DÉSINFORMATION : DE QUOI PARLE-T-ON ?

UN DÉFI CONCEPTUEL

EVIDENCES
LE THINK TANK DÉDIÉ À LA SCIENCE DANS LA SOCIÉTÉ

Synthèse	3
Introduction	5
1. Le problème du lexique et de la taxonomie	6
1.1. Information	6
1.2. Quel est le contraire de l'information ?	7
1.3. Désinformation et Mésinformation	8
1.4. Des catégories insuffisantes au regard du réel à décrire	10
1.5. « The practical rationale is false » (Simion, 2024)	12
2. Le problème des modèles	14
2.1. Le modèle de la contagion	14
2.2. Le modèle de l'auto-support	15
2.3. Le modèle KABP	17
2.4. Le modèle « théâtre impro »	19
3. Pourquoi aucun cadre ne parvient réellement à s'imposer ?	21
3.1. La véracité comme référentiel : un renoncement collectif de fait	22
3.2. Le référentiel de l'authenticité	23
3.3. Le référentiel de l'intégrité	23
3.4. La désinformation met structurellement en échec la méthode scientifique elle-même	24
Références	28

Le think tank Evidences s'adresse à tous ceux qui défendent la valeur politique de l'activité scientifique dans la société, au service de l'émancipation, de la croissance et du progrès. Le lien entre science et société est aujourd'hui au carrefour de multiples enjeux économiques, sociaux et politiques. La France et l'Europe voient leur situation politique, économique et sociale menacée. Notre pays a vu décliner sa position dans plusieurs domaines scientifiques et techniques – et s'accroître d'autant ses dépendances industrielles et technologiques. En parallèle, appuyés sur l'obscurantisme et la désinformation, les populismes progressent et fragilisent notre vie démocratique. C'est cette dynamique globale qu'il s'agit d'enrayer ensemble sans attendre. La puissance émancipatrice du progrès scientifique et technique est au cœur du projet démocratique. Elle passe par la valorisation de la culture scientifique dans la société et par le dialogue éclairé, transparent, confiant et constructif entre citoyens, décideurs et chercheurs ou ingénieurs. Dans un monde économique et géopolitique en profonde transformation, la science est notre atout : elle est l'assurance-vie de nos démocraties, le moteur de la croissance, notre principal outil pour répondre aux grands défis du siècle (énergie, climat, santé, défense…) et la garantie de notre souveraineté pour protéger les générations futures. Et, au-delà de l'assurance de progrès, elle permet de décrire le réel, d'approcher la notion de vérité de plus en plus malmenée ; elle fournit des outils pour la prise de décision et l'anticipation en favorisant la délibération et le débat éclairé.

Désinformation : de quoi parle-t-on ? Un défi conceptuel

La juste information de chacun est la condition de l'émancipation de tous et du bien commun. En miroir, la désinformation est un défi majeur pour nos démocraties. Pourtant aujourd'hui, du côté de la science et de l'action publique, **les outils d'analyse et de régulation sont instables.**

Sommes-nous en train de perdre la main, malgré l'urgence d'agir ?

La littérature révèle de fait un constat clair : aujourd'hui, ni les termes disponibles pour décrire ce qui fait problème, ni les modèles disponibles pour en comprendre la diffusion, ne permettent de partager un cadre de pensée commun efficace.

Dans ce rapport d'Evidences, la politiste **Mélanie Heard** examine les limites des modèles actuels de compréhension et de lutte contre la désinformation et apporte de nouveaux éléments à la réflexion, venus de la littérature internationale.

Principaux constats

La désinformation est un défi majeur pour nos démocraties.

• Un flou lexical : Dés-, més-, mal-information, fake news : les catégories dominantes sont floues et incomplètes. Elles ne captent ni la diversité des contenus problématiques ni la mécanique complexe de leur circulation (témoignages invérifiables, tromperie contextuelle +/- délibérée, informations intègres mais partielles et équivoques, satire gratuite, etc.-…le tout à l'aune des intérêts bien compris des plateformes). Cette imprécision dans les termes fragilise l'action publique.

• Des limites épistémologiques : La désinformation met à l'épreuve la méthode scientifique : contenus éphémères, taxonomies mouvantes, accès partiel aux données, opacité des algorithmes, impossibilité d'audit complet des systèmes numériques.

• Des normes instables : Les critères mobilisés pour qualifier le caractère problématique d'un contenu évoluent : faut-il chercher la véracité ? l'authenticité ? l'intégrité ?

• Des modélisations concurrentes : Toute connaissance, a fortiori placée au service de l'action publique, suppose un cadre d'interprétation : des modèles ou des analogies. Ici les paradigmes disciplinaires multiplient les modèles : contagion (infodémie), paresse cognitive, hiérarchie raison/émotion…Aucun modèle ne rend pleinement compte de la complexité systémique du phénomène, et la concurrence des modèles entrave la construction d'un référentiel partagé.

Ce nouveau modèle révèle combien la désinformation est simultanément un tort qui est fait aux foules et un tort que les foules se font à elles-mêmes. => Et si le sujet de l'action publique concernait surtout le désir qu'ont les gens de participer directement aux récits - petits et grands - qui les concernent ?

• Un enjeu démocratique central : La lutte contre la désinformation ne peut se réduire à la régulation, à la correction des contenus ou à la restriction des accès. Elle implique une stratégie proactive : produire des récits fiables, attractifs et participatifs, capables de restaurer la confiance et de renforcer l'autonomie citoyenne. La désinformation est une performance politique et sociale inscrite dans les architectures numériques qui encouragent l'engagement : l'action publique doit répondre à ce désir de participation en permettant elle aussi à chacun de monter sur la scène d'une juste information émancipatrice.

Recommandations stratégiques

1 **Clarifier les concepts** : élaborer un lexique robuste et partagé pour éviter les confusions normatives et opérationnelles.

2 **Articuler les modèles** : combiner approches cognitives, systémiques et participatives dans une perspective intégrée.

3 **Investir dans l'intégrité des systèmes** : développer des mécanismes effectifs d'audit, de transparence algorithmique et de mitigation des risques.

4 **Renforcer l'autonomie cognitive** : promouvoir la littératie numérique, des dispositifs attentionnels adaptés et des outils collaboratifs.

5 **Occuper l'arène informationnelle** : réinventer la scène publique en créant des espaces narratifs attractifs pour le « vrai », capables de rivaliser avec la scène des récits fallacieux.

La désinformation ne tient pas seulement à la diffusion de contenus trompeurs : elle relève d'un phénomène systémique, participatif et adaptatif. Y répondre suppose de dépasser une logique défensive pour investir dans des compétences techniques publiques et des récits démocratiques capables de redonner à l'espace informationnel juste sa fonction émancipatrice. Il est temps aussi d'interroger nos logiques implicites de répression de l'émotion : la diffusion de la mésinformation joue sur certaines modalités de plaisir et de connivence chez les utilisateurs qu'il s'agit, non pas de blâmer comme autant d'avilissements émotionnels, mais au contraire de rendre compatibles avec un exercice ouvert et participatif de la pensée critique émancipatrice.
L'enjeu politique n'est pas seulement de combattre la désinformation, mais de reconnaître que l'environnement informationnel du XXIe siècle constitue déjà de fait un espace de participation démocratique : aujourd'hui dévoyé, mais potentiellement transformable au service du bien public. L'accès aux Lumières se joue aussi désormais dans l'accès à la lumière des scènes où se racontent les petits et les grands récits collectifs. A nous d'y combattre les voix malignes, mais aussi d'y assurer la justesse des voix profanes qui désirent s'y faire entendre.

