Comprendre les désordres de l'information : état de l'art des mécanismes, vulnérabilités et réponses dans l'environnement informationnel numérique

Marion Seigneurin

Rapport de recherche
mars 2026

Comprendre les désordres de l'information

État de l'art des mécanismes, vulnérabilités et réponses dans l'environnement informationnel numérique.

Marion Seigneurin

Seigneurin, M. (2026). Comprendre les désordres de l'information : état de l'art des mécanismes, vulnérabilités et réponses dans l'environnement informationnel numérique. Rapport de recherche, Laboratoire LITEM

Résumé exécutif

La mésinformation est aujourd’hui largement perçue comme une menace majeure pour les démocraties contemporaines. Pourtant, son fonctionnement et son impact restent souvent mal compris.

Les débats publics tendent à l'expliquer par la crédulité individuelle ou par la seule prolifération de contenus faux, au détriment d'une analyse plus structurelle des conditions dans lesquelles ces contenus circulent, deviennent visibles et s'installent durablement dans l'espace public numérique.

Ce rapport montre que la mésinformation ne peut être appréhendée comme un phénomène homogène ni comme une simple défaillance individuelle. Elle recouvre une diversité de formes, de supports et de techniques de manipulation, dont l'impact ne dépend ni exclusivement de leur fausseté, ni de leur sophistication technique. Des contenus peu élaborés, voire ambigus, peuvent produire des effets importants lorsqu'ils fonctionnent comme des preuves aux yeux des récepteurs, notamment lorsque leur contexte de production ou de circulation est perdu.

L'analyse met en évidence le rôle central des architectures socio-techniques dans la diffusion et l'amplification de la mésinformation. La visibilité disproportionnée de certains contenus résulte d'effets cumulatifs : relais médiatiques, dynamiques algorithmiques orientées vers la captation de l'attention, stratégies artificielles et coordonnées, mais aussi propagation sociale ordinaire. Ces mécanismes contribuent à transformer des contenus parfois marginaux en phénomènes saillants, indépendamment de leur adhésion réelle dans la population.

Les vulnérabilités informationnelles apparaissent ainsi comme relationnelles et contextuelles. Elles émergent de l'interaction entre des mécanismes cognitifs ordinaires, des dynamiques émotionnelles, des logiques sociales et des environnements numériques conçus pour maximiser l'engagement. Partager de la mésinformation ne signifie pas nécessairement y croire ni vouloir tromper, mais s'inscrire dans des usages sociaux où l'exactitude n'est pas toujours prioritaire.

Face à ces constats, le rapport invite à dépasser une lecture exclusivement centrée sur la correction des fausses croyances, sans pour autant opposer réponses systémiques et interventions individuelles. Les contre-mesures systémiques, en agissant sur la sur-visibilité, la rentabilité et l'amplification artificielle des contenus trompeurs, constituent un levier structurant pour réduire les distorsions de l'espace informationnel, bien qu'elles soulèvent des tensions politiques et économiques importantes. Les interventions individuelles, dont l'efficacité demeure souvent limitée et fortement dépendante du contexte, représentent néanmoins des réponses plus immédiatement activables. Leur mise en œuvre, en particulier lorsqu'elles sont intégrées directement aux architectures des plateformes, en fait aujourd'hui l'un des leviers les plus accessibles à court terme pour atténuer certains

effets de la mésinformation.

En proposant un état de l'art synthétique sur les définitions, les formes, les mécanismes d'amplification et les réponses à la mésinformation dans l'environnement numérique, ce rapport entend éclairer les débats actuels et soutenir une approche pragmatique de la réduction des risques, compatible avec la préservation de la liberté d'expression et le fonctionnement du débat démocratique.

Mécanique de diffusion des désordre informationnels

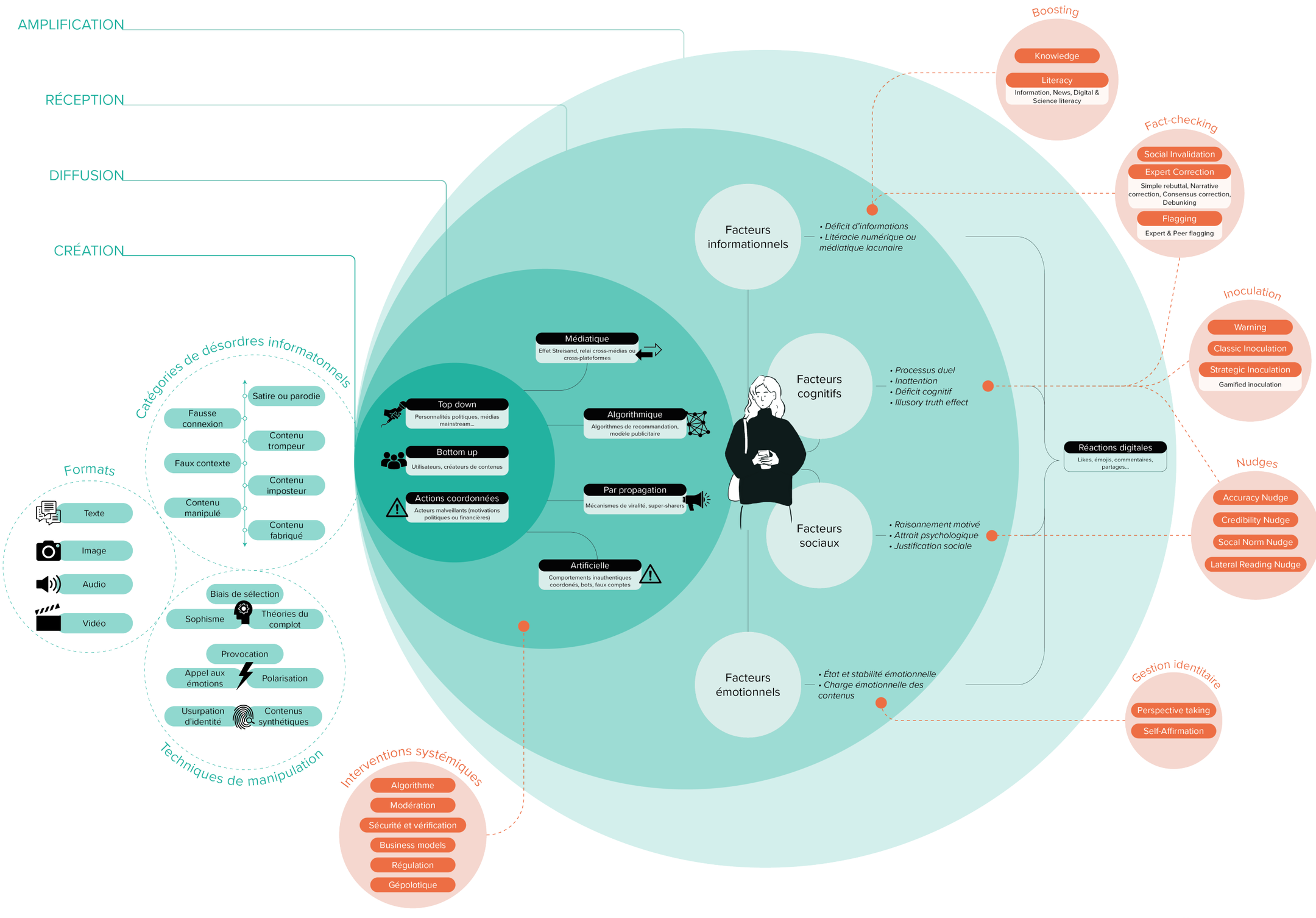


Figure 1 : Mécanique de diffusion des désordres de l'information

Mésinformation, désinformation et autres désordres informationnels

La stabilité de nos démocraties repose sur un principe fondamental : celui de citoyens informés capables d'effectuer des choix éclairés [1]. Aujourd'hui, cette liberté de choix s'exerce lors de processus démocratiques institutionnels tels que les élections, ou dans les actions quotidiennes des citoyens, qu'elles soient directement ou indirectement politiques. Le principe de choix éclairé suppose que les informations dont disposent les citoyens leur permettent de s'impliquer dans la vie publique, d'orienter leurs choix de consommation, ou encore de préserver leur santé et celle de leurs proches. La garantie d'un environnement informationnel sain est donc, avant tout, un enjeu démocratique.

Le partenariat international pour l'information et la démocratie, signé par cinquante sept États, dont la France, établit les principes d'un environnement propice à l'information fiable des citoyens [2]. Dans ce cadre, l'environnement informationnel doit garantir l'exercice de la liberté d'opinion et d'expression, consacré par la Déclaration des droits de l'Homme et du citoyen, ainsi que l'accès à une information libre, plurielle et fiable.

Cependant, si un consensus semble acquis concernant les qualités nécessaires à un environnement informationnel sain, l'opérationnalisation de ces principes, via la distinction entre une information acceptable et une information non acceptable, demeure controversée. La multitude de qualificatifs employés pour désigner les informations entrant en contradiction avec ces principes témoigne de la difficulté à stabiliser une définition commune de l'information non acceptable : fake news, mésinformation, désinformation, propagande, ingérence étrangère, opérations d'influences…

Une vaste littérature cherche ainsi à préciser ces notions, telles qu'elles sont employées par les médias, l'industrie ou les acteurs politiques, et à proposer des cadres conceptuels adaptés à la mise en œuvre de régulations et de politiques de lutte en faveur d'un environnement informationnel sain. L'enjeu définitionnel est inhérent aux ambitions de régulation : pour limiter les risques posés par ces formes d'informations, il est nécessaire d'en définir les contours, de distinguer ce qui est acceptable de ce qui ne l'est pas. L'identification claire d'un phénomène et de ses ressorts conditionne en effet le développement de solutions opérationnelles, qu'il s'agisse de régulations ou de contre-mesures intégrées aux principaux espaces

1 James H. Kuklinski et al., "Misinformation and the Currency of Democratic Citizenship," *The Journal of Politics* 62, no. 3 (2000): 790–816.

2 Forum Information Democracy, "The International Partnership for Information and Democracy," Forum Information Democracy, September 2019.

de diffusion. Comme le soulignent Wardle et Darakhshan (2017) dans leur rapport pour le Conseil de l'Europe :

> *"Les mots que nous choisissons pour décrire la manipulation des médias peuvent conduire à des hypothèses sur la façon dont l'information se propage, sur les personnes qui la propagent et sur celles qui la reçoivent. Ces hypothèses peuvent influencer les types d'interventions ou de solutions qui semblent souhaitables, appropriées ou même possibles."* [3]

Par cet extrait, les chercheurs soulignent la capacité des mots à *comprendre*, c'est à dire à englober, un phénomène : ses formes, ses dynamiques et ses faiblesses.

La question définitionnelle est d'autant plus centrale qu'elle conditionne, par extension, celle de la mesure du phénomène étudié. En effet, des conceptualisations différentes d'un même phénomène entraînent immanquablement des désaccords sur les effets ou la prévalence de ce dernier (Rogers,2020). D'un texte à l'autre, les contenus désignés par le terme "mésinformation" ou "désinformation" varient sensiblement. Le Code de Conduite sur la Désinformation [4] désigne ainsi par le terme "mésinformation" les contenus "*faux ou trompeurs partagés sans intention de nuire, bien que leurs effets puissent néanmoins être préjudiciables*". Dans de nombreuses publications scientifiques, le terme "mésinformation" est toutefois employé pour désigner de manière générique toute forme de contenus faux ou inexacts, sans présupposé d'intention [5]. À l'inverse, le Code de Conduite sur la Désinformation assigne cette définition plus englobante au terme "Désinformation". D'autres expressions telles que "fake news" [6], "informations problématiques" [7] ou "désordres de l'information" [8] ont également été proposées pour désigner de manière globale toute forme d'information susceptible d'induire en erreur ou d'entrer en contradiction avec l'état des connaissances scientifiques socialement

3 Claire Wardle and Hossein Derakhshan, *Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making*, no. 27 (Council of Europe, 2017), 1–107.

4 European Commission, *Code of Conduct on Disinformation [as Amended in October 2024]* (Brussels: European Commission, 2024), 56.

5 Manon Berriche and Sacha Altay, "Internet Users Engage More with Phatic Posts than with Health Misinformation on Facebook," *Palgrave Communications* 6, no. 1 (2020): 1–9.

6 Fabio Giglietto et al., "'Fake News' Is the Invention of a Liar: How False Information Circulates within the Hybrid News System," *Current Sociology* 67, no. 4 (2019): 625–42.

7 Caroline Jack, *Lexicon of Lies. Terms for Problematic Information* (New York City: Data & Society Research Institute, 2017), 1–20; Maria D. Molina et al., "'Fake News' Is Not Simply False Information: A Concept Explication and Taxonomy of Online Content," *American Behavioral Scientist* 65, no. 2 (2021): 180–212.

8 Wardle and Derakhshan, *Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making*.

établies, constituant à ce titre une menace pour l'information des citoyens.

Contrairement à ce que l'usage courant des expressions « mésinformation » et « désinformation » dans la littérature scientifique, les textes juridiques, les codes de pratique ou les discours médiatiques pourrait laisser entendre, il n'existe pas à date de consensus stabilisé autour de la manière de définir ces termes. Tout travail de recherche ou toute action visant à traiter ces phénomènes suppose donc, en amont, de préciser explicitement le terme mobilisé et le périmètre conceptuel qu'il recouvre.

Compte tenu de la part grandissante qu'occupent les réseaux sociaux numériques dans la consommation d'information des citoyens à travers le monde, ce rapport se concentre sur les enjeux propres à l'environnement informationnel numérique, et plus spécifiquement sur les dynamiques à l'œuvre au sein des principales plateformes numériques qualifiées de Très Grandes Plateformes en Ligne (Very large Online Platforms - VLOPs) par le Digital Services Act. En s'appuyant sur les recherches portant sur les définitions effectivement mobilisées par les acteurs de la lutte contre la mésinformation et la désinformation en ligne sur les principaux réseaux sociaux [9], ce rapport désigne par le terme *mésinformation* les contenus qui, sans postulat d'intention, sont susceptibles d'induire en erreur par omission, déformation ou invention d'informations relatives à des sujets d'intérêt public. Ce choix s'inscrit dans une vaste littérature scientifique adoptant une approche similaire, tout en proposant d'en resserrer le périmètre autour des sujets d'intérêt public et, par conséquent, du risque que représentent les contenus trompeurs pour l'espace informationnel et les démocraties.

9 Marion Seigneurin et al., "Navigating Misinformation and Disinformation: How Definition Ambiguity Limits the DSA's Implementation," *European Journal of Communication* 40, no. 6 (2025): 619–46.

Formes de mésinformation et techniques de manipulation

Les formes de la mésinformation

La mésinformation, en tant que contenu circulant sur les plateformes numériques, se manifeste sous des formes variées, allant de l'erreur ponctuelle formulée en quelques mots à des productions audiovisuelles élaborées, comme les *deepfakes* (vidéos réalistes générées à l'aide de techniques d'intelligence artificielle). Afin de rendre compte de cette diversité formelle, la catégorisation proposée par Wardle et Derakhshan (2017) dans leur rapport pour le Conseil de l'Europe constitue aujourd'hui une référence largement mobilisée.

