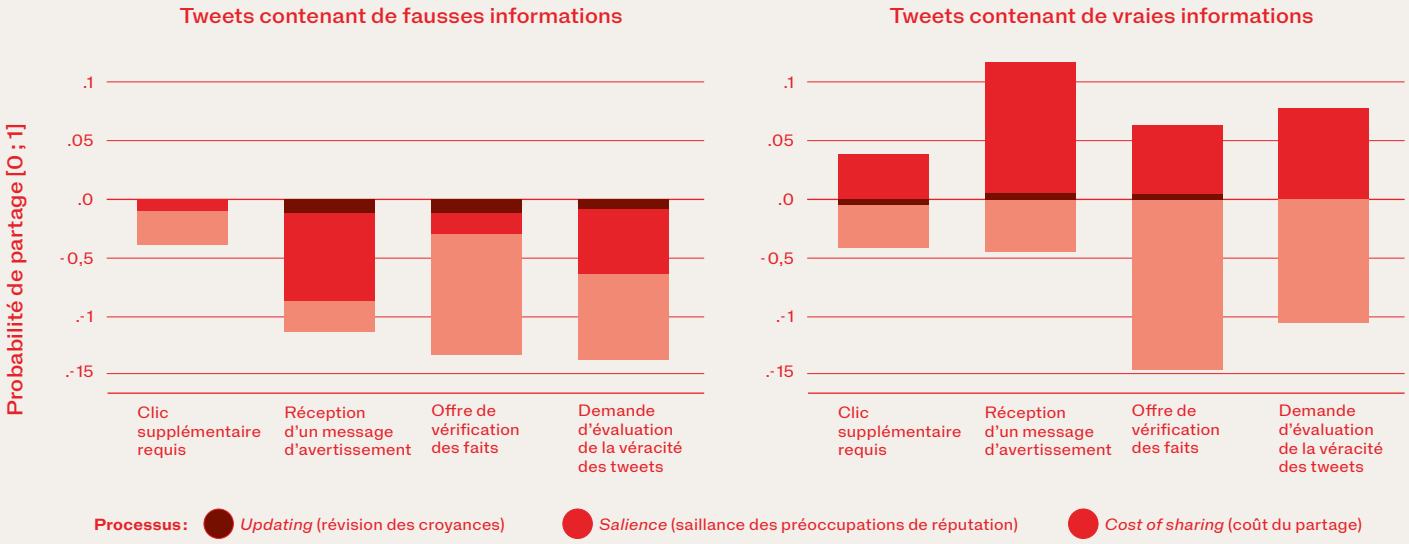
La figure ci-dessus illustre les effets des différents traitements sur le partage des fausses informations (panel de gauche) et vraies (panel de droite). On y constate que tous les traitements ont contribué à réduire le taux de partage des fausses informations. Dans les groupes *Require extra-click*, *Prime fake news circulation* et *Offer fact-check*, les taux de partage sont respectivement de 3,6, 11,5 et 13,6 points inférieurs à celui du groupe de contrôle, sachant que 28 % des membres de ce dernier ont partagé un des tweets contenant de fausses informations. Cependant, tous les traitements ne produisent pas les mêmes effets quant au taux de partage des informations véridiques, qui est de 30 % dans le groupe de contrôle : demander un clic supplémentaire avant de partager n'a aucun effet discernable, offrir l'accès à un *fact-check* diminue le partage de tweets véridiques de 7,8 points de pourcentage, mais envoyer un message de prudence comportementale augmente le taux moyen de partage des tweets véridiques de 8,1 points.

Tous ces résultats établissent une hiérarchie claire de l'efficacité des politiques destinées à améliorer la précision des contenus partagés. Le traitement *Prime fake news circulation*, qui pousse les utilisateurs à réfléchir aux conséquences de leur partage de fausses informations, se révèle plus efficace, car il favorise le « discernement du partage » prôné par Pennycook et Rand : il augmente le partage des vraies informations tout en diminuant le partage des fausses.

L'impact majeur des effets de réputation

Pour comprendre les mécanismes sous-jacents aux effets différenciés de ces traitements sur le partage de vraies et fausses informations, l'étude s'est intéressée aux motifs qui incitent les utilisateurs à partager des informations sur les réseaux sociaux. Elle montre que la perception de la véracité renforce le sentiment de l'utilité du partage pour des raisons de réputation, c'est-à-dire la

DÉCOMPOSITION DES EFFETS DES TRAITEMENTS DIFFÉRENCIÉS DE LUTTE CONTRE LA DÉSINFORMATION



Source : S. Guriev, E. Henry, T. Marquis et E. Zhuravskaya, « Curtailing False News, Amplifying Truth », CEPR Discussion Paper No. 18650, 2023, https://cepr.org/publications/dp18650

La perception de la véracité renforce le sentiment de l'utilité du partage d'informations.

volonté de ne pas paraître mal informé aux yeux de son audience. La concordance des informations avec l'opinion de l'utilisateur accroît aussi son sentiment de satisfaction lorsqu'il les partage, que ce soit pour convaincre son public ou pour signaler son identité politique.