Mélanie Heard
Mélanie Heard, Déléguée générale d'Evidences
Docteure en pensée politique, normalienne formée en philosophie, Mélanie Heard conduit depuis plusieurs années, à Terra Nova puis à Evidences, un travail sur le lien entre connaissance et action publique. Ce rapport est le fruit d'une analyse de la littérature mêlant sources théoriques (en philosophie et épistémologie) et données scientifiques en construction.

Le rapport complet
30 pages · 2026
Le rapport esquisse une revue des réflexions contemporaines sur les enjeux epistémiques et politiques de l'espace informationnel et de ses pathologies actuelles. A partir d'une lecture critique du cadre scientifique et normatif en construction pour décrire et combattre les mécanismes et les effets de la désinformation, il propose des pistes de réflexion politique conçues pour nourrir la conversation publique sur les moyens de combattre ces périls qui menacent aujourd'hui notre idéal démocratique d'émancipation.

Introduction

La circulation en ligne d'énoncés faux ou fallacieux, trompeurs, erronés, mensongers, légers, biaisés, partiels, partiaux, équivoques, suggestifs, insincères, malveillants, agressifs, ou validant des croyances ou des comportements dangereux…constitue une préoccupation majeure pour nos démocraties. Un phénomène couramment désigné comme un problème d'« information » lié à un flux contraire et croissant en ligne de « dés-, més-, ou mal- » information, de « fausses nouvelles », « faits alternatifs », « rumeurs », etc. dont les réseaux sociaux sont le milieu naturel d'expansion croissante, incontrôlée, démesurée. Chercheurs et acteurs publics se mobilisent à l'échelle nationale, européenne et internationale pour comprendre ce phénomène et le combattre. Derrière l'accumulation lexicale du vocabulaire français, l'anglais qui préside aux normes scientifiques et politiques en la matière résume le problème d'un terme à la sémantique riche : le vivre-ensemble est en danger quand l'information qui circule en ligne est « *misleading* ».

De quoi parle-t-on ? De fait, la question lexicale révèle les tensions qui existent entre recherche et régulation. Alors que les sciences proposent des modèles complexes, non linéaires et multidimensionnels, peu accessibles aux profanes, les instruments politiques simplifient fortement les catégories opérationnelles qu'ils en infèrent. Cette simplification ne relève toutefois pas uniquement d'un appauvrissement conceptuel, mais aussi d'une exigence d'intelligibilité et d'explicabilité. Là où la production scientifique vise avant tout la compréhension par des communautés spécialisées, les instruments d'action publique doivent être compréhensibles, discutables et appropriables par l'ensemble des citoyens. La tension ne se situe donc pas seulement entre complexité scientifique et simplification politique, mais entre différents régimes d'explicabilité, chacun répondant à des finalités et des publics distincts. Cette réduction est ainsi moins le symptôme d'une ignorance de la complexité que le produit d'une contrainte structurelle : gouverner suppose de rendre lisible ce qui ne l'est pas spontanément.

Les plateformes privées, quant à elles, utilisent des taxonomies minimales facilitant la modération algorithmique (EDMO, 2024). Le Conseil de l'Europe a, dès 2017, montré dans un rapport (Wardle & Derakhshan, 2017) combien les mots et catégories utilisés pour décrire les phénomènes de désinformation étaient susceptibles de créer des effets cognitifs (sur la formulation du diagnostic) aussi bien que normatifs (sur la formulation des valeurs à défendre) qui peuvent fragiliser à la fois la recherche et l'action.

Car la science politique - en particulier française grâce à l'héritage fondateur de la sociologie des organisations et à l'analyse des politiques publiques, a certes bien démontré que l'idée d'une séquence politique pure allant du diagnostic à la cure n'est en fait qu'une fiction en matière de gestion des problèmes publics ; pour autant, on sait également qu'il ne peut y avoir de mise à l'agenda politique d'un problème public que lorsqu'arrivent à maturité simultanée son **référentiel cognitif** et son **référentiel normatif**, c'est-à-dire d'une part le cadrage descriptif de l'enjeu *(quel est le problème, quelle amplitude, quels critères de mesure et quelles hypothèses en termes de causes et de conséquences ?)* et d'autre part son cadrage au plan des valeurs que la réponse entend protéger *(quelles sont les valeurs mises en cause par le problème et quels principes d'action vont être engagés dans la réponse, pour qui, pour quoi, pour quelle vision du monde* ?).

C'est pourquoi lors d'une **réunion des parties prenantes le 28 octobre 2025** à l'**Elysée**, le président de la République Emmanuel Macron a souligné certes la richesse des énergies mobilisées dans le combat contre la désinformation, mais aussi **l'hétérogénéité** de leurs motivations et de leurs grilles de lecture - « *depuis les menaces qui pèsent sur le cerveau de nos enfants jusqu'aux enjeux géopolitiques* ». Une hétérogénéité majeure décrite aussi tout récemment par exemple par Grégoire Darcy et les chercheurs en sciences cognitives de l'Ecole normale supérieure, qui déplorent « *une absence de cadre d'analyse intégratif, articulant les dynamiques technologiques, géopolitiques et cognitives* ». Or cette hétérogénéité constitue à ce jour un problème politique : pour mieux « *résister* » à la désinformation, le président de la République considère qu'il nous reste à penser le « *cadre de pensée commun* », le creuset cognitif et normatif global qui servira toutes les problématiques concernées et mobilisera toutes les énergies disponibles.

Evidences, le think tank qui défend la démarche scientifique en politique, salue d'abord ici la méthode proposée par l'exécutif. **Le préalable à toute action publique efficace et *evidence-based* est en effet de prendre le temps de construire collectivement un référentiel politique clair et partagé, fondé sur la mise en commun des connaissances disponibles** : c'est la condition pour être **capable d'expliciter avec précision ses raisons, ses valeurs et ses finalités.** D'ailleurs la méthode scientifique nous l'enseigne : on ne comprend que ce que l'on décrit bien, grâce à des hypothèses et à des taxonomies claires et justifiées. Quelle que soit l'urgence à construire les outils qui permettront de répondre à un problème public, il est fondamental de prendre ensemble le temps de dire ce qui fait problème (quels faits ? sur quels critères, quels indicateurs de mesure, quelles données ?) et pourquoi (quelles valeurs sont à défendre, quels idéaux nous guident ?) ; cette démarche est probablement un préalable à l'efficacité de la réponse politique.
Il apparaît dès lors nécessaire de reposer conjointement ces questions fondamentales, nous dit le président de la République, dans la mesure où les actions engagées jusqu'à présent pour lutter contre la désinformation et encadrer la qualité de l'information en ligne — bien que réelles et non négligeables — se révèlent manifestement insuffisantes. Cette insuffisance est d'ailleurs explicitement reconnue par le Président lui-même, qui admet une perte progressive de maîtrise des dynamiques informationnelles : « *nous sommes en train de perdre la main* ».
Il n'est certes plus temps de spéculer ; mais l'urgence à agir, qui crée logiquement des attentes majeures du côté des outils d'intervention et de régulation, ne dispense donc pas de prendre le temps d'une réflexion moins directement opérationnelle, mais cruciale. Sommes-nous tous d'accord sur la juste façon de désigner et mesurer le problème ? Appliquons-nous tous, du pédiatre au politiste ou de l'éducateur à l'élu, les mêmes critères pour discriminer les énoncés dangereux en ligne ? Disposons-nous des mots, des concepts et des critères raisonnés de discernement qui peuvent fédérer nos énergies autour d'un même combat pour défendre un même idéal ?
Il peut sembler paradoxal ou provocant de le dire, mais la réponse est probablement : Non.