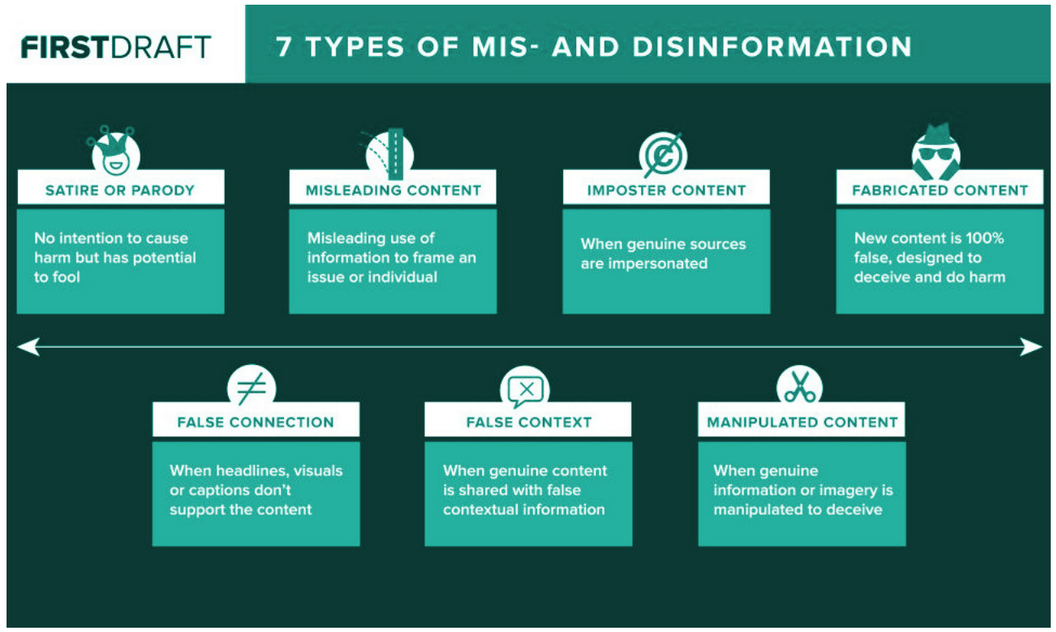


Figure 2 : Claire Wardle, Hossein Derakhshan. 2017. Information Disorder/Toward an interdisciplinary framework for research and policy making. Council of Europe. [initialement produit par Claire Wardle et publié par First Draft]

Cette typologie distingue sept formes de désordres de l'information[10], organisées selon un spectre d'impact négatif progressif, allant de la satire au contenu entièrement fabriqué : la satire ou la parodie, la fausse connexion, le

10 Wardle et Derakhshan (2017) propose l'expression "désordres de l'information" pour désigner les fausses informations diffusées involontairement (mésinformation), les fausses informations diffusées volontairement dans l'intention de nuire (désinformation) et les informations vraies diffusées dans l'intention de nuire (malinformation)

contenu trompeur, le faux contexte, le contenu imposteur, le contenu manipulé et le contenu fabriqué. Elle vise à décrire les différentes modalités de manipulation de l'information en fonction de leur gravité supposée et de leur potentiel de nuisance.

Toutefois, bien que largement utilisée dans la littérature scientifique et les rapports institutionnels, cette catégorisation présente plusieurs limites analytiques. Elle combine de manière hétérogène différents critères tels que l'intention de l'émetteur, le degré de fausseté du contenu et la stratégie formelle mobilisée, sans les distinguer explicitement. Certaines catégories reposent ainsi sur l'identification d'une intention (*tromper, déformer, nuire*), tandis que d'autres n'en précisent pas les conditions. Or, dans les environnements numériques contemporains, l'intention de l'émetteur est rarement ou difficilement accessible [11], ce qui limite la portée opérationnelle de cette typologie.

Par ailleurs, bien que le spectre proposé soit présenté comme étant fondé sur l'impact négatif de la mésinformation, les définitions associées à chaque catégorie demeurent largement dépendantes de la véracité supposée du contenu et de l'intention de son auteur. Cette approche tend à suggérer que les formes de mésinformation ayant l'impact négatif le plus fort seraient nécessairement fausses, techniquement sophistiquées et conçues dans un but explicite de tromper ou de nuire, à l'image des deepfakes générés par l'IA. Or, une partie de la littérature suggère que l'impact de la mésinformation ne peut être expliqué uniquement par la vraisemblance des contenus ou par leur degré de sophistication technique. Plusieurs études soulignent que des formes de manipulation peu élaborées peuvent fonctionner comme des preuves aux yeux des récepteurs et produire des effets importants, indépendamment des techniques mobilisées pour leur fabrication [12]. Ces limites invitent à déplacer l'analyse de la seule production du contenu vers les modalités de réception de la manipulation informationnelle.

Dans cette perspective, Paris et Donovan (2019) distinguent deux grandes modalités de manipulation de l'information : l'expression et la preuve [13]. L'expression renvoie à des formes de manipulation identifiables comme telles par les récepteurs, permettant l'expression d'une opinion, d'une critique politique

11 Joan Donovan and Brian Friedberg, *Source Hacking. Media Manipulation in Practice*, Media Manipulation Research Initiative (Data & Society Research Institute, 2019), 56; Stephan Lewandowsky et al., "Liars Know They Are Lying: Differentiating Disinformation from Disagreement," *Humanities and Social Sciences Communications* 11, no. 1 (2024): 986.

12 Jack, *Lexicon of Lies. Terms for Problematic Information*; Britt Paris and Joan Donovan, *Data & Society — Deepfakes and Cheap Fakes* (New York City: Data & Society Research Institute, 2019), 1–50.

13 Paris and Donovan, *Data & Society — Deepfakes and Cheap Fakes*.

ou d'une prise de position culturelle. Ces formes, telles que la satire, la parodie ou certains détournements visuels, sont généralement reconnues comme manipulées et s'apparentent davantage à des modalités de liberté d'expression ou de participation politique. Henry Jenkins rapproche ainsi ces pratiques de formes de démocratie participative, dans lesquelles les citoyens utilisent la manipulation de contenus comme un moyen de demander des comptes aux institutions ou aux figures de pouvoir [14]. Prolongeant ces courant réflexifs, Andersen et Søe (2020) conceptualisent la mésinformation comme une forme d'action communicative inhérente à l'histoire de la communication politique, qui devient ainsi davantage un élément du débat démocratique auquel les citoyens doivent apprendre à faire face, plutôt qu'un contenu qu'il serait possible de supprimer [15].

À l'inverse, la manipulation de l'information peut être reçue comme une preuve, c'est-à-dire comme un élément factuel venant soutenir une représentation tronquée de la réalité. Dans ce cas, le caractère manipulé du contenu est imperceptible pour les récepteurs, qui l'intègrent comme une information fiable et contribuent ainsi à la diffusion et à la consolidation de fausses croyances. La mésinformation fonctionnant comme preuve limite la possibilité d'une compréhension commune de la réalité et constitue, à ce titre, un risque pour l'espace informationnel et démocratique.

Le danger apparaît de manière accrue lorsque la frontière entre expression et preuve devient floue. Un contenu initialement conçu comme une expression peut être reçu comme une preuve dès lors que ses codes ne sont plus identifiables ou que son contexte de circulation est perdu ce qui, sur les réseaux sociaux, s'apparente davantage à une norme qu'à une exception. Dans ces situations, la réception de l'expression devient équivalente à celle d'une preuve, produisant des effets négatifs comparables.

En nous appuyant sur cette distinction, nous proposons de relire le spectre d'impact négatif développé par Wardle et Derakhshan (2017) non pas à partir de l'intention supposée de l'émetteur, mais en fonction du degré de perceptibilité de la manipulation et du mode de réception du contenu. À une extrémité du spectre, la parodie se caractérise par son inscription dans un registre expressif identifiable, limitant son potentiel de nuisance. À l'autre extrémité, le contenu fabriqué

14 Henry Jenkins, "Photoshop for Democracy," Research, *MIT Technology Review*, April 6, 2004.

15 Jack Andersen and Sille Obelitz Søe, "Communicative Actions We Live by: The Problem with Fact-Checking, Tagging or Flagging Fake News – the Case of Facebook," *European Journal of Communication* 35, no. 2 (2020): 126–39.

correspond à une manipulation imperceptible, susceptible de fonctionner comme une fausse preuve. L'enjeu analytique n'est dès lors plus de déterminer si l'objectif de l'émetteur était de nuire, mais d'identifier si le contenu est susceptible d'être interprété comme une représentation fiable de la réalité.

Cette relecture permet de mieux articuler les formes de la mésinformation avec leurs effets potentiels, et constitue un cadre pertinent pour analyser, dans les sections suivantes, les supports, les techniques de manipulation et les dynamiques de diffusion et d'amplification de la mésinformation en ligne.

Page suivante :

1. Parodie : Le Gorafi. 29/01/2024. Choquée par l'absentéisme des députés, Amélie Oudéa-Castéra ne mettra plus les pieds à l'Assemblée nationale. https://www.legorafi.fr/2024/01/29/choquee-par-labsenteisme-des-deputes-amelie-oudea-castera-nemettra-plus-les-pieds-a-lassemblee-nationale/

2. Fausse connexion : Yaume. 05/11/2022. Non, la Commission Européenne ne peut pas imposer un pass vaccinal à la France. Hoaxbuster. https://www.hoaxbuster.com/covid19/2022/11/07/non-la-commission-europeenne-ne-peut-pasimposer-un-pass-vaccinal-a-la-france

3. Camouflage de sources : Joan Donovan, Brian Friedberg. 04/09/2019. Source Hacking, Media Manipulation in Practice. Data & Society

4. Contenu trompeur : nombre-avortement.fr (@nombreivg). 12/02/2023. Twitter https://x.com/nombreivg/status/1624683760187764737?s=20 Fondation des femmes. 17/01/2024. Mobilisation anti-avortement en France. Quand les réseaux sociaux menacent le droit à l'IVG "les discours anti-avortement s'appuient essentiellement sur des chiffres non officiels ou décontextualisés sur le nombre d'avortements pratiqués chaque année en France et dans le monde pour insinuer que les avortements sont en hausse ou que cette procédure est prétendument banalisée."

5. Faux contexte : Monique NGO MAYAG, AFP Sénégal. 07/02/2024. Attention, cette vidéo ne montre pas les protestations en cours au Sénégal. AFP Factuel. https://factuel.afp.com/doc.afp.com.34HQ3PU

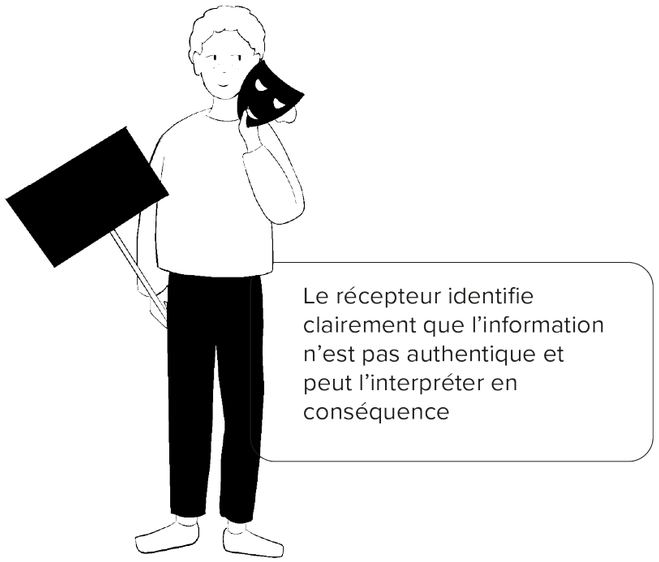
6. Contenu imposteur : JR. 05/12/2021. Le faux profil Twitter d'un politique sud africain fait circuler une fake news internationale. Hoaxbuster.com https://www.hoaxbuster.com/covid19/2021/12/07/faux-profil-twitter-premier-ministre-AF

7. Contenu manipulé : Riddhish Dutta. 22/01/2023. Fact Check: Joe Biden did not touch his granddaughter in an inappropriate manner, edited video goes viral. India Today. https://www.indiatoday.in/fact-check/story/fact-check-joe-biden-did-not-touch-granddaughterinappropriate-manner-edited-video-goes-viral-2324894-2023-01-22 Facebook, Conseil de surveillance. 02/2024. Oversight Board Upholds Meta's Decision in Altered Video of President Biden Case https://oversightboard.com/news/1068824731034762-oversight-board-upholds-meta-s-decision-inaltered-video-of-president-biden-case/

8. Contenu fabriqué : Catalina Marchant de Abreu. 30/01/2024. Taylor Swift deepfake porn inundates X, highlighting dangers of AI-generated content. France 24. https://www.france24.com/en/tv-shows/truth-or-fake/20240130-taylor-swift-deepfake-porn-inundatesx-highlighting-the-dangers-of-ai-generated-content

MANIPULATION PEU ÉLABORÉE — Manipulation détectable

MANIPULATION TRÈS ÉLABORÉE — Manipulation imperceptible



Le récepteur identifie clairement que l'information n'est pas authentique et peut l'interpréter en conséquence

Le récepteur ne peut pas distinguer spontanément le caractère manipulé de l'information et risque de l'interpréter comme vraie

Satire Parodie

Contenu humoristique potentiellement trompeur

1

Fausse connexion

L'en-tête, le titre ou le visuel ne correspondent pas au contenu

2

Cammouflage de sources

La véritable source d'information est dissimulée à un éventuel grand public. On distingue quatre sous-catégories : slogans viraux, reconstitution de fuite de données, assemblage de preuves et accaparement de mots clés.

3

Contenu trompeur

L'information est détournée ou décontextualisée

4

Faux contexte

Contenu repris et partagé avec de fausses informations contextuelles

5

Contenu imposteur

Faux contenu imitant les caractéristiques de contenus fiables

6

Contenu manipulé

Contenu dont un aspect a été volontairement modifié

7

Contenu fabriqué

Contenu fabriqué de toutes pièces

8

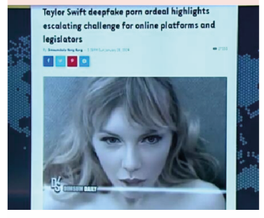


Figure 3 : Catégories des formes de désinformations selon le degré de perceptibilité de la manipulation de l'information

Les supports de la mésinformation

La mésinformation, indépendamment de sa forme ou de son intention, peut être produite et diffusée à travers différents formats : texte, image, audio, vidéo, ou des combinaisons de ces éléments. Ces formats sont transversaux aux catégories de mésinformation précédemment décrites et ne constituent pas, en eux-mêmes, des indicateurs de l'intention ou du degré de manipulation. Une parodie ou une satire peut ainsi mobiliser l'image, la vidéo ou le son, tandis qu'un contenu imposteur peut prendre la forme d'un article textuel imitant un média d'information ou d'un faux compte combinant texte et éléments visuels sur les réseaux sociaux.

L'importance des formats tient avant tout à leurs effets sur la réception. De nombreuses études montrent que les contenus visuels ou multimodaux, combinant texte et image, sont perçus comme plus crédibles [16], suscitent des réactions émotionnelles plus fortes [17] et sont mieux mémorisés que les contenus strictement textuels [18]. La mésinformation visuelle tend ainsi à se diffuser plus rapidement et à produire des effets plus durables sur les perceptions et les comportements, tout en étant plus difficile à corriger. Par ailleurs, certaines recherches suggèrent que l'exposition à des formats fortement manipulés, tels que les deepfakes, peut engendrer une érosion de la confiance des utilisateurs envers les médias et envers leur propre capacité à distinguer le vrai du faux [19].

16 S. Shyam Sundar, "The MAIN Model: A Heuristic Approach to Understanding Technology Effects on Credibility," in *Digital Media, Youth, and Credibility*, ed. Miriam J. Metzger and Andrew J. Flanagin (2008; Cambridge: The MIT Press, 2008), 73–100; Michael Hameleers et al., "A Picture Paints a Thousand Lies? The Effects and Mechanisms of Multimodal Disinformation and Rebuttals Disseminated via Social Media," *Political Communication* 37, no. 2 (2020): 281–301; Tom Dobber et al., "Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes?," *The International Journal of Press/Politics* 26, no. 1 (2021): 69–91.

17 Aarti Iyer et al., "Understanding the Power of the Picture: The Effect of Image Content on Emotional and Political Responses to Terrorism," *Journal of Applied Social Psychology* 44, no. 7 (2014): 511–21; Thomas E. Powell et al., "A Clearer Picture: The Contribution of Visuals and Text to Framing Effects," *Journal of Communication* (United Kingdom) 65, no. 6 (2015): 997–1017; Matthew N. Hannah, "A Conspiracy of Data: QAnon, Social Media, and Information Visualization," *Social Media + Society* 7, no. 3 (2021): 20563051211036064.

18 Doris A. Graber, "Seeing Is Remembering: How Visuals Contribute to Learning from Television News," *Journal of Communication* 40, no. 3 (1990): 134–56; Gillian Murphy and Emma Flynn, "Deepfake False Memories," *Memory* 30, no. 4 (2022): 480–92.

19 Cristian Vaccari and Andrew Chadwick, "Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News," *Social Media + Society* 6, no. 1 (2020): 2056305120903408; Nicholas Diakopoulos and Deborah Johnson, "Anticipating and Addressing the Ethical Implications of Deepfakes in the Context of Elections," *New Media & Society* 23, no. 7 (2021): 2072–98; Teresa Weikmann et al., "After Deception: How Falling for a Deepfake Affects the Way We See, Hear, and Experience Media," *The International Journal of Press/Politics* 30, no. 1 (2025): 187–210.