L'étude confirme qu'il est possible d'influencer le partage à travers trois processus, que nous nommerons respectivement : révision des croyances (*updating*), saillance (*salience*) et coût du partage (*cost of sharing*). Le premier processus amène l'utilisateur à réviser ses croyances sur la véracité ou sur l'alignement partisan d'un contenu. Par exemple, l'exposition au *fact-checking* vise à modifier sa perception de l'exactitude des informations. Le deuxième processus augmente la saillance des préoccupations de réputation par rapport aux motifs partisans, de sorte que l'utilisateur accorde plus d'attention qu'auparavant à la véracité des informations lorsqu'il décide de les partager. Les traitements incitant à la prudence, notamment, sont conçus pour affecter cette saillance. Le troisième processus, qui consiste à demander un clic supplémentaire de confirmation, qu'une information soit vraie ou fausse, augmente le coût du partage pour l'utilisateur. L'on voit ainsi que chaque traitement affecte ce coût.

La figure ci-dessus décompose les effets de ces trois processus. De manière surprenante, les traitements destinés à faire réviser les croyances sur la véracité des informations, comme le fact-checking, ont peu d'impact. En réalité, l'effet global de chaque traitement découle d'une combinaison entre la saillance des préoccupations de réputation, les motifs partisans et le coût du partage. La saillance explique en particulier la différence entre les effets des traitements sur le partage des vraies et des fausses informations. Améliorer (ou protéger) sa réputation augmente le partage de vraies informations et réduit le partage de fausses. Tous les traitements, à des degrés divers, augmentent la saillance, celui incitant à la prudence ayant l'effet le plus important. Simultanément, les frictions liées aux différents traitements réduisent tout autant le partage des vraies informations que celui des fausses. Les coûts supplémentaires du traitement incitant à la prudence sont considérablement inférieurs à

ceux de l'offre de *fact-checking*, ce qui rend ce type d'intervention plus efficace pour augmenter le partage de vraies informations.

Une question d'efficience

Les résultats de cette étude ont deux implications pour les politiques de lutte contre la désinformation. Premièrement, ils confirment l'efficacité d'actions à court terme consistant à inciter l'utilisateur à réfléchir aux conséquences de la circulation des fausses informations, comme l'ont préconisé Pennycook et Rand. Cette méthode réduit le partage de fausses informations et augmente celui de faits véridiques, sans diminuer l'engagement global des utilisateurs de réseaux sociaux. Deuxièmement, ces résultats montrent que, grâce au *fact-checking*, l'utilisateur partage moins les fausses informations, non parce qu'il découvre qu'elles sont fausses, mais parce qu'au moment de partager, il prend conscience de la nécessité de vérifier la véracité des faits. Par conséquent, bien qu'entraînant des investissements importants, la vérification des faits par des vérificateurs professionnels pourrait se révéler moins efficace qu'un *fact-checking* réalisé par un algorithme, plus rapide (intervenant plus tôt dans le processus de partage) et moins coûteux, quoique plus sujet à erreurs. Dans ce dernier cas, l'utilisateur est vite informé que le contenu est signalé comme suspect par l'algorithme, ce qui suscite une préoccupation accrue pour la véracité.

Bien entendu, ces politiques à court terme sont complémentaires et non pas substituables à des politiques de long terme comme l'éducation au numérique. L'étude montre d'ailleurs un mécanisme intéressant, qui souligne cette complémentarité : si l'utilisateur, soucieux de sa réputation, sait que son public est plus alerte quant aux questions de désinformation du fait d'une meilleure éducation, cela le rend d'autant moins susceptible de répandre de la désinformation. Notons toutefois que les politiques de court terme sont susceptibles de créer un phénomène d'accoutumance qui risque de réduire leur efficacité. Il pourrait être judicieux de les employer uniquement durant les périodes de risques accrus, comme les campagnes électorales.

Enseigne d'un restaurant de hot dogs à Chicago, après le débat télévisé pour la présidentielle du 11 septembre 2024, durant lequel Donald Trump affirma qu'à Springfield, Ohio, les immigrants enlevaient les chiens et les chats de la ville pour les manger.

RÉFÉRENCES

→ Barrera O., Guriev S., Henry É. et Zhuravskaya E., « Facts, Alternative Facts, and Fact Fhecking in Times of Post-truth Politics », *Journal of Public Economics*, 182, 2020, p. 104-123.

→ Guess A. M., Lerner M., Lyons B., Montgomery J. M., Nyhan B., Reifler J. et Sircar N., « A Digital Media Literacy Intervention Increases Discernment between Mainstream and False News in the United States and India », *Proceedings of the National Academy of Sciences of the United States of America*, 117 (27), 2020, p. 15536-15545.

→ Guriev S., Henry É., Marquis T. et Zhuravskaya E., « Curtailing False News, Amplifying Truth », CEPR Discussion Paper N°. 18650, 2023, https://cepr.org/publications/dp18650

→ Henry É., Zhuravskaya E. et Guriev S., « Checking and Sharing Alt-Facts », *American Economic Journal : Economic Policy*, 14 (3), 2022, p. 55-86.

→ Nyhan B., « Facts and Myths about Misperceptions », *Journal of Economic Perspectives*, 34 (3), 2020, p. 220-236.

→ Pennycook G. et Rand D., « Accuracy Prompts are a Replicable and Generalizable Approach for Reducing the Spread of Misinformation », *Nature Communications*, 13 (2333), 2022.

→ Persily N. et Joshua A. Tucker, *Social Media and Democracy*, Cambridge, Cambridge University Press, 2020.