1. Le problème du lexique et de la taxonomie

Les catégories employées par les institutions et par les plateformes (essentiellement : *misinformation, disinformation, malinformation, fake news*…) sont trop pauvres pour capturer la complexité du phénomène. D'une part, comme le retient par exemple le Conseil de l'Europe avec Wardle & Derakhshan (2017), parce qu'elles sont souvent **floues et incohérentes**. D'autre part, parce qu'elles ne capturent **pas de façon exhaustive** le phénomène et conduisent à laisser de côté des pans entiers d'énoncés problématiques.

1.1. Information

Hugo Mercier et Dan Sperber ont montré dans leur livre *L'Enigme de la raison* qu'il est crucial de donner à la notion d'information la densité d'un élément essentiel de notre humanité et de nos facultés de raisonnement : « *Les humains se distinguent des autres animaux par la richesse et l'amplitude des informations qu'ils communiquent entre eux et par le degré auquel ils dépendent de cette communication. Pour devenir un adulte compétent, chacun de nous a dû énormément apprendre des autres. Nos savoir-faire et nos connaissances générales doivent moins à notre expérience individuelle qu'à la transmission sociale. La plupart de nos activités quotidiennes, en famille, au travail, en amour ou dans nos divertissements, dépendent massivement de* ***ce que nous avons appris d'autrui****. Les avantages énormes, indispensables, que nous tirons de la communication vont de pair avec une* ***extrême vulnérabilité à la désinformation*** » (Mercier & Sperber, 2021).
Essentielle donc à toute vie en société, la circulation de l'information est singulièrement cruciale en **démocratie**. Elle est d'abord **un droit** reconnu par la Déclaration universelle

des droits de l'homme en son article 19 : le droit, désigné comme corollaire nécessaire de la liberté d'expression et d'opinion, « *de chercher, de recevoir et de répandre (…) les informations et les idées* ». Elle est aussi un moyen crucial au service de la qualité du débat public, sans lequel il est impossible à une communauté de s'emparer librement et démocratiquement de son destin.

On trouve par exemple chez Kant dans *Qu'est-ce que les Lumières ?* un socle clair en ce sens, auquel Emmanuel Macron se réfère parfois. Dans l'état que Kant appelle de « *minorité* » et d'assujettissement obscurantiste dans l'Ancien régime, les hommes indexent leur existence à une tutelle univoque : « *Il est si commode d'être mineur. Si j'ai un livre pour me tenir lieu d'entendement, un directeur pour ma conscience, un médecin pour mon régime... je n'ai pas besoin de me fatiguer moi-même* ». A l'inverse, la sortie de cet état de minorité, cette émancipation qui est tout l'objet des Lumières, passe immanquablement par la circulation libre des idées. Ce que produit la liberté, c'est la possibilité pour chacun d'un « *usage public de la raison* » qui « *répand les lumières parmi les hommes* » parce que chacun devient alors à même de « *communiquer au public ses réflexions* ***soigneusement pesées et bien intentionnées*** ». Et cette libre communication des raisons est au fondement du libéralisme politique : « *Quand la nature a ainsi fait éclore, sous la dure carapace, le germe dont elle prend soin le plus tendrement, à savoir l'inclinaison et la vocation pour la pensée libre, cette tendance influe peu à peu en retour sur la mentalité du peuple (ce qui le rend progressivement plus apte à agir librement) et, finalement, sur les principes mê mes du gouvernement, lequel juge profitable pour lui-mê me de* ***traiter l'homme, qui dès lors est plus qu'une simple machine, conformément à sa dignité*** » (Kant. 1784).

1.2. Quel est le contraire de l'information ?

Le terme « information » est-il nécessairement connoté positivement au sens où il contiendrait logiquement l'idée de véracité et de connaissance vraie, facteur d'émancipation ? Les énoncés « informatifs » sont-ils strictement limités à ceux qui édifient la raison ?
Ce n'est pas le cas dans l'usage commun : le terme *information* désigne aujourd'hui tout contenu - données, énoncés, images, signaux - transmis, reçu ou manipulé, *sans distinction de nature ni de véracité*. Cette polysémie, apparue avec les médias de masse et les dispositifs de communication modernes, conjugue imprécision conceptuelle et efficacité pratique : elle permet de nommer globalement ce qui « circule », sans en spécifier la nature. Mais le bénéfice de cette souplesse a un coût : le mot s'emploie couramment de manière contradictoire : le même locuteur peut tantôt considérer qu'en disant « j'ai reçu l'information que x vient » il signale qu'il pense qu'on peut tenir pour vrai que x va venir (considérant donc ici implicitement que le terme « information » est synonyme de véracité), et tantôt considérer à l'inverse que l'énoncé « ce témoin a livré des informations contradictoires » est lui aussi cohérent, alors qu'il suppose cette fois que chaque information exige du discernement puisque certaines sont nécessairement fausses. Cette coexistence de sens incompatibles ne trouble pourtant guère la communication, le contexte pragmatique suffisant de fait à lever les ambiguïtés.
Reste que le dictionnaire n'offre pas d'antonyme au mot "information" (source : CNRTL). La théorie informatique définira l'information comme une réduction de l'imprévisibilité du signal et lui opposera la notion de « bruit » au sens d'un brouillage de la transmission accroissant l'incertitude. Héritée des travaux fondateurs de la théorie de l'information, cette approche conçoit l'information comme une grandeur statistique abstraite, indépendante de toute considération sémantique, normative ou intentionnelle. Elle vise avant tout à modéliser, mesurer et optimiser la transmission de messages dans des canaux potentiellement bruités, sans se prononcer sur leur signification, leur véracité ou leurs effets cognitifs et sociaux.
Cette neutralité formelle, qui constitue l'un des principaux atouts de ces modèles dans le champ de l'ingénierie et de l'informatique théorique, en marque également les limites dès lors qu'il s'agit de penser les usages sociaux de l'information. Elle permet de traiter indistinctement des messages factuellement exacts, trompeurs ou mensongers, pourvu