Au-delà des formats, les modalités de production de la mésinformation varient en fonction des technologies utilisées. Certaines manipulations très élaborées reposent sur des techniques d'intelligence artificielle avancées et sont communément appelées *deep fakes*. À l'inverse, on qualifie communément de *cheap fakes* les manipulations plus accessibles qui reposent sur des procédés simples, comme le recadrage d'images, le ralentissement ou le montage rudimentaire de vidéos existantes, ou encore l'ajout de légendes trompeuses à l'aide d'outils grand public. Certaines de ces manipulations consistent uniquement à recontextualiser des contenus existants, en ajoutant par exemple une légende différente du contexte initial, et relèvent alors de la catégorie du faux contexte.

Une partie de la littérature scientifique souligne toutefois l'absence de corrélation directe entre le degré de sophistication technique des formes de mésinformation et leur impact effectif, invitant ainsi à relativiser l'attention accordée aux technologies d'intelligence artificielle dans le débat public. L'essor récent de l'IA générative a néanmoins ravivé les inquiétudes relatives à la production automatisée de mésinformation qui est désormais identifiée comme un risque majeur dans plusieurs rapports prospectifs [20]. Les systèmes d'IA générative peuvent en effet être mobilisés pour produire des contenus trompeurs dans l'ensemble des formats : textuels, visuels, audio ou audiovisuels [21]. Il est désormais possible de générer ou de modifier de manière réaliste des visages, des scènes ou des voix n'ayant jamais existé, à l'aide de modèles génératifs grand public. Des technologies similaires existent pour l'audio, notamment via des systèmes de synthèse vocale (text-to-speech) ou de conversion vocale (voice conversion), utilisés dans plusieurs contextes de conflits internationaux. Enfin, les modèles de langage de grande taille (Large Language Models - LLMs) permettent la génération de contenus textuels difficiles à distinguer des productions humaines, et déjà largement mobilisés dans des campagnes coordonnées de diffusion de fausses informations sur les réseaux sociaux.

Enfin, la question des formats se pose également dans le développement de mesures de lutte efficaces contre la mésinformation. Les capacités de détection automatisée demeurent très inégales selon les formats : les contenus textuels sont généralement mieux identifiés que les images, les vidéos ou les fichiers audio [22],

20 World Economic Forum, *Global Risks Report 2025*, no. 20, Global Risks Report (Genève: World Economic Forum, 2025), 104.

21 Noémi Bontridder and Yves Poullet, "The Role of Artificial Intelligence in Disinformation," *Data & Policy* 3 (January 2021): e32.

22 Kalina Bontcheva et al., *Generative AI and Disinformation: Recent Advances, Challenges, and Opportunities*, ed. Kalina Bontcheva (TITAN, AI4Media, AI4Trust, 2024), 38.

bien que certains textes générés par des modèles de langage avancés puissent également s'avérer difficiles à distinguer de productions humaines.

En outre, la littérature ne permet pas d'établir de manière concluante quels formats de correction ou de debunking sont les plus efficaces (vidéo, étiquetage visuel, datavisualisation…), soulignant la nécessité de poursuivre les recherches sur l'articulation entre formats, réception et amplification [23].

23 Viorela Dan et al., "Visual Mis- and Disinformation, Social Media, and Democracy," *Journalism & Mass Communication Quarterly* 98, no. 3 (2021): 641–64.

Les techniques de manipulation

Les catégories de mésinformation décrites précédemment permettent de caractériser les formes que prennent les contenus manipulés et les supports mobilisés pour leur diffusion. Elles renseignent toutefois peu sur les mécanismes par lesquels ces contenus produisent des effets sur les récepteurs. L'analyse des techniques de manipulation vise ainsi à compléter cette approche formelle en s'intéressant aux procédés rhétoriques, cognitifs et émotionnels mobilisés au niveau du contenu.

La mésinformation se caractérise par une faible dépendance aux standards de preuve, indépendamment de l'intention initiale de ses émetteurs. Cette relative absence de contrainte d'exactitude permet aux contenus qui la composent de mobiliser efficacement certaines préférences cognitives et émotionnelles des récepteurs, qu'il s'agisse de productions stratégiques ou de reprises involontaires.

Les techniques de manipulation analysées ci-dessous ne relèvent pas uniquement de la falsification intentionnelle de faits. Elles reposent sur des procédés rhétoriques, cognitifs ou émotionnels qui peuvent être mobilisés délibérément dans des stratégies de désinformation, mais aussi être reproduits, amplifiés ou réinterprétés par des utilisateurs ordinaires s'inscrivant alors involontairement dans des dynamiques de mésinformation.

Technique de manipulation	Définition
Usurpation d'identité	L'usurpation d'identité consiste à utiliser de manière frauduleuse l'identité d'un individu ou d'une organisation afin d'accroître la crédibilité perçue d'un contenu. Elle repose sur l'exploitation de la confiance accordée aux sources et peut prendre la forme de faux comptes, de sites imitant des médias ou des institutions, ou encore de contenus attribuant de prétendus propos à des experts, des scientifiques ou des personnalités publiques. (voir notamment Howard & Reiss, 2021)
Raisonnements fallacieux	Les raisonnements fallacieux consistent à mobiliser une structure argumentative apparemment logique pour défendre des conclusions trompeuses ou infondées. Ils incluent notamment la confusion entre corrélation et causalité , les attaques ad hominem , la déformation de données ou d'arguments scientifiques , ainsi que l'appel à des normes naturelles présentées comme intrinsèquement supérieures, en particulier dans les domaines de la santé ou de l'environnement. (Zimmerman et al., 2005 ; Kata, 2012 ; Cook, 2020 ; Lewandowsky et al., 2021)
Appel aux émotions	L'appel aux émotions vise à susciter des réactions affectives fortes, telles que la peur, la colère ou l'indignation, afin de court-circuiter l'évaluation analytique de l'information. En accentuant l'urgence ou la menace perçue, cette technique favorise l'engagement et le partage au détriment de la vérification. (Brady et al., 2017)
Polarisation	La polarisation consiste à accentuer les clivages existants en opposant de manière simplifiée des positions politiques, idéologiques ou affectives . Elle peut viser des enjeux politiques explicites , ou prendre une forme affective en renforçant l'hostilité envers des groupes sociaux ou des individus, contribuant ainsi à rigidifier les cadres interprétatifs. (Van der Linden, 2023 ; Lewandowsky et al., 2024)

Tableau 1 : Exemples de techniques de manipulation mobilisées par la désinformation d'après une synthèse de la littérature

Technique de manipulation	Définition
Théories du complot	Les théories du complot reposent sur des récits expliquant des événements complexes par l'existence de complots secrets attribués à des acteurs puissants et malveillants. Elles se caractérisent par une suspicion généralisée, une résistance aux preuves contradictoires et une tendance à réinterpréter les faits de manière cohérente avec le narratif complotiste, pouvant structurer durablement la vision du monde des individus. (Van Prooijen and Douglas, 2018)
Provocation (Trolling)	La provocation, parfois qualifiée de trolling, vise à susciter des réactions émotionnelles ou conflictuelles par la diffusion de contenus volontairement choquants, polarisants ou insultants . L'objectif principal n'est pas la persuasion, mais la perturbation des échanges et la manipulation de l'attention collective. (McCosker, 2014 ; Roozenbeek and Van der Linden, 2020)
Exigences irréalistes	Les exigences irréalistes consistent à formuler des critiques sur la base de standards de sécurité, d'efficacité ou de cohérence impossibles à satisfaire. Cette technique est fréquemment mobilisée dans les débats liés à la santé publique, à l'immigration ou à la sécurité, où elle permet de délégitimer des politiques ou des institutions en posant des critères inatteignables. (Cook et al., 2023)
Sélection biaisée d'information (Cherry picking)	La sélection biaisée de l'information, souvent désignée par le terme cherry picking, consiste à ne retenir que certains éléments de données, d'images ou d'extraits audiovisuels afin de soutenir une interprétation particulière, tout en omettant les informations contradictoires ou contextuelles nécessaires à une compréhension complète. (Cook et al. 2022)
Deepfakes and contenus manipulés	Les deepfakes et autres contenus synthétiques désignent la manipulation ou la fabrication numérique de contenus textuels, visuels ou sonores dans le but d'induire en erreur. Ils reposent sur des technologies permettant de générer ou de modifier de manière réaliste des visages, des voix ou des scènes, rendant la distinction entre production humaine et production artificielle de plus en plus difficile. (Paris and Donovan, 2019)

Diffuser et amplifier : architectures sociotechniques

Espace informationnel et prépondérance de la mésinformation

La mésinformation se manifeste dans un vaste espace mondialisé permettant le partage de l'information : l'environnement informationnel. Celui-ci peut être compris comme un ensemble d'écosystèmes d'information partiellement imbriqués, au sein desquels interagissent des acteurs, des technologies et des contenus [24]. Un écosystème informationnel renvoie ainsi à toute configuration dans laquelle des individus produisent, traitent et partagent de l'information à l'aide de dispositifs techniques, et dont les dynamiques conditionnent la visibilité et la diffusion des contenus.

La mésinformation est présentée comme un phénomène massif largement diffusé dans l'écosystème de l'information. Le vocabulaire employé pour désigner la présence de ce phénomène informationnel en ligne - "infodémie" [25], "ère post vérité" [26], "guerre de l'information" [27] - révèle la dangerosité perçue et l'étendue du phénomène. En 2023, selon un sondage Ipsos mené sur un échantillon représentatif de la population dans 16 pays différents, 85% de la population estimait que la mésinformation représentait une menace réelle [28]. En 2025, la mésinformation et la désinformation figuraient pour la deuxième année consécutive en tête des risques globaux à court terme selon le World Economic Forum [29].

24 Alicia Wanless et al., *Assessing National Information Ecosystems* (Washington, DC: Carnegie Endowment for International Peace, 2025), 27.

25 Gunther Eysenbach, "Infodemiology: The Epidemiology of (Mis)Information," *The American Journal of Medicine* 113, no. 9 (2002): 763–65; Organisation Mondiale De La Santé OMS, "Première conférence de l'OMS sur l'infodémiologie," Conférence, Première conférence de l'OMS sur l'infodémiologie, June 29, 2020.

26 Stephan Lewandowsky et al., "Beyond Misinformation: Understanding and Coping with the 'Post-Truth' Era.," *Journal of Applied Research in Memory and Cognition* 6, no. 4 (2017): 353–69; Sergio Sismondo, "Post-Truth?," *Social Studies of Science* 47, no. 1 (2017): 3–6; Jayson Harsin, "Post-Truth and Critical Communication Studies," in *Oxford Research Encyclopedia of Communication*, by Jayson Harsin (Oxford University Press, 2018).

27 A. Wanless and J. Pamment, "How Do You Define a Problem Like Influence?," *Journal of Information Warfare* 18, no. 3 (2019): 1–14.

28 Ipsos, *Elections & Social Media: The Battle against Disinformation and Trust Issues | Ipsos* (UNESCO, 2023), 35.

29 World Economic Forum, *Global Risks Report 2025*.

La question de la prépondérance réelle de la mésinformation a également été saisie par la recherche académique, malgré des difficultés de mesure intrinsèque à l'architecture des réseaux sociaux et aux difficultés de définition entourant la mésinformation.

Proposant d'étudier la diffusion de 126,000 rumeurs identifiées par six sites de fact-checking entre 2006 à 2017, Vosoughi et al. (2018) ont ainsi montré que ces dernières, partagées par près de 3 millions d'utilisateurs, se diffusaient plus vite et plus largement que les vraies informations. Le top 1% des fausses informations touchant entre 1000 et 100,000 personnes contre rarement plus de 1000 pour les vraies informations [30] . Dans une étude menée par King et al. auprès de 1010 américains utilisant régulièrement les RSN, 93,3% déclaraient avoir déjà vu, publié ou partagé de la mésinformation sur les réseaux sociaux [31].

Néanmoins, une part grandissante des recherches invite à rationaliser la prévalence de la mésinformation en ligne dans la part globale des contenus consommés quotidiennement, ainsi que son impact direct sur les croyances ou actions des utilisateurs. Budak et al. (2024) ont ainsi mis en évidence un décalage entre les discours publics, qui supposent une forte prévalence et un impact significatif de la mésinformation, et les résultats de la recherche académique, qui révèlent une exposition en réalité limitée, souvent restreinte à une fraction marginale de la population [32]. Dès 2016, les médias comme le New York Times ou Buzzfeed ont respectivement mis en avant des chiffres jugés affolants sur la diffusion de la mésinformation, faisant respectivement état sur Facebook de 3,000 publicités issus de faux comptes russes pour un total de 100,000$ et de 8,711,000 partages, réactions et commentaires générés par les 20 fake news les plus populaires du réseau social entre le 1er août et le jour des élections présidentielles américaines de 2016. Néanmoins, ramené à l'échelle des réseaux sociaux, ces chiffres colossaux ne représentent qu'une fraction des revenus et contenus consommés des plateformes : soit 0.1% des revenus quotidien de Facebook pour les publicités russes et seulement 0.006% de l'engagement utilisateur global [33].

30 Soroush Vosoughi et al., "The Spread of True and False News Online," *Science* 359, no. 6380 (2018): 1146–51.

31 Catherine King et al., "A Path Forward on Online Misinformation Mitigation Based on Current User Behavior," *Scientific Reports* 15, no. 1 (2025): 9475.

32 Ceren Budak et al., "Misunderstanding the Harms of Online Misinformation," *Nature* 630, no. 8015 (2024): 45–53.

33 Duncan J. Watts and David M. Rothschild, "Don't Blame the Election on Fake News. Blame It on the Media.," Columbia Journalism Review, May 5, 2025.

Modalités de diffusion de la mésinformation

Les principaux réseaux sociaux numériques ne constituent pas un écosystème informationnel fermé. Ils s'inscrivent dans un environnement médiatique hybride, au sein duquel circulent en permanence des contenus issus, d'une part, des médias traditionnels dont la presse écrite, la radio et la télévision et, d'autre part, de réseaux d'information alternatifs tels que les blogs, les forums, ou les groupes de discussion ouverts, semi-ouverts ou fermés hébergés par des applications de messagerie et des plateformes comme WhatsApp, Telegram ou Reddit. Ces différents canaux participent conjointement à la captation, à la mise en circulation et à la reformulation des événements qui structurent l'espace public démocratique.

L'articulation entre ces espaces informationnels a été profondément transformée par l'essor des réseaux sociaux numériques. Là où les médias traditionnels reposent historiquement sur des processus de vérification et de hiérarchisation introduisant une latence dans la diffusion de l'information, les plateformes numériques favorisent l'instantanéité et la participation directe des utilisateurs. Cette transformation a fait émerger une première modalité de diffusion de la mésinformation reposant sur des dynamiques *bottom-up*[34], dans lesquelles les citoyens eux-mêmes jouent un rôle central. La mésinformation circule alors à travers des pratiques ordinaires de partage ou de commentaire, socialement et politiquement motivées [35], ou sans intention de tromper, sous l'effet de mécanismes attentionnels, émotionnels ou cognitifs [36].

Une seconde modalité de diffusion repose sur des dynamiques *top-down*[37],

34 littéralement, "*de la base vers le sommet*"

35 Tom Buchanan and James Kempley, "Individual Differences in Sharing False Political Information on Social Media: Direct and Indirect Effects of Cognitive-Perceptual Schizotypy and Psychopathy," *Personality and Individual Differences* 182 (November 2021): 111071; Jay Van Bavel et al., "Political Psychology in the Digital (Mis)Information Age: A Model of News Belief and Sharing," *Social Issues and Policy Review* 15, no. 1 (2021): 84–113; Sacha Altay et al., "A Survey of Expert Views on Misinformation: Definitions, Determinants, Solutions, and Future of the Field," *Harvard Kennedy School Misinformation Review*, ahead of print, July 27, 2023.

36 Yanqing Sun and Juan Xie, "Who Shares Misinformation on Social Media? A Meta-Analysis of Individual Traits Related to Misinformation Sharing," *Computers in Human Behavior* 158 (September 2024): 108271.