qu'ils soient transmissibles et exploitables du point de vue statistique.
L'épistémologie sociale, quant à elle, appréhende l'information en tension avec les notions d'ignorance, d'asymétrie cognitive ou de « désordre informationnel », mettant l'accent sur les conditions sociales de production, de circulation et d'appropriation des savoirs. Dans d'autres contextes encore, notamment philosophiques ou politiques, les concepts de tromperie, de pseudo-information, d'opacité ou de manipulation sont mobilisés afin de qualifier des usages intentionnels de l'information et leurs effets sur les croyances individuelles et collectives.
Ce décalage conceptuel se retrouve dans les dispositifs numériques contemporains, largement fondés sur des approches algorithmiques héritées de ces cadres formels. Les systèmes informatiques qui organisent aujourd'hui la circulation de l'information — qu'il s'agisse de moteurs de recherche, de systèmes de recommandation ou de modération automatisée — sont principalement optimisés pour gérer des flux en termes de volume, de vitesse, d'attention ou de prédictibilité, bien plus que pour évaluer la qualité, la fiabilité ou la portée normative des contenus diffusés.
C'est dans ce contexte que l'usage courant et les instruments d'action publique mobilisent désormais des catégories explicitement normatives telles que « désinformation », « mésinformation » ou « malinformation », mais aussi « fake news » ou « faits alternatifs ». Ces catégories visent à rendre opératoire, dans l'urgence de l'action, une distinction entre différents régimes de circulation de l'information. Elles révèlent toutefois un écart persistant entre des cadres théoriques qui permettent de penser l'information sans le vrai, sans le faux et sans l'intention, et d'autre part des phénomènes sociaux qui exigent au contraire précisément de se prononcer sur ces dimensions.
Mais dans tout ceci, donc, **pas de concept clair de ce que serait en propre l'*anti-information***. Quel est le problème politique à combattre, en quoi ce qui se produit sous nos yeux est-il contraire à ce qui devrait-être et à ce que nous devons défendre ? Est-ce le faux, en propre, ou bien seulement le partiellement-vrai qui prête à confusion ? Est-ce la perte de repères collectifs permettant de discerner le faux, est-ce la malveillance qui préside à sa diffusion, ou bien la simple légèreté qui le propage sans même y croire, pour rire ? Les questions ici pour partitionner précisément ce qui *fait problème* à l'aune de l'intérêt général ne sont pas aussi clairement cartographiées qu'on le voudrait.
On peut toutefois retenir ici avec la philosophe **Mona Simion** de l'Université de Glasgow **l'importance de l'antonymie entre « information » et « ignorance ».** L'information, décrite comme un processus, est fonction de son résultat épistémique : quand la capacité fonctionnelle de s'informer est exercée à bon escient, avec un résultat positif, elle produit de la connaissance - alors que quand elle est exercée à front renversé, avec un résultat négatif, elle produit de l'ignorance. Le problème qui nous occupe, en ce sens, tient au fait que, quand se produit *l'inverse* de l'information, **c'est notre statut d'agent capable de connaissance qui est défait**. Ainsi, s'intéresser à l'information et à son contraire revient à s'intéresser à « *des choses qui ont la capacité de générer ou d'augmenter l'ignorance - c'est-à-dire de **dépouiller totalement ou partiellement une personne de son statut de connaissant**, de bloquer son accès à la connaissance, ou de réduire sa proximité au savoir* » (Simion, 2024).

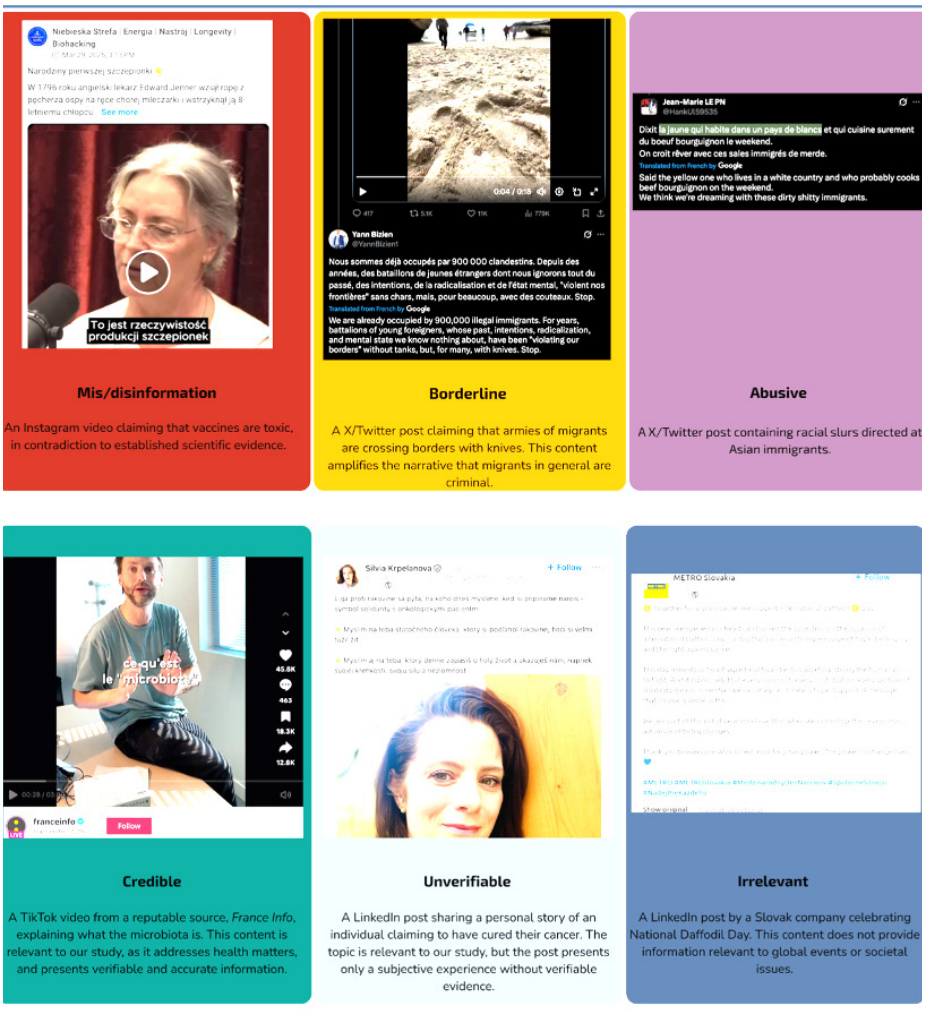
1.3. Désinformation et Mésinformation

Mais le terme "désinformation" convient-il vraiment dans cette entreprise ? La réponse de la littérature scientifique à ce jour est, là encore, massivement : Non.
Comme le remarquait par exemple Tim Hayward en introduction d'un article intitulé « The Problem of Disinformation » publié dans la revue *Social Epistemology* en 2025, « *le terme « désinformation » est utilisé de diverses manières, et bien qu'il soit normalement compris comme présentant un problème, il existe des points de vue contrastés sur la nature de ce problème* ». Le cadre lexical qui fait référence est fondé sur la distinction entre désinformation et mésinformation. Ainsi, la Commission européenne a fixé dans la communication relative au European Democracy Action Plan (EDAP) de décembre 2020 qu'« *il importe d'opérer une distinction entre différents phénomènes communément appelés « désinformation » afin de permettre l'élaboration de mesures appropriées* », et elle propose les définitions de référence suivantes :

« *on entend par «**mésinformation**» des contenus faux ou trompeurs transmis sans intention de nuire, même si leurs effets peuvent néanmoins être préjudiciables; c'est notamment le cas lorsque des personnes partagent de bonne foi de fausses informations avec des amis ou des membres de leur famille;*
*on entend par «**désinformation**» des contenus faux ou trompeurs diffusés avec l'intention de tromper ou dans un but lucratif ou politique et susceptibles de causer un préjudice public* ».
Ce cadre lexical distingue donc la *dés*information de la *més*information sur un **critère d'intentionnalité** de tromper ou de nuire. Les erreurs innocentes (on les dira "authentiques") n'y sont pas considérées comme de la désinformation, mais elles sont tout de même jugées suffisamment problématiques pour justifier l'usage d'un terme spécifique : la mésinformation.
Cette distinction lexicale importante, sur un critère d'intentionnalité, est aujourd'hui largement diffusée, mais il faut bien reconnaître aussi qu'une large partie de la littérature, tant chez les chercheurs que chez les régulateurs, choisit encore de passer outre et emploie en fait le terme « désinformation » de façon totalisante : pour qualifier tout énoncé qui fait problème au sens où il est, volontairement ou non, « *misleading* ». Référence prééminent, le Code of Practice on Disinformation que la Commission européenne a intégré en février 2025 au Digital Services Act choisit clairement cette acception « parapluie » qui englobe largement : la désinformation au sens strict (définition EDAP), mais aussi la mésinformation, les « information influence operation » et les « foreign interference in the information space ». En pratique, pour la Commission le mot est donc très large : « *la désinformation nuit à notre société en* érodant la confiance dans les institutions et les médias, en mettant en danger les processus électoraux, en entravant la capacité des citoyens à prendre des décisions éclairées, en portant atteinte à la liberté d'expression ».
Pourquoi le critère d'intentionnalité de la nuisance, qui fonde la distinction lexicale entre désinformation et mésinformation, semble-t-il de moins en moins pertinent aujourd'hui ? C'est que, comme le notait le Parlement britannique dans une note du Parliamentary Office of Science and Technology en 2024, « *Il n'est pas toujours possible de faire la distinction entre désinformation et mésinformation. Un contenu peut être créé à des fins de désinformation et être partagé à l'insu de tous comme mésinformation* ».
Surtout, cette opposition binaire rend mal compte de stratégies informationnelles dont l'objectif n'est plus tant de mentir que de produire de la confusion, de saturer l'espace informationnel ou d'instiller un doute systématique, sapant toute possibilité de débat public et de décision fondée sur des faits partagés. En réalité, rappellent les auteurs du rapport sur la désinformation de l'Institut Jean Nicod de l'Ecole normale supérieure (Darcy et al. 2025) le mensonge est rare, coûteux et souvent inutile : « *la désinformation n'est que la partie émergée de l'iceberg : la plus visible, la plus facile à désigner, mais ni la plus répandue, ni la plus dangereuse. Les manipulations les plus fréquentes reposent moins sur des contre-vérités que sur des procédés bien plus subtils* ».
Ainsi, Caroline Jack, chercheuse américaine spécialisée dans la désinformation et les médias numériques, affiliée à l'institut de recherche Data & Society**,** a proposé dès 2017 dans son très remarqué *Lexicon of Lies* une typologie de ces tactiques de disruption informationnelle, qui ne relèvent ni du mensonge ni de la persuasion classique sur le modèle familier de la propagande, mais plutôt d'une pollution délibérée et diffuse de tout l'écosystème communicationnel. L'exemple paradigmatique souvent cité en est le *xuanchuan* pratiqué en Chine par « l›armée des cinquante centimes », dont la fonction assumée n'est pas de défendre le régime par l'argumentation, mais « simplement » d'inonder les espaces de discussion de contenus anodins ou hors-sujet, détournant ainsi l'attention des sujets sensibles et saturant les capacités d'engagement critique des publics.
Le besoin de rendre compte de ce **floutage des responsabilités et des intentions entre manipulateurs et manipulés** provoque ce que certains commentateurs décrivent aujourd'hui comme **une mutation profonde du cadre d'analyse** : le lexique de référence, qui donne un rôle clé dans ses termes fondamentaux à la notion d'intentionnalité de nuire et repose sur un modèle émetteur/récepteur, n'est vraiment plus bien articulé aux caractéristiques dont il s'agit de rendre compte.

1.4. Des catégories insuffisantes au regard du réel à décrire

Une autre critique que l'on peut formuler contre la taxonomie désinformation/mésinformation, c'est qu'en réalité elle est loin de couvrir l'ensemble des natures d'énoncés qui posent problème. Citons ici par exemple une toute récente et importante publication scientifique, issue du projet européen SIMODS fédérant huit organisations de fact-checking et de recherche de premier plan pour mesurer la désinformation sur les principales plateformes : à la question basique « que mesurer ? », la réponse n'est pas aussi claire qu'on voudrait le penser, comme le montre la figure ci-dessous.



Captures d'écran d'exemples de publications annotées comme appartenant à chacune des catégories principales.

Dans cette figure, les auteurs de l'étude retiennent sous la catégorie qu'ils appellent « dis/misinformation » **un énoncé caractérisé à la fois par sa portée générale assumée, par sa fausseté manifeste, et par une intention délibérée de nuire** : « les vaccins sont toxiques ». Ils excluent en revanche de cette catégorie d'autres contenus qu'ils disent seulement « borderline », « abusifs » ou « invérifiables » - même s'ils en retiennent finalement aussi le caractère « problématique ». La catégorie « dis/misinformation » est donc clairement insuffisante. Mais quel lexique choisir pour classer l'ensemble des contenus problématiques ?