37 par opposition à *bottom-up*, désigne un mouvement "*du sommet vers la base*"

portées par des acteurs disposant d'une forte visibilité médiatique et d'un capital d'autorité symbolique, tels que des personnalités politiques, des figures publiques ou certains médias [38]. Les prises de parole des personnalités publiques, qu'elles soient diffusées à la télévision, à la radio ou directement sur les réseaux sociaux, peuvent ainsi constituer des vecteurs puissants de mésinformation [39]. Les personnalités publiques situées aux extrémités de l'échiquier politique ont notamment investi les plateformes numériques comme des outils privilégiés de communication directe, contournant ainsi en partie les médiations journalistiques traditionnelles. En France comme aux États-Unis, les stratégies de certaines formations politiques, en particulier à droite, visant à imposer des thématiques marginales à l'agenda médiatique et politique via les réseaux sociaux sont désormais bien documentées [40].

La diffusion top-down de la désinformation peut également être alimentée de manière non intentionnelle par des médias traditionnels désireux de démentir des contenus de mésinformation ou les relayant par erreur [41]. Toutefois, ces médias se distinguent par l'existence de cadres déontologiques et de procédures de correction transparentes, garantissant en principe une rectification rapide des erreurs factuelles et une transparence quant aux processus de vérification [42]. Si ces mécanismes n'éliminent pas tout risque de mésinformation, ils en limitent généralement la persistance et l'amplification.

Enfin, une troisième modalité de diffusion se distingue des dynamiques *top-down* et *bottom-up* par son caractère organisé et stratégique. Contrairement aux circulations émergentes ou aux prises de parole individuelles relayées à postériori,

38 Nicolas Berlinski et al., "The Effects of Unsubstantiated Claims of Voter Fraud on Confidence in Elections," *Journal of Experimental Political Science* 10, no. 1 (2023): 34–49; Shane Littrell et al., "Who Knowingly Shares False Political Information Online?," *Harvard Kennedy School Misinformation Review*, ahead of print, August 25, 2023.

39 Data for Good et al., *Cartographie de La Désinformation Climatique Dans Les Médias Français et Brésiliens* (2025), Rapport. 120 p.

40 Yochai Benkler et al., *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics* (Oxford University Press, 2018); Tristan Boursier, "Influenceurs d'extrême droite : le moteur caché du succès du RN," The Conversation, The Conversation, June 20, 2024; Kamilia Khalilizadeh, *Retour sur la campagne vue des réseaux sociaux | Ipsos* (Ipsos, 2024); Gaël Stephan and Stéphanie Wojcik, "Engagement et Ethos de l'extrême Droite En Ligne : Militantes et Militants de Reconquête! Sur Instagram:," *Quaderni* n° 111, no. 1 (2024): 83–102.

41 Cherilyn Ireton, Julie Posetti, *Journalism, fake news & disinformation: handbook for journalism education and training* (2018).

42 Emma Margolin, *10 Tips for Reporting on Disinformation*, Tip sheet (Data & Society Research Institute, 2020), 4; Regina Cazzamatta, "Decoding Correction Strategies: How Fact-Checkers Uncover Falsehoods Across Countries," *Journalism Studies*, March 3, 2025, 1–23.

ces dispositifs coordonnés visent à orienter ou perturber délibérément la circulation de l'information. Ces actions reposent sur une coordination stratégique des prises de parole, des temporalités de diffusion et des canaux mobilisés, afin d'augmenter la visibilité de certains contenus ou de saturer l'environnement informationnel. Ces campagnes peuvent être portées par des acteurs aux statuts hétérogènes, allant d'institutions étatiques [43]– on parlera alors communément d'*ingérences étrangères*–à des entrepreneurs de désinformation opérant à des fins économiques [44].

En France, l'importance de ces phénomènes a été soulignée par les autorités en charge de la protection de l'espace informationnel, notamment à travers les travaux du service de vigilance et de protection contre les ingérences numériques étrangères–Viginum–ayant identifié pas moins de vingt-cinq tentatives d'ingérence lors des élections européennes s'étant déroulées en France en 2025 [45]. Plus largement, les incitations économiques à la production et à la diffusion de mésinformation ont favorisé l'émergence d'écosystèmes transnationaux impliquant des acteurs variés : réseaux de faux sites d'information, influenceurs exploitant des récits climatosceptiques, masculinistes ou complotistes, ou encore faux experts dans le domaine de la santé. Dans certains contextes, les motivations politiques et les incitations financières se combinent, à l'image de la campagne d'influence s'étant déroulée lors de l'élection présidentielle roumaine de 2025 et reposant sur des réseaux d'influenceurs [46].

43 Samantha Bradshaw and Philip N. Howard, "The Global Organization of Social Media Disinformation Campaigns," *Journal of International Affairs* 71, no. 1.5 (2018): 23–32; Martin Echeverria et al., *Introduction. Deceiving from the Top: State-Sponsored Disinformation as a Contemporary Global Phenomenon* (2025), 1–17.

44 Donovan and Friedberg, *Source Hacking. Media Manipulation in Practice*; Rafael Grohmann and Jonathan Corpus Ong, "Disinformation-for-Hire as Everyday Digital Labor: Introduction to the Special Issue," *Social Media + Society* 10, no. 1 (2024): 20563051231224723.

45 Vie Publique, "Ingérences étrangères et IA : une menace pour les démocraties ? | vie-publique.fr," Information, vie-publique.fr, December 29, 2025.

46 Viginum, *Manipulation d'algorithmes et instrumentalisation d'influenceurs : enseignements de l'élection présidentielle en Roumanie & risques pour la France*, Enjeux systémiques (SGDSN, 2025), 14.

De la diffusion à l'amplification

L'impact de la mésinformation, loin de dépendre uniquement des contenus eux-mêmes, repose largement sur les mécanismes d'amplification qui conditionnent sa visibilité, sa circulation et sa persistance dans l'espace informationnel. Ces mécanismes opèrent à différents niveaux et peuvent se combiner, produisant dès lors des effets cumulatifs à l'origine de la sensation de prévalence de la mésinformation dans l'espace informationnel numérique.

Un des premiers mécanismes d'amplification de la mésinformation est médiatique. Des contenus initialement marginaux, produits ou diffusés sur des plateformes à faible audience, peuvent acquérir une visibilité massive lorsqu'ils sont repris, commentés ou dénoncés par des médias généralistes. Ce mécanisme est connu sous le nom d'effet Streisand : les tentatives de suppression, de censure ou de démenti public peuvent paradoxalement accroître l'attention portée à un contenu. La médiatisation de ces contenus, même sous une forme critique ou factuelle, peut contribuer à leur légitimation en les inscrivant à l'agenda médiatique et politique, et à leur diffusion auprès de publics qui n'y auraient autrement pas été exposés. Ce phénomène a été observé à de nombreuses reprises lors de la couverture médiatique de rumeurs, de contenus complotistes ou de fausses informations circulant initialement sur des forums ou des réseaux sociaux alternatifs.

Néanmoins, pleinement conscients de ces enjeux, de nombreux organismes de recherche et de fact-checking ont formulé des recommandations visant précisément à limiter les possibles effets contre-productifs de la couverture médiatique de la mésinformation. Ces lignes directrices insistent notamment sur l'importance de ne pas reproduire inutilement les récits trompeurs, d'éviter les titres sensationnalistes, de contextualiser les contenus problématiques sans en amplifier les éléments narratifs centraux, et de privilégier une approche axée sur les faits plutôt que sur la polémique. Des structures telles que l'European Digital Media Observatory (EDMO) ou Data & Society ont ainsi contribué à formaliser des bonnes pratiques destinées aux journalistes, aux chercheurs et aux plateformes, visant à réduire les risques d'amplification involontaire tout en maintenant une fonction d'information et de correction [47]. L'effet Streisand apparaît dès lors moins

47 Whitney Phillips, *The Oxygen of Amplification. Better Practices for Reporting on Extremists, Antagonists, and Manipulators.* (Data & Society Research Institute, 2018), 45; Margolin, *10 Tips for Reporting on Disinformation*; Janine Zacharia and Andrew Grotto, *How to Responsibly Report on Hacks and Disinformation* (California: Stanford Cyber Policy Center, 2020), 16; American Psychological Association, "Recommendations

comme une fatalité que comme un risque conditionné par les modalités concrètes de traitement médiatique des contenus trompeurs.

L'amplification algorithmique est un autre mécanisme d'amplification, qui repose sur les algorithmes de recommandations utilisés par les plateformes numériques pour conditionner la visibilité de certains contenus et maximiser l'engagement des utilisateurs à des fins économiques [48]. Ces systèmes de recommandation sont conçus pour capter l'attention des utilisateurs et augmenter les métriques d'engagement comme les clics, les partages, les commentaires, ou le temps de visionnage, car ces indicateurs conditionnent directement les revenus publicitaires.

Or, de nombreux travaux montrent que les contenus suscitant de fortes réactions émotionnelles, telles que la peur [49], la colère ou l'aversion [50], performent mieux selon ces métriques. La mésinformation, en mobilisant fréquemment ces registres émotionnels, bénéficie ainsi d'un avantage structurel de diffusion, indépendamment de sa véracité ou de son intentionnalité. Cette logique produit des dynamiques informationnelles délétères [51]. Elle engendre notamment les bulles de filtres, qui enferment les utilisateurs dans un ensemble homogène de contenus, parmi lesquels des thématiques marginales, interprétations biaisées ou récits trompeurs apparaissent comme cohérents et socialement validés. Selon une dynamique proche, les *rabbit-holes*, littéralement trous de lapin en anglais, qualifient des recommandations dont le caractère polarisant et extrême s'intensifie à mesure que les utilisateurs consomment les contenus suggérés. Ces logiques algorithmiques interagissent avec les biais cognitifs préexistants des utilisateurs et jouent un rôle d'amplification déterminant, sans toutefois être seules responsables

for Countering Misinformation," Apa.Org, American Psychological Association, November 29, 2023.

48 Miriam Fernández et al., "Analysing the Effect of Recommendation Algorithms on the Amplification of Misinformation," arXiv:2103.14748, preprint, arXiv, March 26, 2021; Royal Pathak et al., "Understanding the Contribution of Recommendation Algorithms on Misinformation Recommendation and Misinformation Dissemination on Social Networks," *ACM Trans. Web* 17, no. 4 (2023): 35:1-35:26; Pau Muñoz et al., "The Role of Recommendation Algorithms in the Formation of Disinformation Networks," *Information Processing & Management* 62, no. 6 (2025): 104243.

49 Carola Salvi et al., "Going Viral: How Fear, Socio-Cognitive Polarization and Problem-Solving Influence Fake News Detection and Proliferation During COVID-19 Pandemic," *Frontiers in Communication* 5 (January 2021).

50 Michael MacKuen et al., "Civic Engagements: Resolute Partisanship or Reflective Deliberation," *American Journal of Political Science* 54, no. 2 (2010): 440–58; Brian E. Weeks, "Emotions, Partisanship, and Misperceptions: How Anger and Anxiety Moderate the Effect of Partisan Bias on Susceptibility to Political Misinformation," *Journal of Communication* 65, no. 4 (2015): 699–719.

51 David Chavalarias, *Toxic data: comment les réseaux manipulent nos opinions* (Paris: Flammarion, 2025).

des phénomènes de bulle de filtre [52].

L'amplification algorithmique de la mésinformation ne relève donc pas d'un dysfonctionnement ponctuel, mais d'un effet systémique du modèle de captation de l'attention mis en place par les plateformes.

Un troisième mécanisme d'amplification de la mésinformation repose sur des dynamiques artificielles et stratégiques, mises en œuvre par des réseaux coordonnés étatiques ou privés. Ces dispositifs exploitent les signaux sociaux pris en compte par les algorithmes (volume d'interactions, rapidité de diffusion, popularité apparente) afin d'accroître artificiellement la présence en ligne de certains contenus et de simuler un consensus ou une opinion publique factice. Ils s'appuient la plupart du temps sur l'utilisation de faux comptes et de réseaux de bots, dont la présence en ligne soulève des enjeux de modération relatifs à l'équilibre entre la réduction de la diffusion de contenus inauthentiques et la préservation de la liberté d'expression [53].

Dans ces configurations, l'amplification ne résulte pas de pratiques ordinaires de partage, mais d'une manipulation directe des infrastructures de visibilité des plateformes numériques. Certaines campagnes mobilisent également des stratégies d'exploitation des vides informationnels (*data voids*), consistant à produire massivement des contenus trompeurs associés à des termes, événements ou expressions encore peu documentés en ligne [54]. En saturant ces espaces informationnels naissants, les acteurs malveillants orientent les résultats des moteurs de recherche ou les dynamiques de circulation sur les réseaux sociaux, et imposent des récits faux ou biaisés avant l'émergence de sources fiables. Cette stratégie d'amplification s'avère particulièrement problématique dans des contextes d'urgence comme les crises environnementales et sanitaires ou les conflits armés [55].

Les trois mécanismes précédemment décrits montrent que l'amplification de la mésinformation ne peut être réduite à un simple déficit de jugement individuel.

52 Franziska Zimmer et al., "Fake News in Social Media: Bad Algorithms or Biased Users?," *Journal of Information Science Theory and Practice* 7, no. 2 (2019): 40–53.

53 Lynnette Hui Xian Ng and Kathleen M. Carley, "A Global Comparison of Social Media Bot and Human Characteristics," *Scientific Reports* 15, no. 1 (2025): 10973.

54 Danah boyd and Michael Golebiewski, *Data Voids: Where Missing Data Can Easily Be Exploited* (New York City: Data & Society Research Institute, 2019), 1–51.

55 Sonya Hilberts et al., "The Impact of Misinformation on Social Media in the Context of Natural Disasters: Narrative Review," *JMIR Infodemiology* 5, no. 1 (2025): e70413.

Elle résulte de dynamiques médiatiques (relais cross-médias, effet Streisand), de logiques algorithmiques structurées par la captation de l'attention (bulles de filtre, *rabbit holes*), ainsi que de dispositifs artificiels et stratégiques reposant sur des comportements inauthentiques coordonnés. Ces mécanismes contribuent à accroître la visibilité et la persistance de certains contenus, indépendamment de leur adhésion réelle dans la population.

Cependant, l'amplification de la mésinformation ne repose pas uniquement sur ces infrastructures ou sur des acteurs organisés. Elle s'inscrit aussi dans des dynamiques de propagation sociale, directement liées aux usages ordinaires des plateformes numériques. Certains contenus connaissent ainsi une diffusion disproportionnée du fait de l'activité de *super-sharers*, des utilisateurs particulièrement actifs ou bien connectés, capables de faire circuler rapidement des récits trompeurs au-delà de leurs cercles immédiats [56]. Plus largement, la circulation de la mésinformation est favorisée par des logiques d'homophilie et de polarisation, dans lesquelles les contenus circulent prioritairement au sein de communautés partageant des cadres interprétatifs proches, renforçant leur persistance sans nécessairement atteindre un public majoritaire [57].

Ces dynamiques invitent à déplacer le regard : si les utilisateurs participent activement à la diffusion de la mésinformation, ce n'est pas uniquement parce qu'ils se trompent ou adhèrent naïvement aux contenus partagés. Comprendre les ressorts de cette propagation suppose d'analyser les conditions de réception, les mécanismes cognitifs, sociaux et émotionnels mobilisés, ainsi que les fonctions que ces contenus peuvent remplir pour ceux qui les relaient. C'est à l'examen de ces déterminants que la section suivante est consacrée.

56 Sahar Baribi-Bartov et al., "Supersharers of Fake News on Twitter," *Science* 384, no. 6699 (2024): 979–82; Sander Van der Linden and Yara Kyrychenko, "A Broader View of Misinformation Reveals Potential for Intervention," *Science* 384, no. 6699 (2024): 959–60.

57 Zimmer et al., "Fake News in Social Media."

Réception et vulnérabilités

4.1 Facteurs informationnels

Les facteurs informationnels renvoient aux conditions dans lesquelles les individus accèdent à l'information, la comprennent et l'interprètent dans leur environnement informationnel. Une première hypothèse, historiquement dominante dans les sciences de l'information et de la communication, repose sur le modèle du déficit informationnel [58]. Selon ce paradigme, la mésinformation se diffuser principalement parce que les individus sont insuffisamment ou mal informés, ou encore privés d'accès à des sources fiables. Les désordres de l'information résulteraient alors d'une mauvaise compréhension des faits ou d'un manque de connaissances, conduisant les utilisateurs à former et à relayer des croyances erronées.