La catégorie dite « invérifiable » pose par exemple de sérieux problèmes. L'étude classe l'énoncé « *J'ai guéri du cancer grâce à…* » comme « invérifiable », et non comme « dés/mésinformation », au motif qu'il s'agit d'un témoignage personnel. Or on sait bien cependant qu'un post qui ferait de cette façon la promotion captieuse du charlatanisme peut fort bien, s'il circule massivement, finir par nuire gravement à de nombreux malades en les détournant des traitements éprouvés et de la confiance dans la science.
L'équipe de Kate Starbird à l'Université de Washington a ainsi montré notamment que c'est à partir d'un post énoncé sous la forme d'un simple « témoignage » personnel et « authentique » que la rumeur selon laquelle les migrants mangeraient chats et chiens s'est répandue aux Etats-Unis, jusqu'à devenir un marqueur idéologique de la campagne présidentielle en 2024. Bien avant que le candidat Trump n'y fasse référence, la rumeur commence avec le « témoignage » sur Facebook d'une résidente de Springfield, repris au sein de groupes Facebook locaux avant qu'un leader néonazi local en parle lors d'une réunion publique enregistrée : la vidéo est alors propulsée par des influenceurs, largement relayée sur des applications comme Telegram ou Reddit et devient peu à peu un nouveau terrain de jeu politique, comportant un élément narratif inédit, pour des communautés anti-immigration. Des mèmes, souvent produits par IA sur le ton de l'humour et mettant en scène le candidat Trump protégeant chiens et chats, se diffusent alors rapidement, attirant l'attention de responsables politiques comme Ted Cruz et Marjorie Taylor Greene, qui les relaient en les reliant à leur récit politique plus large sur la nécessité de stopper l'immigration à la frontière mexicaine. C'est alors que le candidat président va jusqu'à dire « ***they're eating the pets*** » lors de son débat électoral avec Kamala Harris. Cet exemple désormais bien connu est utilisé par Kate Starbird pour illustrer la mécanique des flux collaboratifs et quasi-improvisés qui, relayés par des influenceurs aux intentions idéologiques pernicieuses assumées, peuvent transformer un simple « témoignage personnel » ambigu en marqueur symbolique d'une « désinformation » à très grande échelle (Tomson & Starbird, 2024). Dès lors, où situer le critère qui permet de discerner quels énoncés font problème ? L'idée qu'il faille se contenter de classer certains énoncés captieux dans la catégorie des « invérifiables » parce qu'ils sont subjectifs pose manifestement une série d'apories : est-ce alors la reprise massive, qui, sur un critère de volume (mais avec quel seuil ?), fera d'un énoncé qui, en lui-même, n'était qu'« invérifiable », une « désinformation » dangereuse ? Ou bien est-ce, en bout de chaîne, et quelle que soit la nature de l'énoncé de départ, le fait que des croyances et des comportements objectivables et dangereux en soient le résultat tangible ?
Autre trou lexical dans la raquette de la taxonomie disponible : la notion d'**information *insuffisante* et *décontextualisée***. C'est un point sur lequel l'OCDE insiste, posant depuis 2022 (Lesher, Pawelec & Desai, 2022) le constat que les catégories « désinformation » et « mésinformation » ne couvrent pas l'ensemble des énoncés problématiques. La notion de « *contextual deception* » ou **tromperie contextuelle** est donc avancée en complément ; elle désigne « *l'usage d'informations vraies, mais pas nécessairement pertinentes, pour* ***cadrer*** un événement, un sujet ou une personne (par exemple, un titre qui ne correspond pas à l'article associé), ou la ***déformation*** de faits afin de soutenir un récit (par exemple, *supprimer délibérément une information constituant un élément de contexte indispensable à la compréhension du sens initial)* ». Bien que les faits mobilisés soient exacts (contrairement à la désinformation) et non fabriqués (contrairement à la mésinformation), la manière dont ils sont utilisés est fallacieuse et procède d'une intention de manipuler le public ou de nuire. **Ce type de pratiques met en évidence les limites d'une approche centrée sur la véracité factuelle des énoncés, en montrant que l'effet trompeur peut résider moins dans ce qui est dit que dans ce qui est tu, déplacé ou mis en saillance.** Il souligne ainsi la nécessité de penser l'information non seulement en termes de contenu, mais aussi de **cadre, de sélection, de hiérarchisation et de mise en contexte**, dimensions largement invisibilisées par les taxonomies dominantes. Cette catégorie est d'ailleurs souvent l'apanage des médias plus traditionnels, participant ainsi à la polarisation des débats faute d'espace commun de compréhension du réel.
Enfin, l'OCDE ajoute encore une autre catégorie de contenus que les concepts de désinformation et de mésinformation ne parviennent pas à recouvrir : il s'agit de la **satire**. Elle se définit par le recours à l'humour et à l'exagération pour critiquer des personnes ou des idées, souvent sous la forme d'un commentaire social ou politique ; l'OCDE

souligne que la spécificité de la satire est que son statut humoristique, établi en principe par le contexte de sa production initiale, s'efface aux yeux du public à mesure que le contenu est partagé et repartagé, ce qui favorise l'ignorance : « *le lien contextuel disparaît parfois, intentionnellement ou non, du fait de la personne qui le diffuse, conduisant de nouveaux publics à mal interpréter le sens initial* ».

Au terme de cette analyse, force est de considérer que le lexique disponible dans la littérature tant académique qu'institutionnelle produit une impression de flottement conceptuel, qui brouille ne serait-ce que l'étape basique de la description du problème. Chaque nouvelle catégorie produite semble mettre en lumière dans le problème décrit un trait caractéristique que les autres catégories manquent.

Absence de véracité, intention maligne, cadrage contextuel, quantité suffisante d'informations, présence d'une instrumentalisation manipulatrice, mais aussi ampleur du flux de diffusion, reprise politique, caractère dangereux des croyances et comportements engendrés… : le constat est celui d'une grande difficulté à combiner tous ces traits et à conceptualiser de manière cohérente ce qui au juste constitue le problème de la désinformation. On repère au total qu'au moins **trois conceptions du problème que pose la désinformation se retrouvent couramment dans les usages comme dans la littérature académique : l'information trompeuse** (l'énoncé, faux, partiel ou décontextualisé, est *fallacieux)***, la tromperie intentionnelle** (il est *intentionnellement* diffusé en tant que tel) **et les conséquences dommageables de la tromperie** (la *circulation* de cet énoncé comporte un *risque* de *dommages* sur les croyances et/ou les comportements des récepteurs)**.** Bien que ces trois dimensions puissent parfois être réunies dans un même énoncé, il s'agit de trois types de problèmes qui sont distincts, chacun valant problème à lui seul.

Sur cette base, l'ensemble de la littérature révèle finalement une variété de cadrages et de présupposés différents, chacun décrivant le même problème général sous des formulations distinctes qui privilégient l'un ou l'autre de ces trois traits caractéristiques. **Le constat est dès lors celui d'une réelle imprécision des termes du problème de la désinformation. Les catégories utilisées tant par l'action publique que par la recherche 1) sont floues puisque leurs limites sont mouvantes et 2) ne capturent ni avec précision ni avec exhaustivité le problème qu'elles veulent désigner.**

1.5. « The practical rationale is false » (Simion, 2024)

En réponse au président de la République qui soulignait, on l'a dit, l'hétérogénéité des motivations qui guident les acteurs et celle de leurs grilles de lecture, on peut donc retenir à ce stade pour hypothèse que cette hétérogénéité tient peut-être en tout premier lieu à une imprécision du vocabulaire qui circule pour décrire ce dont il est question dans ce combat commun. Notons au passage qu'à cela s'ajoute en France, on l'aura également noté, le recours permanent à l'anglais qui semble bien inévitable - en dépit des efforts récents du ministère de la Culture à fournir un lexique de sa façon.