Ce modèle a structuré de nombreuses interventions visant à lutter contre la mésinformation, en particulier celles reposant sur l'apport d'informations correctives ou sur la production de contenus explicatifs plus accessibles. Des travaux empiriques ont montré que la mise à disposition d'informations contextualisées, présentées de manière claire et visuelle, pouvait réduire certaines fausses croyances, notamment dans des environnements peu politisés [59]. Toutefois, la littérature récente a largement remis en question le caractère suffisant de cette approche. Plusieurs études montrent que l'accès à l'information fiable, bien qu'indispensable, ne permet ni d'expliquer à lui seul la croyance dans la mésinformation, ni d'anticiper les comportements de partage [60]. Des individus fortement exposés à des sources d'information de qualité, et disposant de compétences informationnelles élevées, peuvent aussi bien participer activement à la diffusion de contenus trompeurs.

Ces résultats invitent à déplacer l'analyse du seul manque d'information vers les conditions structurelles de circulation et de réception des contenus. Dans l'environnement numérique, l'information est souvent fragmentée, décontextualisée et consommée sous forme d'extraits, de titres ou de visuels isolés.

58 Molly J. Simis et al., "The Lure of Rationality: Why Does the Deficit Model Persist in Science Communication?," *Public Understanding of Science (Bristol, England)* 25, no. 4 (2016): 400–414; Laura D. Scherer and Gordon Pennycook, "Who Is Susceptible to Online Health Misinformation?," *American Journal of Public Health* 110, no. S3 (2020): S276–77.

59 Brendan Nyhan and Jason Reifler, "The Roles of Information Deficits and Identity Threat in the Prevalence of Misperceptions," *Journal of Elections, Public Opinion and Parties*, April 3, 2019, world.

60 Ullrich K. H. Ecker et al., "The Psychological Drivers of Misinformation Belief and Its Resistance to Correction," *Nature Reviews Psychology* 1, no. 1 (2022): 13–29; Sander Van Der Linden, "Misinformation: Susceptibility, Spread, and Interventions to Immunize the Public," *Nature Medicine* 28, no. 3 (2022): 460–67.

Cette fragmentation affaiblit les repères nécessaires à l'évaluation de la fiabilité d'un contenu et favorise des interprétations erronées, en particulier lorsque les éléments contextuels comme la source, la date, ou les conditions de production, sont absents ou difficiles à identifier. La mésinformation s'insère ainsi dans des espaces d'incertitude informationnelle, où les connaissances disponibles sont incomplètes, instables ou trop complexes pour être immédiatement interprétées.

La littératie médiatique et numérique constitue ainsi un volet central des facteurs informationnels. Définie comme la capacité à accéder, analyser, évaluer et produire des messages dans des contextes variés, elle est régulièrement mobilisée pour expliquer la susceptibilité aux désordres de l'information [61]. De nombreuses études montrent une association entre faibles niveaux de littératie numérique et exposition accrue à la mésinformation [62]. Néanmoins, les résultats restent ambivalents : si certaines compétences améliorent le discernement, elles n'expliquent pas systématiquement les intentions de partage [63].

61 Elena Broda and Jesper Strömbäck, "Misinformation, Disinformation, and Fake News: Lessons from an Interdisciplinary, Systematic Literature Review," *Annals of the International Communication Association* 48, no. 2 (2024): 139–66.

62 Andrew M. Guess et al., "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* 117, no. 27 (2020): 15536–45.

63 Nathaniel Sirlin et al., "Digital Literacy Is Associated with More Discerning Accuracy Judgments but Not Sharing Intentions," *Harvard Kennedy School Misinformation Review*, ahead of print, December 6, 2021.

4.2. Facteurs cognitifs

Les facteurs cognitifs renvoient aux mécanismes de traitement de l'information mobilisés par les individus lors de la réception et de l'évaluation des contenus. Une large partie de la littérature s'appuie sur les théories de la rationalité limitée et sur les modèles du raisonnement dual pour expliquer la susceptibilité aux désordres de l'information [64]. Contrairement au modèle de l'homo economicus [65], ces approches considèrent que les individus disposent de capacités cognitives contraintes et qu'ils recourent à des stratégies de simplification pour traiter l'information dans des environnements complexes et saturés.

Le modèle des processus duels distingue ainsi un système de raisonnement intuitif, rapide et peu coûteux cognitivement (Système 1), et un système analytique, lent et exigeant (Système 2) [66]. Dans les environnements numériques, caractérisés par la surcharge informationnelle, la vitesse de circulation des contenus et la multiplication des stimuli attentionnels, le traitement intuitif tend à dominer. Les utilisateurs évaluent alors les contenus à partir d'indices heuristiques plutôt qu'à partir d'un examen approfondi de leur exactitude [67].

Ces heuristiques – telles que la familiarité, la disponibilité, la cohérence ou l'approbation sociale – constituent des mécanismes cognitifs ordinaires, fonctionnels dans de nombreux contextes. Toutefois, dans l'environnement des réseaux sociaux, elles peuvent conduire à des biais systématiques. L'effet de vérité illusoire illustre ce phénomène : la répétition d'une information augmente sa véracité perçue en facilitant son encodage et sa récupération en mémoire, indépendamment de son exactitude. Ce mécanisme est particulièrement saillant dans des environnements où les mêmes récits circulent sous des formes légèrement différentes, donnant l'impression d'une convergence informationnelle [68].

64 Gordon Pennycook and David G. Rand, "The Psychology of Fake News," *Trends in Cognitive Sciences* 25, no. 5 (2021): 388–402.

65 john Von Neumann, *Theory Of Games And Economic Behavior* (1944); Joseph Persky, "The Ethology of Homo Economicus," *Journal of Economic Perspectives* 9, no. 2 (1995): 221–31.

66 Amos Tversky and Daniel Kahneman, "Judgment under Uncertainty: Heuristics and Biases," *Science* 185, no. 4157 (1974): 1124–31; Jonathan Evans and Keith Stanovich, "Dual-Process Theories of Higher Cognition," *Perspectives on Psychological Science* 8 (May 2013): 223–41.

67 Thomas Gilovich et al., eds., *Heuristics and Biases: The Psychology of Intuitive Judgment* (Cambridge: Cambridge University Press, 2002).

68 Ian Maynard Begg et al., "Dissociation of Processes in Belief: Source Recollection, Statement Familiarity, and the Illusion of Truth," *Journal of Experimental Psychology: General* (US) 121, no. 4 (1992): 446–58; Gordon Pennycook et al., "Prior Exposure Increases Perceived Accuracy of Fake News," *Journal of Experimental Psychology: General* (US) 147, no. 12 (2018): 1865–80; Ecker et al., "The Psychological Drivers of Misinformation Belief and Its Resistance to Correction."

Les facteurs cognitifs permettent également de distinguer croyance et partage. Plusieurs travaux montrent que les individus peuvent partager des contenus qu'ils jugent inexacts ou douteux, notamment lorsque l'attention n'est pas explicitement orientée vers la précision [69]. L'inattention apparaît alors comme un facteur central : dans les usages ordinaires des plateformes, les objectifs poursuivis ne sont pas prioritairement épistémiques. Le partage peut répondre à des logiques de divertissement, de réaction émotionnelle ou d'interaction sociale, reléguant l'évaluation de l'exactitude au second plan.

Enfin, ces mécanismes cognitifs sont renforcés par les architectures socio-techniques des plateformes. Les métriques d'engagement, la visibilité cumulative et les recommandations algorithmiques produisent des signaux interprétés par les utilisateurs comme des indices de crédibilité ou d'importance [70]. Les vulnérabilités cognitives doivent ainsi être comprises comme le produit de l'interaction entre des mécanismes mentaux ordinaires et un environnement informationnel conçu pour minimiser la friction cognitive et maximiser la réactivité.

69 Gordon Pennycook and David G. Rand, "Lazy, Not Biased: Susceptibility to Partisan Fake News Is Better Explained by Lack of Reasoning than by Motivated Reasoning," *Cognition*, The Cognitive Science of Political Thought, vol. 188 (July 2019): 39–50.

70 Hugo Mercier, *Not Born Yesterday | Princeton University Press* (2020).

4.3. Facteurs sociaux

Les facteurs sociaux déplacent l'analyse de la relation individu–contenu vers les dynamiques d'interaction et d'appartenance qui structurent la circulation de l'information. La littérature montre que le partage de mésinformation ne suppose pas nécessairement une adhésion à la véracité des contenus. Dans de nombreux cas, la diffusion de mésinformation répond à des fonctions sociales, identitaires ou relationnelles, indépendantes des finalités informatives.

La théorie du raisonnement motivé constitue l'un des cadres explicatifs les plus mobilisés pour comprendre ces dynamiques [71]. Contrairement aux biais cognitifs associés au traitement intuitif, le raisonnement motivé repose sur des processus réflexifs orientés vers la protection de croyances, de valeurs ou d'identités socialement saillantes [72]. Dans ce cadre, les individus mobilisent leurs capacités analytiques non pour évaluer objectivement une information, mais pour défendre une position préexistante, réduire une dissonance cognitive ou préserver leur appartenance à un groupe.

Ces mécanismes sont particulièrement visibles dans les contextes de polarisation politique. De nombreuses études montrent que la concordance idéologique d'un contenu constitue un prédicteur plus fort de son partage que son exactitude perçue [73]. Les réseaux sociaux renforcent ces dynamiques en favorisant l'exposition aux contenus partagés par des connaissances directes ou des utilisateurs idéologiquement proches. Les réactions d'engagement visibles (like, partage, commentaire) sur un contenu donné jouent alors davantage un rôle social, à savoir signaler l'appartenance à certains groupes ou soutenir une personnalité ou des idées, le tout indépendamment de leur validité factuelle.

Ces mêmes dynamiques se manifestent lors du partage de certains types

71 Ziva Kunda, "The Case for Motivated Reasoning," *Psychological Bulletin* (US) 108, no. 3 (1990): 480–98; Dan M. Kahan, "The Politically Motivated Reasoning Paradigm, Part 1: What Politically Motivated Reasoning Is and How to Measure It," in *Emerging Trends in the Social and Behavioral Sciences* (John Wiley & Sons, Ltd, 2016), 1–16.

72 Eva Jonas et al., "Chapter Four - Threat and Defense: From Anxiety to Approach," in *Advances in Experimental Social Psychology*, vol. 49, ed. James M. Olson and Mark P. Zanna (Academic Press, 2014), 219–86.

73 Nir Grinberg et al., "Fake News on Twitter during the 2016 U.S. Presidential Election," *Science (New York, N.Y.)* 363, no. 6425 (2019): 374–78; Gordon Pennycook et al., "Shifting Attention to Accuracy Can Reduce Misinformation Online," *Nature* 592, no. 7855 (2021): 590–95.

de mésinformation à visée humoristique ou expressive endossant des fonctions phatiques. Les contenus circulent alors comme supports d'interaction, permettant de maintenir des liens sociaux [74]. Dans ces configurations, l'information n'est pas évaluée principalement pour sa véracité, mais pour sa capacité à produire du lien, de l'émotion ou de la reconnaissance sociale. Des travaux montrent ainsi que certains contenus largement qualifiés de mésinformation sont partagés pour leur dimension relationnelle ou humoristique, sans intention explicite de tromper.

Enfin, l'explication sociale du partage de l'information, et plus largement les théories autour de l'image de soi sur les réseaux sociaux, permettent de comprendre les coûts associés aux pratiques de correction de la mésinformation en ligne. Corriger une fausse information en ligne peut en effet être perçu comme risqué : peur de fragiliser une relation avec un proche, de s'exposer à une communauté hostile etc. Ces contraintes relationnelles expliquent en partie la rareté des comportements correctifs observés, malgré une approbation largement exprimée à l'égard des pratiques sociales de correction de la mésinformation [75].

74 Berriche and Altay, "Internet Users Engage More with Phatic Posts than with Health Misinformation on Facebook."

75 Catherine King et al., "A Path Forward on Online Misinformation Mitigation Based on Current User Behavior," *Scientific Reports* 15, no. 1 (2025): 9475.

4.4. Facteurs émotionnels

Les facteurs émotionnels occupent une place centrale dans la réception et la diffusion des désordres de l'information. Les recherches montrent que les émotions suscitées par un contenu, tout comme l'état émotionnel préalable du récepteur, influencent significativement le traitement cognitif de l'information [76]. Les individus s'appuient fréquemment sur leur ressenti affectif pour évaluer la pertinence ou l'importance d'un contenu, en particulier dans des contextes de surcharge informationnelle.

La littérature souligne que les contenus générant des émotions intenses tendent à susciter davantage d'engagement. La mésinformation mobilise fréquemment ces registres émotionnels, en s'affranchissant des contraintes de nuance, de proportionnalité ou de vérification propres aux contenus journalistiques. Toutefois, la relation entre émotion et susceptibilité à la mésinformation n'est pas univoque. Certaines émotions négatives, comme l'anxiété ou la tristesse, peuvent accroître la vigilance et la réflexivité [77], tandis que des émotions positives comme la joie peuvent favoriser une crédulité accrue [78].

Plusieurs travaux proposent de dépasser cette opposition en montrant que l'effet des émotions dépend du système cognitif mobilisé. Dans un mode de traitement intuitif, les émotions agissent comme des heuristiques influençant directement les jugements. Dans un mode réflexif, elles peuvent au contraire servir d'indices permettant une évaluation plus critique. Ainsi, un même état émotionnel peut soit faciliter, soit inhiber le partage de mésinformation, selon les ressources attentionnelles disponibles et le contexte d'usage [79].

Les émotions jouent également un rôle dans la persistance des croyances,

76 Ecker et al., "The Psychological Drivers of Misinformation Belief and Its Resistance to Correction."

77 Joseph Forgas, "Don't Worry, Be Sad! On the Cognitive, Motivational, and Interpersonal Benefits of Negative Mood," *Current Directions in Psychological Science* 22 (June 2013): 225–32.

78 Joseph P. Forgas and Rebekah East, "On Being Happy and Gullible: Mood Effects on Skepticism and the Detection of Deception," *Journal of Experimental Social Psychology* (Netherlands) 44, no. 5 (2008): 1362–67; Alex S. Koch and Joseph P. Forgas, "Feeling Good and Feeling Truth: The Interactive Effects of Mood and Processing Fluency on Truth Judgments," *Journal of Experimental Social Psychology* (Netherlands) 48, no. 2 (2012): 481–85.

79 Richard E. Petty and Pablo Briñol, "Emotion and Persuasion: Cognitive and Meta-Cognitive Processes Impact Attitudes," *Cognition and Emotion*, January 2, 2015, world.

en particulier lors de l'association entre une expérience ou connaissance donnée et émotion négative [80]. Les contenus suscitant des émotions négatives, comme la mésinformation, sont donc susceptibles d'être mieux mémorisés et plus facilement réactivés, ce qui contribue à la durabilité de certains récits trompeurs, même après exposition à des corrections. Cette asymétrie affective entre contenus trompeurs et contenus correctifs constitue un défi majeur pour les stratégies de lutte contre la mésinformation [81].

80 Elizabeth A. Kensinger, "Remembering the Details: Effects of Emotion," *Emotion Review: Journal of the International Society for Research on Emotion* 1, no. 2 (2009): 99–113.

81 Ecker et al., "The Psychological Drivers of Misinformation Belief and Its Resistance to Correction."

Réponses et contre-mesures

Les sections précédentes montrent que la mésinformation n'est pas la résultante d'actions de partage réalisées par des individus simplement crédules. Au contraire, elle devient visible et persistante via des mécanismes variés d'amplification technique (médiatiques, algorithmiques, artificiels) et via des dynamiques de réception (informationnelles, cognitives, sociales, émotionnelles) qui orientent ce que les utilisateurs voient, retiennent et partagent. Les contre-mesures déployées en réponse à la mésinformation doivent donc agir à deux niveaux complémentaires : sur l'architecture de circulation (niveau systémique) et sur les comportements et compétences des individus (niveau individuel) [82].

82 Carolin-Theresa Ziemer and Tobias Rothmund, "Psychological Underpinnings of Misinformation Countermeasures," *Journal of Media Psychology* 36, no. 6 (2024): 397–409.

5.1. Contre-mesures systémiques

Les contre-mesures systémiques agissent sur les conditions de visibilité et de circulation de la mésinformation, donc par extension sur l'audience finale des contenus. Elles visent moins à apporter corrections aux contenus consommés par les utilisateurs des réseaux sociaux qu'à réduire la sur-visibilisation produite par l'architecture des plateformes et par leur modèle économique publicitaire.

Les contre-mesures systémiques passent d'abord par l'ajustement des algorithmes de recommandation et de classement : *downranking*[83], limites de recommandation, signaux de qualité, ou réducteurs de viralité [84]. L'objectif n'est pas d'éliminer tous les contenus trompeurs, mais d'éviter qu'ils bénéficient d'un avantage structurel lié à l'engagement et au temps de visionnage. Concrètement, il s'agit de réduire la capacité d'un contenu douteux à passer d'un cercle restreint à une audience massive, et de limiter les enchaînements de recommandations qui alimentent *rabbit holes* et bulles de filtres [85].

Une autre possibilité consiste à agir sur les incitations économiques au cœur du modèle de captation de l'attention : démonétisation de certaines pages ou comptes, restrictions publicitaires, baisse de distribution de contenus typiquement produits pour « faire du clic » (titres sensationnalistes, récits polarisants recyclés), ou perturbation des chaînes de valeur qui transforment l'attention en revenu. L'enjeu est de couper la boucle production–engagement–monétisation qui permet à la mésinformation d'être rentable et attire ainsi des mécaniques de diffusion et d'amplification allant à l'encontre d'un environnement informationnel sain. Cette typologie de contre-mesures s'inscrit dans le cadre des engagements du Code de conduite sur la désinformation, intégré au dispositif réglementaire du Digital Services Act européen, visant à lutter contre les risques systémiques en ligne, notamment en lien avec la mésinformation [86].

83 En français, traductible par dé-référencement

84 Laura Courchesne et al., "Review of Social Science Research on the Impact of Countermeasures against Influence Operations," *Harvard Kennedy School Misinformation Review*, ahead of print, September 13, 2021.

85 David Chavalarias et al., "Can a Single Line of Code Change Society? The Systemic Risks of Optimizing Engagement in Recommender Systems on Global Information Flow, Opinion Dynamics and Social Structures," *Journal of Artificial Societies and Social Simulation* 27, no. 1 (2024): 9.

86 European Commission, *Code of Conduct on Disinformation*.

Enfin, une partie de ces réponses vise la dimension artificielle et stratégique de l'amplification : détection et perturbation de comportements inauthentiques coordonnés (bots, réseaux de comptes, opérations de manipulation), sécurisation des leviers de visibilité (hashtags, automatisation, multi-comptes), et réduction de l'effet « popularité simulée » sur les signaux pris en compte par les plateformes [87].

Les modélisations scientifiques cherchant à mesurer l'efficacité de ces différentes contre-mesures constatent surtout un point opérationnel : pour être efficace individuellement, une contre-mesure systémique doit intervenir tôt et de manière très stricte. En revanche, une combinaison modérée de chaque levier d'action permettrait théoriquement de produire un effet cumulatif très supérieur à chaque intervention prise individuellement [88]. L'enjeu est moins de trouver « la » bonne mesure que de rendre la sur-visibilité systématiquement plus coûteuse et moins probable en cumulant les réponses au niveau systémique.

87 Jon Bateman, and Dean Jackson, *Countering Disinformation Effectively: An Evidence-Based Policy Guide - Carnegie Endowment for International Peace* (Washington, DC: Carnegie Endowment for International Peace, 2024), 1–130.

88 Joseph B. Bak-Coleman et al., "Combining Interventions to Reduce the Spread of Viral Misinformation," *Nature Human Behaviour* 6, no. 10 (2022): 1372–80.

5.2. Contre-mesures individuelles

Les contre-mesures individuelles visent à agir directement sur les pratiques informationnelles des utilisateurs, c'est-à-dire sur ce qu'ils croient, retiennent et partagent. Elles interviennent à différents moments du cycle informationnel : de manière préemptive, en amont de l'exposition, ou selon une logique corrective, au moment de l'exposition, du partage, voire après la diffusion du contenu. Ces interventions, nombreuses et hétérogènes, font l'objet de terminologies fluctuantes. Comme pour la mésinformation elle-même, cette instabilité conceptuelle complique l'évaluation comparative de leur efficacité [89]. À titre d'exemple, les termes labels, flagging ou content labeling sont fréquemment utilisés pour désigner des dispositifs très proches, voire identiques, sans que les distinctions soient toujours explicitées dans les études empiriques.

Les interventions correctives constituent la forme la plus visible de contre-mesure individuelle. Le fact-checking, entendu ici comme l'ensemble des dispositifs réactifs de correction, recouvre un large spectre allant de simples signalements visuels indiquant le caractère faux ou contesté d'un contenu, à des corrections plus élaborées reposant sur des stratégies de réfutation structurées, narratives ou fondées sur le rappel d'un consensus scientifique [90]. Malgré cette diversité de formats, l'objectif poursuivi demeure relativement stable : corriger des croyances factuellement inexactes et, dans certains cas, réduire les intentions de partage ou les niveaux d'engagement associés aux contenus trompeurs. La littérature souligne toutefois des limites récurrentes. Les corrections circulent généralement moins que les allégations qu'elles visent à contredire, leur efficacité dépend fortement de leur visibilité et de leur mise en forme, et leurs effets restent le plus souvent circonscrits aux affirmations ciblées, sans se généraliser à d'autres contenus [91]. Les travaux de Kahan et al. (2016), qui ont mis en évidence l'existence de réactions défensives lorsque des preuves contredisent des croyances fortement ancrées dans des contextes controversés, ont contribué à structurer un débat important

89 Anastasia Kozyreva et al., "Toolbox of Individual-Level Interventions against Online Misinformation," *Nature Human Behaviour* 8, no. 6 (2024): 1044–52.

90 R. Kelly Garrett and Shannon Poulsen, "Flagging Facebook Falsehoods: Self-Identified Humor Warnings Outperform Fact Checker and Peer Warnings," *Journal of Computer-Mediated Communication* 24, no. 5 (2019): 240–58; Lewandowsky et al., "The Debunking Handbook 2020."

91 Lisa Fazio et al., "Combating Misinformation: A Megastudy of Nine Interventions Designed to Reduce the Sharing of and Belief in False and Misleading Headlines," preprint, OSF, June 23, 2024.

autour des effets potentiellement contre-productifs du fact-checking [92]. Toutefois, les recherches ultérieures, fondées sur des protocoles variés et des contextes empiriques diversifiés, suggèrent que ces effets négatifs ne constituent pas un résultat systématique. Plusieurs études montrent en effet que les corrections factuelles n'entraînent généralement pas d'effets délétères mesurables, y compris auprès de publics idéologiquement sensibles [93].

Dans le prolongement de ces dispositifs correctifs, certaines interventions reposent sur la correction par les pairs. Celles-ci recouvrent des modalités distinctes selon leur degré de formalisation. L'invalidation sociale renvoie à des corrections produites directement dans les espaces conversationnels, notamment dans les commentaires, où l'efficacité dépend étroitement du style adopté et des coûts relationnels associés [94]. Corriger un autre utilisateur expose en effet à des risques de conflictualité, de perte de statut ou d'évitement, ce qui limite la fréquence et la portée de ces pratiques.

Le fact-checking crowdsourcé, à l'image des dispositifs de type Community Notes présents sur la plateforme X, formalise cette correction par les pairs en l'inscrivant dans une infrastructure dédiée [95]. Cette institutionnalisation permet de réduire certains coûts sociaux et d'augmenter la visibilité des corrections une fois publiées. Toutefois, leur impact reste fortement contraint par la latence du processus de validation collective des notes proposées par les contributeurs, qui laisse aux contenus le temps d'atteindre l'essentiel de leur audience, limitant son efficacité globale [96].

92 Dan M. Kahan et al., "Motivated Numeracy and Enlightened Self-Government," *Behavioural Public Policy* 1, no. 1 (2017): 54–86.

93 Philipp Schmid and Cornelia Betsch, "Effective Strategies for Rebutting Science Denialism in Public Discussions," *Nature Human Behaviour* 3, no. 9 (2019): 931–39; Brendan Nyhan, "Why the Backfire Effect Does Not Explain the Durability of Political Misperceptions," *Proceedings of the National Academy of Sciences* 118, no. 15 (2021): e1912440117; Briony Swire-Thompson et al., "The Backfire Effect after Correcting Misinformation Is Strongly Associated with Reliability," *Journal of Experimental Psychology. General* 151, no. 7 (2022): 1655–65.

94 Mohsen Mosleh et al., "Perverse Downstream Consequences of Debunking: Being Corrected by Another User for Posting False Political News Increases Subsequent Sharing of Low Quality, Partisan, and Toxic Content in a Twitter Field Experiment," *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA), CHI '21, mai 2021, 1–13.

95 Ariadna Matamoros-Fernández and Nadia Jude, "The Importance of Centering Harm in Data Infrastructures for 'Soft Moderation': X's Community Notes as a Case Study," *New Media & Society* 27, no. 4 (2025): 1986–2011.

96 Yuwei Chuai et al., "Did the Roll-Out of Community Notes Reduce Engagement With Misinformation on X/Twitter?," *Proc. ACM Hum.-Comput. Interact.* 8, no. CSCW2 (2024): 428:1-428:52; Thomas Renault et al., "Collaboratively Adding Context to Social Media Posts Reduces the Sharing of False News," arXiv:2404.02803, preprint, arXiv, April 3, 2024; Roberta Braga et al., *A Deep Dive into X's Community Notes: An Analysis of English and Spanish Contributions Between 2021 and 2025* (Digital Democracy Institute of the Americas, 2025), 1–21.

D'autres contre-mesures individuelles n'agissent pas directement sur les croyances, mais sur l'attention. Les nudges, ou incitations discrètes, cherchent à réorienter temporairement le traitement de l'information vers des critères épistémiques, tels que l'exactitude ou la crédibilité des contenus. Ces dispositifs prennent généralement la forme de rappels, de questions ou d'indices contextuels intégrés à l'interface, visant à interrompre des routines de partage largement automatisées dans les usages des réseaux sociaux. De nombreuses études montrent que ces interventions peuvent réduire la propension à partager des contenus trompeurs, en particulier lorsque l'attention est explicitement dirigée vers la question de la véracité [97]. Leur efficacité repose toutefois sur des effets de courte durée et sur des changements de comportement modestes. Elles agissent peu sur les croyances elles-mêmes et tendent à perdre en efficacité lorsque l'exposition est répétée ou lorsque les contenus sont fortement alignés avec des identités politiques ou sociales saillantes [98].

Concernant les capacités concrètes des utilisateurs, les interventions regroupées sous le terme de *"boosting"* visent à renforcer les compétences informationnelles des individus afin de favoriser une évaluation plus autonome des contenus [99]. Elles reposent sur l'hypothèse selon laquelle une partie de la vulnérabilité à la mésinformation tient à des compétences insuffisantes, qu'il s'agisse de connaissances factuelles, de capacités d'évaluation des sources ou de maîtrise des codes informationnels numériques. Ces interventions prennent des formes variées, allant de l'éducation aux médias et à l'information à des dispositifs plus ciblés tels que les conseils de vérification ou l'entraînement au raisonnement analytique [100]. La littérature montre qu'elles peuvent améliorer le discernement entre contenus fiables et trompeurs, y compris dans des contextes politisés. Leurs effets apparaissent toutefois dépendants de la durée et de l'intensité de l'exposition, et se révèlent souvent plus robustes sur l'évaluation de l'exactitude que sur les comportements de partage. Certaines études soulignent par ailleurs des effets secondaires, comme une surestimation de ses propres compétences,

97 Pennycook et al., "Shifting Attention to Accuracy Can Reduce Misinformation Online."

98 Fazio et al., "Combating Misinformation."

99 Kozyreva et al., "Toolbox of Individual-Level Interventions against Online Misinformation."

100 Sacha Altay et al., "Media Literacy Tips Promoting Reliable News Improve Discernment and Enhance Trust in Traditional Media," *Communications Psychology* 2, no. 1 (2024): 1–9.

susceptible de fragiliser le jugement dans des situations complexes [101].

Les approches par inoculation s'inscrivent quant à elles dans une logique explicitement préventive. Elles consistent à exposer les individus à des versions affaiblies ou explicitées de techniques de manipulation afin de renforcer leur résistance future à la mésinformation [102]. Contrairement au fact-checking, ces dispositifs ne ciblent pas des contenus spécifiques, mais des stratégies récurrentes telles que les sophismes, la polarisation ou les narratifs complotistes. La littérature indique que ces interventions peuvent produire des effets relativement durables sur le discernement, en particulier lorsqu'elles sont conçues de manière interactive ou ludique. Leur efficacité varie néanmoins selon les publics et les contextes, et reste conditionnée à la capacité de transférer les compétences acquises à des situations nouvelles. Leur déploiement à grande échelle pose en outre des défis opérationnels importants, notamment en termes de format et d'intégration dans les usages ordinaires des plateformes.

Enfin, certaines contre-mesures ciblent explicitement les dimensions identitaires et relationnelles de la réception de l'information. Ces approches partent du constat que la persistance de la mésinformation ne relève pas uniquement d'un déficit cognitif, mais tient aussi à son inscription dans des cadres moraux, politiques ou identitaires fortement investis. Les dispositifs regroupés sous le terme de self-affirmation visent ainsi à renforcer, en amont de l'exposition, le sentiment d'estime de soi et de cohérence identitaire des utilisateurs, afin de réduire les réactions défensives face à des contenus perçus comme menaçants sur le plan symbolique ou politique et de rétablir des conditions plus favorables à la réévaluation de l'information [103].

Les stratégies de mise en perspective, qui reposent sur des postures d'écoute et de reconnaissance du point de vue de l'interlocuteur – parfois médiatisées par

101 Thomas Nygren et al., "Boosting Teenagers' Skills to Fact-Check Photos and Problems of Overconfidence: Impact and Side Effects in Education against Misinformation from Ukraine," preprint, Open Science Framework, September 24, 2024.

102 Josh Compton et al., "Inoculation Theory in the Post-truth Era: Extant Findings and New Frontiers for Contested Science, Misinformation, and Conspiracy Theories," *Social and Personality Psychology Compass* 15, no. 6 (2021): e12602; Stephan Lewandowsky and Sander Van Der Linden, "Countering Misinformation and Fake News Through Inoculation and Prebunking," *European Review of Social Psychology* 32, no. 2 (2021): 348–84; Trisha Harjani et al., "Gamified Inoculation Against Misinformation in India: A Randomized Control Trial," *Journal of Trial & Error* 3, no. 1 (2023).

103 Ecker et al., "The Psychological Drivers of Misinformation Belief and Its Resistance to Correction"; Ziemer and Rothmund, "Psychological Underpinnings of Misinformation Countermeasures"; Camille J. Saucier et al., "Leveraging Motivations to Curb Misinformation," *Journal of Media Psychology* 37, no. 5 (2025): 316–22.

des agents conversationnels – s'inscrivent dans une logique similaire. Les résultats empiriques suggèrent que ces interventions peuvent atténuer certains effets de polarisation et faciliter l'acceptation de corrections. Leur efficacité demeure toutefois étroitement dépendante du contexte d'interaction et de la crédibilité perçue des interlocuteurs, ce qui limite leur transposition à grande échelle dans des environnements de communication de masse [104].

104 Sacha Yesilaltay, "Understanding Misinformation and Fighting for Information" (phdthesis, Université Paris sciences et lettres, 2021); Thomas H. Costello et al., "Durably Reducing Conspiracy Beliefs through Dialogues with AI," preprint, OSF, April 3, 2024.

Conclusion

La mésinformation ne constitue ni une anomalie marginale, ni le simple produit de comportements individuels déficients. Elle s’inscrit dans un environnement informationnel structuré par des architectures techniques, des modèles économiques et des dynamiques sociales qui conditionnent sa circulation, sa visibilité et sa persistance.