Ce problème de lexique est largement souligné dans la littérature. Le philosophe de l'information et épistémologue Don Fallis, professeur à l'University of Arizona, a notamment montré dans *What is Disinformation?* que les définitions existantes sont tantôt trop larges, tantôt trop étroites, et que cette indétermination conceptuelle menace notre capacité de recherche, de détection et d'action sur la désinformation et l'information. De même, dans *The importance of epistemology for the study of misinformation* (2024), le politologue de l'Université de Miami Joseph Uscinski, spécialiste du complotisme, déplore avec ses coauteurs que la recherche sur la désinformation travaille aujourd'hui « *without a reliable epistemology for determining the truth of the information and beliefs in question* », ce qui laisse les chercheurs s'en remettre à des définitions lâches et à une « *naïve epistemology* ».

Pourrait-on arguer malgré tout, en bons utilitaristes, que ces faiblesses conceptuelles importent peu au regard de l'urgence à agir ? Que, même si le lexique n'est pas conceptuellement parfait, il est suffisamment *utile* en pratique pour l'action publique, la détection algorithmique, la régulation des plateformes ? Mona Simion montre que

non. Parce qu'un concept mal formé n'est en fait jamais réellement utile en pratique. En réalité, une mauvaise définition ne sert pas les systèmes de détection ; elle produit des erreurs de classification ; elle compromet la recherche empirique ; elle crée de la confusion politique ; elle brouille la responsabilité normative.

Le flou des catégories est particulièrement dommageable parce qu'il freine la translation du savoir à l'action : que mesure-t-on dans les études en guise de « désinformation » problématique dont l'action publique devrait s'emparer ? Les critères retenus pour juger de ce qui fait problème sont-ils conformes aux responsabilités et aux finalités qui sont celles des acteurs publics ? Des grilles mouvantes ne peuvent qu'entraver la coopération nécessaire entre chercheurs et acteurs publics.

Pire, un lexique déficient peut aussi freiner, voire dissuader, l'adhésion du public en accréditant la connotation d'une censure arbitraire où les critères de repérage des contenus problématiques sont à la discrétion opaque du régulateur. Le soupçon que les mots masquent des biais politiques est ici un réel risque.

Enfin, la cartographie mouvante des concepts est d'autant plus surprenante et dommageable si l'on songe à l'intérêt que suscitent désormais de fait, chez les développeurs autant que les acteurs publics, les systèmes de détection algorithmique automatisée de la désinformation et de la mésinformation. Comment espérer coder ce que l'on ne pense pas encore très clairement ? Quel scoring produire concernant une fiabilité dont le périmètre reste à discerner ?

Au total, et comme le note Mona Simion, « *il est surprenant que les spécialistes de l'information, de la communication et de l'informatique n'aient pas consacré davantage d'efforts à établir une caractérisation correcte de la nature de la désinformation, alors même que les enjeux deviennent de plus en plus graves : propagation de contenus menaçant nos démocraties, affaiblissement de la confiance dans l'expertise, baisse de l'adhésion aux dispositifs de santé publique, et atteintes à la cohésion sociale. (…) Étant donné l'ampleur vertigineuse de notre dépendance épistémique — une dépendance que les technologies récentes n'ont fait que renforcer — pouvoir rendre compte de façon adéquate de la nature de la mésinformation et de la désinformation, afin de pouvoir les identifier de manière fiable et s'en prémunir, est devenu essentiel* ».

L'une des réponses possibles à cette difficulté est de s'intéresser à présent, non plus à la terminologie en elle-même, mais aux modèles qui permettent de penser le phénomène de la désinformation en tant que système. S'il faudra certes tôt ou tard s'accorder sur des pipelines de détection qui doivent, par construction, reposer sur des taxonomies fixes, avec des labels binaires et des critères objectivables, peut-être est-il toutefois nécessaire de poser d'abord la question de la façon dont on parvient à **modéliser la désinformation comme système**. C'est que la désinformation semble bien avoir en effet pour caractéristique de constituer un processus et un *système* adaptatif et mouvant dont la démarche taxonomique échoue à rendre compte parce qu'elle manque ses dimensions *dynamiques*. **Penser la désinformation comme un système implique alors de déplacer le regard, des contenus pris isolément vers les interactions, les boucles de rétroaction et les dynamiques d'amplification qui structurent leur circulation.** Une telle approche permet de rendre compte du rôle conjoint des architectures techniques, des incitations économiques, des pratiques sociales et des asymétries cognitives dans la production d'effets informationnels délétères. **Elle invite ainsi à dépasser une logique essentiellement classificatoire au profit de cadres analytiques capables de saisir l'évolution, l'adaptabilité et la résilience des écosystèmes de désinformation.**

2. Le problème des modèles

On sait que, si elle a certes besoin de taxonomies claires, **la connaissance progresse aussi en s'appuyant sur des modèles**, c'est-à-dire des schémas simplifiés ou des analogies qui permettent de décrire, d'expliquer et parfois de prédire un phénomène. Qu'il s'agisse de métaphores (épidémiologiques, économiques, militaires…) ou de formalismes plus explicites, les sciences ont besoin de ces représentations construites du réel. En ce sens, "bien décrire pour bien comprendre" signifie d'abord choisir *quel type de modèle* on mobilise, et donc *quels aspects* du phénomène on décide de rendre visibles ou, symétriquement, de laisser dans l'ombre.
Dans le cas de la désinformation, plusieurs modèles sont aujourd'hui en concurrence sans bien s'articuler.

2.1. Le modèle de la contagion

Symbolisé par le terme « infodémie » forgé par l'OMS, un premier ensemble de travaux pense la désinformation à partir de la métaphore de la **contagion** : un agent pathogène circule dans l'environnement informationnel en se transmettant de personne à personne via des vecteurs cognitifs et psychologiques, sur des supports techniques. On distinguera alors, comme en épidémiologie, la susceptibilité des agents récepteurs, la transmissibilité du pathogène, la fréquence et l'intensité de l'exposition, l'étendue du réservoir infectieux, la probabilité des contacts, la sévérité des symptômes, etc. Ce modèle a l'avantage de rendre saisissables la dynamique temporelle de la diffusion, de rendre compte de la désinformation en tant que cinétique liée à une pluralité de variables. Il aide notamment à penser la désinformation comme déterminée à la fois par la pathogénicité des énoncés et par la susceptibilité du public exposé. **Sur le plan technique, ce cadre a trouvé des traductions formelles dans des modèles issus de l'épidémiologie computationnelle et de la science des réseaux** (Chavalarias, 2023), où la diffusion de contenus est modélisée à l'aide de variantes des modèles SIR ou SIS, parfois enrichies de paramètres cognitifs ou attentionnels. Ces approches permettent de simuler des scénarios de propagation, d'estimer des seuils de diffusion, ou encore d'identifier des points de levier — individus, communautés ou moments critiques — sur lesquels une intervention pourrait avoir un effet disproportionné.
Le mérite de cette description est de fournir aussi une représentation pertinente de ce que peut être une politique de « prévention » par **immunisation**. C'est le sens des travaux sur l'**inoculation** psychologique qui font désormais référence et fondent une partie de la réponse publique : avec les méthodes dites de ***pre-bunking***, van der Linden et ses collègues par exemple proposent ainsi de « vacciner » les individus contre la désinformation en les exposant à de petites doses de procédés rhétoriques fallacieux qu'ils apprennent ainsi à reconnaître (van der Linden et al., 2017). L'idée que suggère à ce stade la littérature étant que cette prophylaxie pré-exposition, qui renforce les défenses cognitives du public, est plus efficace que le *debunking* curatif ex post (Heard 2025). **Ces stratégies s'accompagnent par ailleurs d'un intérêt croissant pour des outils numériques automatisés**, tels que des jeux sérieux, des interfaces interactives ou des modules intégrés aux plateformes, conçus pour exposer les utilisateurs à des schémas de manipulation typiques. Elles illustrent une tentative de traduction technique d'un objectif cognitif — renforcer la vigilance épistémique — dans des dispositifs scalables, compatibles avec les logiques industrielles des environnements numériques.
Un autre avantage significatif du modèle de la contagion, c'est qu'il rend non seulement bien compte de la propagation, mais aussi des **mécanismes d'isolement informationnel des publics** dans des réseaux polarisés typiquement représentés comme des « **clusters** » épidémiques. Ce phénomène combine trois notions, parfois différenciées, parfois amalgamées : les **echo chambers** (chambres d'écho où les individus sont exposés quasi exclusivement à des contenus conformes à leurs croyances), les **filter bubbles** (bulles de filtrage produites (en partie) par la personnalisation algorithmique et les clusters **homophily-driven** (regroupement spontané de personnes partageant des affinités cognitives,