Les débats publics tendent pourtant à surestimer le rôle de la crédulité individuelle et à sous-estimer les effets cumulatifs de l'amplification médiatique, algorithmique et stratégique. En réalité, les vulnérabilités informationnelles sont fortement relationnelles et contextuelles. Elles résultent de l'interaction entre des mécanismes cognitifs ordinaires, des dynamiques émotionnelles, des logiques d'appartenance sociale et des environnements techniques conçus pour capter l'attention. Dans ce cadre, partager un contenu trompeur ne signifie pas nécessairement y adhérer, ni chercher à tromper autrui, mais s'inscrire dans des usages sociaux où l'exactitude n'est pas toujours la norme prioritaire.

Les réponses à la mésinformation gagnent ainsi à être pensées comme des ajustements de l'environnement informationnel plutôt que comme des tentatives de correction morale ou cognitive des individus. Les contre-mesures systémiques, en agissant sur la sur-visibilité et la rentabilité des contenus trompeurs, constituent un levier central de la lutte contre la désinformation. Leur mise en œuvre demeure toutefois conflictuelle, dans la mesure où elles entrent en tension directe avec les intérêts économiques des plateformes. Les interventions individuelles, bien que produisant des effets limités, dépendants du contexte, du format et des dynamiques identitaires, jouent ainsi un rôle complémentaire à ne pas négliger.

Il est tentant de fantasmer un environnement informationnel lisse, utopiquement tourné vers un débat démocratique apaisé. Or, les recherches sur l'histoire de la communication montrent que la mésinformation n'est pas née avec les réseaux sociaux et ne disparaitra pas avec eux. L'enjeu n'est donc pas de faire disparaître toute conflictualité informationnelle, mais de limiter les distorsions qu'introduit la mésinformation dans la perception de l'opinion publique sur les enjeux contemporains tels que le changement climatique, les inégalités sociales ou la santé ; en particulier lors de moments démocratiques critiques tels que les élections ou les mobilisations collectives. En proposant une synthèse des connaissances sur les formes, les mécanismes de diffusion et les réponses à la mésinformation, ce rapport vise à contribuer à une mise en œuvre plus éclairée et opérationnelle des contre-mesures à la mésinformation.

Bibliographie

Altay, Sacha, Manon Berriche, Hendrik Heuer, Johan Farkas, and Steven Rathje. "A Survey of Expert Views on Misinformation: Definitions, Determinants, Solutions, and Future of the Field." *Harvard Kennedy School Misinformation Review*, ahead of print, July 27, 2023.

Altay, Sacha, Andrea De Angelis, and Emma Hoes. "Media Literacy Tips Promoting Reliable News Improve Discernment and Enhance Trust in Traditional Media." *Communications Psychology* 2, no. 1 (2024): 1–9.

American Psychological Association. "Recommendations for Countering Misinformation." Apa.Org, American Psychological Association, November 29, 2023.

Andersen, Jack, and Sille Obelitz Søe. "Communicative Actions We Live by: The Problem with Fact-Checking, Tagging or Flagging Fake News – the Case of Facebook." *European Journal of Communication* 35, no. 2 (2020): 126–39.

Bak-Coleman, Joseph B., Ian Kennedy, Morgan Wack, et al. "Combining Interventions to Reduce the Spread of Viral Misinformation." *Nature Human Behaviour* 6, no. 10 (2022): 1372–80.

Baribi-Bartov, Sahar, Briony Swire-Thompson, and Nir Grinberg. "Supersharers of Fake News on Twitter." *Science* 384, no. 6699 (2024): 979–82.

Bateman, Jon, and Dean Jackson. *Countering Disinformation Effectively: An Evidence-Based Policy Guide - Carnegie Endowment for International Peace*. Washington, DC: Carnegie Endowment for International Peace, 2024.

Begg, Ian Maynard, Ann Anas, and Suzanne Farinacci. "Dissociation of Processes in Belief: Source Recollection, Statement Familiarity, and the Illusion of Truth." *Journal of Experimental Psychology: General* (US) 121, no. 4 (1992): 446–58.

Benkler, Yochai, Robert Farris, and Hal Roberts. *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. Oxford University Press, 2018.

Berlinski, Nicolas, Margaret Doyle, Andrew M. Guess, et al. "The Effects of Unsubstantiated Claims of Voter Fraud on Confidence in Elections." *Journal of Experimental Political Science* 10, no. 1 (2023): 34–49.

Berriche, Manon, and Sacha Altay. "Internet Users Engage More with Phatic Posts than with Health Misinformation on Facebook." *Palgrave Communications* 6, no. 1 (2020): 1–9.

Bontcheva, Kalina, Symeon Papadopoulous, Filareti Tsalakanidou, et al. *Generative AI and Disinformation: Recent Advances, Challenges, and Opportunities*. Edited by Kalina Bontcheva. TITAN, AI4Media, AI4Trust, 2024.

Bontridder, Noémi, and Yves Poullet. "The Role of Artificial Intelligence in Disinformation." *Data & Policy* 3 (January 2021): e32.

Boursier, Tristan. "Influenceurs d'extrême droite : le moteur caché du succès du RN." The Conversation, The Conversation, June 20, 2024.

boyd, Danah, and Michael Golebiewski. *Data Voids: Where Missing Data Can Easily Be Exploited*. New York City: Data & Society Research Institute, 2019.

Bradshaw, Samantha, and Philip N. Howard. "The Global Organization of Social Media Disinformation Campaigns." *Journal of International Affairs* 71, no. 1.5 (2018): 23–32.

Brady, William J., Julian A. Wills, John T. Jost, Joshua A. Tucker, and Jay J. Van Bavel. "Emotion Shapes the Diffusion of Moralized Content in Social Networks." *Proceedings of the National Academy of Sciences* 114, no. 28 (2017): 7313–18.

Braga, Roberta, Cristina Tardáguila, and Marcelo Soares. *A Deep Dive into X's Community Notes: An Analysis of English and Spanish Contributions Between 2021 and 2025*. Digital Democracy Institute of the Americas, 2025.

Broda, Elena, and Jesper Strömbäck. "Misinformation, Disinformation, and Fake News: Lessons from an Interdisciplinary, Systematic Literature Review." *Annals of the International Communication Association* 48, no. 2 (2024): 139–66.

Buchanan, Tom, and James Kempley. "Individual Differences in Sharing False Political Information on Social Media: Direct and Indirect Effects of Cognitive-Perceptual Schizotypy and Psychopathy." *Personality and Individual Differences* 182 (November 2021): 111071.

Budak, Ceren, Brendan Nyhan, David M. Rothschild, Emily Thorson, and Duncan J. Watts. "Misunderstanding the Harms of Online Misinformation." *Nature* 630, no. 8015 (2024): 45–53.

Cazzamatta, Regina. "Decoding Correction Strategies: How Fact-Checkers Uncover Falsehoods Across Countries." *Journalism Studies*, March 3, 2025, 1–23.

Chavalarias, David. *Toxic data: comment les réseaux manipulent nos opinions*. Paris: Flammarion, 2025.

Chavalarias, David, Paul Bouchaud, and Maziyar Panahi. "Can a Single Line of Code Change Society? The Systemic Risks of Optimizing Engagement in Recommender Systems on Global Information Flow, Opinion Dynamics and Social Structures." *Journal of Artificial Societies and Social Simulation* 27, no. 1 (2024): 9.

Cherilyn Ireton, Julie Posetti. *Journalism, fake news & disinformation: handbook for journalism education and training*. 2018.

Chuai, Yuwei, Haoye Tian, Nicolas Pröllochs, and Gabriele Lenzini. "Did the Roll-Out of Community Notes Reduce Engagement With Misinformation on X/Twitter?" *Proc. ACM Hum.-Comput. Interact.* 8, no. CSCW2 (2024): 428:1-428:52.

CISA. *Tactics of Di Sinformation*. Cybersecurity and Infrastructure Security Agency (CISA), 2022.

Compton, Josh, Sander Van Der Linden, John Cook, and Melisa Basol. "Inoculation Theory in the Post-truth Era: Extant Findings and New Frontiers for Contested Science, Misinformation, and Conspiracy Theories." *Social and Personality Psychology Compass* 15, no. 6 (2021): e12602.

Cook, John. "Deconstructing Climate Science Denial." In *Research Handbook on Communicating Climate Change*, edited by David C. Holmes and Lucy M. Richardson. Edward Elgar Publishing, 2020.

Cook, John, Ullrich K. H. Ecker, Melanie Trecek-King, et al. "The Cranky Uncle Game—Combining Humor and Gamification to Build Student Resilience against Climate Misinformation." *Environmental Education Research* 29, no. 4 (2023): 607–23.

Cook, John, Chelsey Lepage, Kathryn L. Hopkins, et al. "Co-Designing and Pilot Testing a Digital Game to Improve Vaccine Attitudes and Misinformation Resistance in Ghana." *Human Vaccines & Immunotherapeutics* 20, no. 1 (2024): 2407204.

Costello, Thomas H., Gordon Pennycook, and David Rand. "Durably Reducing Conspiracy Beliefs through Dialogues with AI." Preprint, OSF, April 3, 2024.

Courchesne, Laura, Julia Ilhardt, and Jacob N. Shapiro. "Review of Social Science Research on the Impact of Countermeasures against Influence Operations." *Harvard Kennedy School Misinformation Review*, ahead of print, September 13, 2021.

Dan, Viorela, Britt Paris, Joan Donovan, et al. "Visual Mis- and Disinformation, Social Media, and Democracy." *Journalism & Mass Communication Quarterly* 98, no. 3 (2021): 641–64.

Data for Good, Quota Climat, and Science Feedback. *Cartographie de La Désinformation Climatique Dans Les Médias Français et Brésiliens*. 2025.

Diakopoulos, Nicholas, and Deborah Johnson. "Anticipating and Addressing the Ethical Implications of Deepfakes in the Context of Elections." *New Media & Society* 23, no. 7 (2021): 2072–98.

Dobber, Tom, Nadia Metoui, Damian Trilling, Natali Helberger, and Claes De Vreese. "Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes?" *The International Journal of Press/Politics* 26, no. 1 (2021): 69–91.

Donovan, Joan, and Brian Friedberg. *Source Hacking. Media Manipulation in Practice*. Media Manipulation Research Initiative. Data & Society Research Institute, 2019.

Echeverria, Martin, Sara García Santamaría, and Daniel Hallin. *Introduction. Deceiving from the Top: State-Sponsored Disinformation as a Contemporary Global Phenomenon*. 2025.

Ecker, Ullrich K. H., Stephan Lewandowsky, John Cook, et al. "The Psychological Drivers of Misinformation Belief and Its Resistance to Correction." *Nature Reviews Psychology* 1, no. 1 (2022): 13–29.

European Commission. *Code of Conduct on Disinformation [as Amended in October 2024]*. Brussels: European Commission, 2024.

Evans, Jonathan, and Keith Stanovich. "Dual-Process Theories of Higher Cognition." *Perspectives on Psychological Science* 8 (May 2013): 223–41.

Eysenbach, Gunther. "Infodemiology: The Epidemiology of (Mis)Information." *The American Journal of Medicine* 113, no. 9 (2002): 763–65.

Fazio, Lisa, David Rand, Stephan Lewandowsky, et al. "Combating Misinformation: A Megastudy of Nine Interventions Designed to Reduce the Sharing of and Belief in False and Misleading Headlines." Preprint, OSF, June 23, 2024.

Fernández, Miriam, Alejandro Bellogín, and Iván Cantador. "Analysing the Effect of Recommendation Algorithms on the Amplification of Misinformation." arXiv:2103.14748. Preprint, arXiv, March 26, 2021.

Forgas, Joseph. "Don't Worry, Be Sad! On the Cognitive, Motivational, and Interpersonal Benefits of Negative Mood." *Current Directions in Psychological Science* 22 (June 2013): 225–32.

Forgas, Joseph P., and Rebekah East. "On Being Happy and Gullible: Mood Effects on Skepticism and the Detection of Deception." *Journal of Experimental Social Psychology* (Netherlands) 44, no. 5 (2008): 1362–67.

Forum Information Democracy. "The International Partnership for Information and Democracy." Forum Information Democracy, September 2019.

Garrett, R. Kelly, and Shannon Poulsen. "Flagging Facebook Falsehoods: Self-Identified Humor Warnings Outperform Fact Checker and Peer Warnings." *Journal of Computer-Mediated Communication* 24, no. 5 (2019): 240–58.

Giglietto, Fabio, Laura Iannelli, Augusto Valeriani, and Luca Rossi. "'Fake News' Is the Invention of a Liar: How False Information Circulates within the Hybrid News System." *Current Sociology* 67, no. 4 (2019): 625–42.

Gilovich, Thomas, Dale Griffin, and Daniel Kahneman, eds. *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge: Cambridge University Press, 2002.

Graber, Doris A. "Seeing Is Remembering: How Visuals Contribute to Learning from Television News." *Journal of Communication* 40, no. 3 (1990): 134–56.

Grinberg, Nir, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. "Fake News on Twitter during the 2016 U.S. Presidential Election." *Science (New York, N.Y.)* 363, no. 6425 (2019): 374–78.

Grohmann, Rafael, and Jonathan Corpus Ong. "Disinformation-for-Hire as Everyday Digital Labor: Introduction to the Special Issue." *Social Media + Society* 10, no. 1 (2024): 20563051231224723.

Guess, Andrew M., Michael Lerner, Benjamin Lyons, et al. "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* 117, no. 27 (2020): 15536–45.

Hameleers, Michael, Thomas E. Powell, Toni G. L. A. Van Der Meer, and Lieke Bos. "A Picture Paints a Thousand Lies? The Effects and Mechanisms of Multimodal Disinformation and Rebuttals Disseminated via Social Media." *Political Communication* 37, no. 2 (2020): 281–301.

Hannah, Matthew N. "A Conspiracy of Data: QAnon, Social Media, and Information Visualization." *Social Media + Society* 7, no. 3 (2021): 20563051211036064.

Harjani, Trisha, Melisa-Sinem Basol, Jon Roozenbeek, and Sander van der Linden. "Gamified Inoculation Against Misinformation in India: A Randomized Control Trial." *Journal of Trial & Error* 3, no. 1 (2023).

Harsin, Jayson. "Post-Truth and Critical Communication Studies." In *Oxford Research Encyclopedia of Communication*, by Jayson Harsin. Oxford University Press, 2018.

Hilberts, Sonya, Mark Govers, Elena Petelos, and Silvia Evers. "The Impact of Misinformation on Social Media in the Context of Natural Disasters: Narrative Review." *JMIR Infodemiology* 5, no. 1 (2025): e70413.

Howard, Jonathan, and Dorit Rubinstein Reiss. "The Anti-Vaccine Movement: A Litany of Fallacy and Errors." In *Pseudoscience: The Conspiracy Against Science*, edited by Allison B. Kaufman and James C. Kaufman, 0. The MIT Press, 2018.

Hugo Mercier. *Not Born Yesterday | Princeton University Press*. 2020.

Ipsos. *Elections & Social Media: The Battle against Disinformation and Trust Issues | Ipsos*. UNESCO, 2023.

Iyer, Aarti, Joanna Webster, Matthew J. Hornsey, and Eric J. Vanman. "Understanding the Power of the Picture: The Effect of Image Content on Emotional and Political Responses to Terrorism." *Journal of Applied Social Psychology* 44, no. 7 (2014): 511–21.

Jack, Caroline. *Lexicon of Lies. Terms for Problematic Information*. New York City: Data & Society Research Institute, 2017.

Jenkins, Henry. "Photoshop for Democracy." Research. *MIT Technology Review*, April 6, 2004.

Jonas, Eva, Ian McGregor, Johannes Klackl, et al. "Chapter Four - Threat and Defense: From Anxiety to Approach." In *Advances in Experimental Social Psychology*, vol. 49, edited by James M. Olson and Mark P. Zanna, 219–86. Academic Press, 2014.

Kahan, Dan M. "The Politically Motivated Reasoning Paradigm, Part 1: What Politically Motivated Reasoning Is and How to Measure It." In *Emerging Trends in the Social and Behavioral Sciences*, 1–16. John Wiley & Sons, Ltd, 2016.

Kahan, Dan M., Ellen Peters, Erica Cantrell Dawson, and Paul Slovic. "Motivated Numeracy and Enlightened Self-Government." *Behavioural Public Policy* 1, no. 1 (2017): 54–86.

Kata, Anna. "Anti-Vaccine Activists, Web 2.0, and the Postmodern Paradigm – An Overview of Tactics and Tropes Used Online by the Anti-Vaccination Movement." *Vaccine*, Special Issue: The Role of Internet Use in Vaccination Decisions, vol. 30, no. 25 (2012): 3778–89.

Kensinger, Elizabeth A. "Remembering the Details: Effects of Emotion." *Emotion Review: Journal of the International Society for Research on Emotion* 1, no. 2 (2009): 99–113.

Khalilizadeh, Kamilia. *Retour sur la campagne vue des réseaux sociaux | Ipsos*. Ipsos, 2024.

King, Catherine, Samantha C. Phillips, and Kathleen M. Carley. "A Path Forward on Online Misinformation Mitigation Based on Current User Behavior." *Scientific Reports* 15, no. 1 (2025): 9475.

———. "A Path Forward on Online Misinformation Mitigation Based on Current User Behavior." *Scientific Reports* 15, no. 1 (2025): 9475.

Koch, Alex S., and Joseph P. Forgas. "Feeling Good and Feeling Truth: The Interactive Effects of Mood and Processing Fluency on Truth Judgments." *Journal of Experimental Social Psychology* (Netherlands) 48, no. 2 (2012): 481–85.

Kozyreva, Anastasia, Philipp Lorenz-Spreen, Stefan M. Herzog, et al. "Toolbox of Individual-Level Interventions against Online Misinformation." *Nature Human Behaviour* 8, no. 6 (2024): 1044–52.

Kuklinski, James H., Paul J. Quirk, Jennifer Jerit, David Schwieder, and Robert F. Rich. "Misinformation and the Currency of Democratic Citizenship." *The Journal of Politics* 62, no. 3 (2000): 790–816.

Kunda, Ziva. "The Case for Motivated Reasoning." *Psychological Bulletin* (US) 108, no. 3

(1990): 480–98.

Lewandowsky, S., Al-Rawi, A. K., Gavaruzzi, T., et al. *The COVID-19 Vaccine Communication Handbook. A practical guide for improving vaccine communication and fighting misinformation.* 2021.

Lewandowsky, Stephan, John Cook, Ullrich Ecker, et al. "The Debunking Handbook 2020." *Copyright, Fair Use, Scholarly Communication, Etc.*, January 1, 2020, 19.

Lewandowsky, Stephan, Ullrich K. H. Ecker, and John Cook. "Beyond Misinformation: Understanding and Coping with the 'Post-Truth' Era." *Journal of Applied Research in Memory and Cognition* 6, no. 4 (2017): 353–69.

———. "Liars Know They Are Lying: Differentiating Disinformation from Disagreement." *Humanities and Social Sciences Communications* 11, no. 1 (2024): 986.

Lewandowsky, Stephan, and Sander Van Der Linden. "Countering Misinformation and Fake News Through Inoculation and Prebunking." *European Review of Social Psychology* 32, no. 2 (2021): 348–84.

Littrell, Shane, Casey Klofstad, Amanda Diekman, et al. "Who Knowingly Shares False Political Information Online?" *Harvard Kennedy School Misinformation Review*, ahead of print, August 25, 2023.

MacKuen, Michael, Jennifer Wolak, Luke Keele, and George E. Marcus. "Civic Engagements: Resolute Partisanship or Reflective Deliberation." *American Journal of Political Science* 54, no. 2 (2010): 440–58.

Margolin, Emma. *10 Tips for Reporting on Disinformation*. Tip sheet. Data & Society Research Institute, 2020.

Matamoros-Fernández, Ariadna, and Nadia Jude. "The Importance of Centering Harm in Data Infrastructures for 'Soft Moderation': X's Community Notes as a Case Study." *New Media & Society* 27, no. 4 (2025): 1986–2011.

McCosker, Anthony. "Trolling as Provocation: YouTube's Agonistic Publics." *Convergence: The International Journal of Research into New Media Technologies* 20, no. 2 (2014): 201–17.

Molina, Maria D., S. Shyam Sundar, Thai Le, and Dongwon Lee. "'Fake News' Is Not Simply False Information: A Concept Explication and Taxonomy of Online Content." *American Behavioral Scientist* 65, no. 2 (2021): 180–212.

Mosleh, Mohsen, Cameron Martel, Dean Eckles, and David Rand. "Perverse Downstream Consequences of Debunking: Being Corrected by Another User for Posting False Political News Increases Subsequent Sharing of Low Quality, Partisan, and Toxic Content in a Twitter Field Experiment." *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA), CHI '21, mai 2021, 1–13.

Muñoz, Pau, Raúl Barba-Rojas, Fernando Díez, and Alejandro Bellogín. "The Role of Recommendation Algorithms in the Formation of Disinformation Networks." *Information Processing & Management* 62, no. 6 (2025): 104243.

Murphy, Gillian, and Emma Flynn. "Deepfake False Memories." *Memory* 30, no. 4 (2022): 480–92.

Neumann, john Von. *Theory Of Games And Economic Behavior*. 1944.

Ng, Lynnette Hui Xian, and Kathleen M. Carley. "A Global Comparison of Social Media Bot and Human Characteristics." *Scientific Reports* 15, no. 1 (2025): 10973.

Nygren, Thomas, Markus Al-Afifi, and Carl-Anton Werner Axelsson. "Boosting Teenagers' Skills to Fact-Check Photos and Problems of Overconfidence: Impact and Side Effects in Education against Misinformation from Ukraine." Preprint, Open Science Framework, September 24, 2024.

Nyhan, Brendan. "Why the Backfire Effect Does Not Explain the Durability of Political Misperceptions." *Proceedings of the National Academy of Sciences* 118, no. 15 (2021): e1912440117.

Nyhan, Brendan, and Jason Reifler. "The Roles of Information Deficits and Identity Threat in the Prevalence of Misperceptions." *Journal of Elections, Public Opinion and Parties*, April 3, 2019.

World.

OMS, Organisation Mondiale De La Santé. "Première conférence de l'OMS sur l'infodémiologie." Conférence. Première conférence de l'OMS sur l'infodémiologie, June 29, 2020.

Paris, Britt, and Joan Donovan. *Data & Society — Deepfakes and Cheap Fakes*. New York City: Data & Society Research Institute, 2019.

Pathak, Royal, Francesca Spezzano, and Maria Soledad Pera. "Understanding the Contribution of Recommendation Algorithms on Misinformation Recommendation and Misinformation Dissemination on Social Networks." *ACM Trans. Web* 17, no. 4 (2023): 35:1-35:26.

Pennycook, Gordon, Tyrone D. Cannon, and David G. Rand. "Prior Exposure Increases Perceived Accuracy of Fake News." *Journal of Experimental Psychology: General* (US) 147, no. 12 (2018): 1865–80.

Pennycook, Gordon, Ziv Epstein, Mohsen Mosleh, Antonio A. Arechar, Dean Eckles, and David G. Rand. "Shifting Attention to Accuracy Can Reduce Misinformation Online." *Nature* 592, no. 7855 (2021): 590–95.

Pennycook, Gordon, and David G. Rand. "Lazy, Not Biased: Susceptibility to Partisan Fake News Is Better Explained by Lack of Reasoning than by Motivated Reasoning." *Cognition*, The Cognitive Science of Political Thought, vol. 188 (July 2019): 39–50.

———. "The Psychology of Fake News." *Trends in Cognitive Sciences* 25, no. 5 (2021): 388–402.

Persky, Joseph. "The Ethology of Homo Economicus." *Journal of Economic Perspectives* 9, no. 2 (1995): 221–31.

Petty, Richard E., and Pablo Briñol. "Emotion and Persuasion: Cognitive and Meta-Cognitive Processes Impact Attitudes." *Cognition and Emotion*, January 2, 2015. World.

Phillips, Whitney. *The Oxygen of Amplification. Better Practices for Reporting on Extremists, Antagonists, and Manipulators.* Data & Society Research Institute, 2018.

Powell, Thomas E., Hajo G. Boomgaarden, Knut De Swert, and Claes H. de Vreese. "A Clearer Picture: The Contribution of Visuals and Text to Framing Effects." *Journal of Communication* (United Kingdom) 65, no. 6 (2015): 997–1017.

Renault, Thomas, David Restrepo Amariles, and Aurore Troussel. "Collaboratively Adding Context to Social Media Posts Reduces the Sharing of False News." arXiv:2404.02803. Preprint, arXiv, April 3, 2024.

Roozenbeek, Jon, and Sander Van der Linden. "Breaking Harmony Square: A Game That 'Inoculates' against Political Misinformation." *Harvard Kennedy School Misinformation Review*, ahead of print, November 6, 2020.

Salvi, Carola, Paola Iannello, Alice Cancer, et al. "Going Viral: How Fear, Socio-Cognitive Polarization and Problem-Solving Influence Fake News Detection and Proliferation During COVID-19 Pandemic." *Frontiers in Communication* 5 (January 2021).

Saucier, Camille J., Christopher Calabrese, and Nathan Walter. "Leveraging Motivations to Curb Misinformation." *Journal of Media Psychology* 37, no. 5 (2025): 316–22.

Scherer, Laura D., and Gordon Pennycook. "Who Is Susceptible to Online Health Misinformation?" *American Journal of Public Health* 110, no. S3 (2020): S276–77.

Schmid, Philipp, and Cornelia Betsch. "Effective Strategies for Rebutting Science Denialism in Public Discussions." *Nature Human Behaviour* 3, no. 9 (2019): 931–39.

Schopenhauer, Arthur. *L'art d'avoir toujours raison: la dialectique éristique*. Translated by Dominique-Laure Miermont-Grente. With Didier Raymond. La petite collection. Paris: Mille et une nuits, 2021.

Seigneurin, Marion, Christine Balagué, and Inna Lyubareva. "Navigating Misinformation and Disinformation: How Definition Ambiguity Limits the DSA's Implementation." *European Journal of Communication* 40, no. 6 (2025): 619–46.

Simis, Molly J., Haley Madden, Michael A. Cacciatore, and Sara K. Yeo. "The Lure of Rationality:

Why Does the Deficit Model Persist in Science Communication?" *Public Understanding of Science (Bristol, England)* 25, no. 4 (2016): 400–414.

Sirlin, Nathaniel, Ziv Epstein, Antonio A. Arechar, and David G. Rand. "Digital Literacy Is Associated with More Discerning Accuracy Judgments but Not Sharing Intentions." *Harvard Kennedy School Misinformation Review*, ahead of print, December 6, 2021.

Sismondo, Sergio. "Post-Truth?" *Social Studies of Science* 47, no. 1 (2017): 3–6.

Stephan, Gaël, and Stéphanie Wojcik. "Engagement et Ethos de l'extrême Droite En Ligne : Militantes et Militants de Reconquête! Sur Instagram:" *Quaderni* n° 111, no. 1 (2024): 83–102.

Stolle, Lucas B., Rohit Nalamasu, Joseph V. Pergolizzi, et al. "Fact vs Fallacy: The Anti-Vaccine Discussion Reloaded." *Advances in Therapy* 37, no. 11 (2020): 4481–90.

Sun, Yanqing, and Juan Xie. "Who Shares Misinformation on Social Media? A Meta-Analysis of Individual Traits Related to Misinformation Sharing." *Computers in Human Behavior* 158 (September 2024): 108271.

Sundar, S. Shyam. "The MAIN Model: A Heuristic Approach to Understanding Technology Effects on Credibility." In *Digital Media, Youth, and Credibility*, edited by Miriam J. Metzger and Andrew J. Flanagin, 73–100. 2008; Cambridge: The MIT Press, 2008.

Swire-Thompson, Briony, Nicholas Miklaucic, John P. Wihbey, David Lazer, and Joseph DeGutis. "The Backfire Effect after Correcting Misinformation Is Strongly Associated with Reliability." *Journal of Experimental Psychology. General* 151, no. 7 (2022): 1655–65.

Tversky, Amos, and Daniel Kahneman. "Judgment under Uncertainty: Heuristics and Biases." *Science* 185, no. 4157 (1974): 1124–31.

Vaccari, Cristian, and Andrew Chadwick. "Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News." *Social Media + Society* 6, no. 1 (2020): 2056305120903408.

Van Bavel, Jay, Elizabeth Harris, Philip Pärnamets, Steve Rathje, Kimberly Doell, and Joshua Tucker. "Political Psychology in the Digital (Mis)Information Age: A Model of News Belief and Sharing." *Social Issues and Policy Review* 15, no. 1 (2021): 84–113.

Van der Linden, Sander. *Foolproof: Why We Fall for Misinformation and How to Build Immunity*. London: 4th Estate, 2023.

Van Der Linden, Sander. "Misinformation: Susceptibility, Spread, and Interventions to Immunize the Public." *Nature Medicine* 28, no. 3 (2022): 460–67.

Van der Linden, Sander, and Yara Kyrychenko. "A Broader View of Misinformation Reveals Potential for Intervention." *Science* 384, no. 6699 (2024): 959–60.

Van Prooijen, Jan-Willem, and Karen M. Douglas. "Belief in Conspiracy Theories: Basic Principles of an Emerging Research Domain." *European Journal of Social Psychology* 48, no. 7 (2018): 897–908.

Vie Publique. "Ingérences étrangères et IA : une menace pour les démocraties ? | vie-publique.fr." Information. vie-publique.fr, December 29, 2025.

Viginum. *Manipulation d'algorithmes et instrumentalisation d'influenceurs : enseignements de l'élection présidentielle en Roumanie & risques pour la France*. Enjeux systémiques. SGDSN, 2025.

Vosoughi, Soroush, Deb Roy, and Sinan Aral. "The Spread of True and False News Online." *Science* 359, no. 6380 (2018): 1146–51.

Wanless, A., and J. Pamment. "How Do You Define a Problem Like Influence?" *Journal of Information Warfare* 18, no. 3 (2019): 1–14.

Wanless, Alicia, Samantha Lai, and John Hicks. *Assessing National Information Ecosystems*. Washington, DC: Carnegie Endowment for International Peace, 2025.

Wardle, Claire, and Hossein Derakhshan. *Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making*. No. 27. Council of Europe, 2017.

Watts, Duncan J., and David M. Rothschild. "Don't Blame the Election on Fake News. Blame It on the Media." Columbia Journalism Review, May 5, 2025.

Wawrzuta, Dominik, Mariusz Jaworski, Joanna Gotlib, and Mariusz Panczyk. "Characteristics of Antivaccine Messages on Social Media: Systematic Review." *Journal of Medical Internet Research* 23, no. 6 (2021): e24564.

Weeks, Brian E. "Emotions, Partisanship, and Misperceptions: How Anger and Anxiety Moderate the Effect of Partisan Bias on Susceptibility to Political Misinformation." *Journal of Communication* 65, no. 4 (2015): 699–719.

Weikmann, Teresa, Hannah Greber, and Alina Nikolaou. "After Deception: How Falling for a Deepfake Affects the Way We See, Hear, and Experience Media." *The International Journal of Press/Politics* 30, no. 1 (2025): 187–210.

World Economic Forum. *Global Risks Report 2025*. No. 20. Global Risks Report. Genève: World Economic Forum, 2025.

Yesilaltay, Sacha. "Understanding Misinformation and Fighting for Information." Phdthesis, Université Paris sciences et lettres, 2021.

Zacharia, Janine, and Andrew Grotto. *How to Responsibly Report on Hacks and Disinformation*. California: Stanford Cyber Policy Center, 2020.

Ziemer, Carolin-Theresa, and Tobias Rothmund. "Psychological Underpinnings of Misinformation Countermeasures." *Journal of Media Psychology* 36, no. 6 (2024): 397–409.

Zimmer, Franziska, Katrin Scheibe, Mechtild Stock, and Wolfgang G. Stock. "Fake News in Social Media: Bad Algorithms or Biased Users?" *Journal of Information Science Theory and Practice* 7, no. 2 (2019): 40–53.

Zimmerman, Richard K., Robert M. Wolfe, Dwight E. Fox, et al. "Vaccine Criticism on the World Wide Web." *Journal of Medical Internet Research* 7, no. 2 (2005): e17.

Zollo, Fabiana, Alessandro Bessi, Michela Del Vicario, et al. "Debunking in a World of Tribes." *PLOS ONE* 12, no. 7 (2017): e0181821