politiques ou identitaires). **Ces phénomènes ont été largement étudiés à l'aide d'outils de graph mining et d'analyse des réseaux complexes**, qui permettent de caractériser la modularité des graphes sociaux, la densité intra-cluster, ou encore la faiblesse des liens inter-communautés. Du point de vue algorithmique, ils interrogent directement les effets de la recommandation personnalisée, de l'optimisation de l'engagement et des mécanismes de renforcement itératif, qui peuvent accentuer la segmentation de l'espace informationnel. Les préférences révélées d'un individu et son appartenance à des réseaux contribuent à renforcer le type de contenus auxquels chacun est exposé. Lorsque les utilisateurs interagissent de façon répétée avec certains contenus au sein de leurs réseaux sociaux, ou les partagent, et que ces contenus confortent leurs croyances, des chambres d'écho qui confirment leurs biais préexistants apparaissent et se développent (Karsten et West, 2016). Des travaux récents de modélisation de la diffusion de la mésinformation sur les plateformes de médias sociaux suggèrent en outre que les bulles de filtre peuvent effectivement contribuer à expliquer sa propagation (Acemoglu, Ozdaglar et Siderius, 2021). Ces phénomènes d'exposition, désormais bien décrits, sont très bien captés par le modèle épidémiologique ; il rend visible la polarisation problématique qu'ils engendrent en traitant la désinformation dans un **réseau de transmission fermé** où les flux de représentations circulent intensément au point de faire cluster. En outre, ce modèle permet aussi de rendre compte, comme dans certaines épidémies, de l'existence de « superspreaders » de désinformation, dont la recherche a bien montré l'importance - par exemple dans la diffusion sur Twitter/X de la propagande russe durant l'élection américaine de 2016, avec seulement 1% d'utilisateurs responsable de 70% des expositions aux énoncés fallacieux (Eady et al. 2023). **D'un point de vue technique, ces résultats ont nourri des travaux visant à détecter automatiquement ces nœuds à forte centralité**, en combinant mesures structurelles (degré, centralité d'intermédiarité) et signaux comportementaux. Ils alimentent également des débats normatifs et opérationnels sur le ciblage préférentiel de certaines interventions — modération, ralentissement de diffusion, ou démonétisation — qui posent toutefois d'importantes questions éthiques et politiques.

En dépit de ces atouts, on voit bien aussi que ce modèle tend à naturaliser la circulation de la désinformation comme un processus quasi biologique, subi par le public qui s'en trouve victime à la manière d'une fatalité. Pareille représentation risque de nous conduire à minimiser en partie les dimensions stratégiques et intentionnelles, à perdre de vue que la désinformation ne tombe pas du ciel mais dépend à la fois d'une ingénierie du chaos déterminée et pugnace, et d'une participation volontaire et active du public.

2.2. Le modèle de l'auto-support cognitif

La participation active du public à la désinformation qui l'affecte est un trait distinctif dont les modèles doivent s'attacher à rendre compte. C'est pourquoi un deuxième modèle se focalise sur la désinformation comme objet de consommation à risque, sorte de toxicomanie cognitive et compulsive qu'il s'agirait d'apprendre à réguler. On propose de l'appeler ici le **modèle de l'auto-support cognitif** : un peu comme, à la fin du XXe siècle, un consensus s'est formé autour de l'idée d'auto-support pour contrôler l'incidence du VIH chez les usagers de drogues : constatant que le modèle régulateur top-down et sécuritaire de la « guerre à la drogue » s'avérait inefficace voire contre-productif, la logique de l'auto-support s'est imposée pour favoriser l'*empowerment* des usagers et la mise à disposition d'outils adaptés pour les aider à réduire leurs risques.

De même, dans le modèle abordé maintenant, la consommation de dés/mésinformation est interprétée comme un risque épistémique, lié à des vices ou biais cognitifs, dont des mécanismes d'auto-support peuvent réduire l'impact en soutenant chez les usagers les vertus intellectuelles (esprit critique) qui permettent de s'en prémunir. La prémisse étant que le risque est là, impondérable : s'il est impossible de prévenir tout contact avec la substance à risque, il est possible d'en mitiger les dangers chez ceux qui la consomment.

La désinformation y est analysée comme une forme spécifique d'*erreur épistémique* dans la consommation de l'information : les énoncés erronés sont problématiques parce que les processus de traitement sont défaillants chez leurs consommateurs. On renonce peu

ou prou à empêcher tant l'exposition que la consommation : mais ce qu'il faut annuler, ce sont les dangers qui pourraient résulter de ce risque tenu pour impondérable. L'idée, comme dans les stratégies dites de réduction des risques face aux drogues, est que, si l'exposition au risque de consommer est de fait inévitable, au moins faut-il renforcer les capacités individuelles et/ou communautaires d'en réduire au maximum les dommages.
Une littérature importante en psychologie et en sciences cognitives soutient cette lecture, autour par exemple des concepts **de paresse cognitive ou de surcharge attentionnelle**. Ainsi, des travaux récents montrent que, contrairement à ce que l'on pourrait croire, ce n'est pas le partisanisme et le raisonnement politiquement motivé qui expliquent pourquoi les individus se laissent tromper par les informations fausses, mais des déficiences de traitement - qu'un support adapté permettrait d'améliorer en dépit de la surcharge tentatrice à laquelle les usagers sont inexorablement exposés.
La prémisse est ici que les usagers s'exposent volontiers au risque d'être trompés alors même qu'u fond ils sont conscients des dangers et prêts à s'en défendre si on les y aide. Des auteurs comme le canadien Gordon Pennycook et David Rand au MIT démontrent ainsi que la faible capacité à discerner le vrai du faux est davantage liée à un manque de raisonnement analytique et de connaissances pertinentes qu'à des parti-pris idéologiques ou identitaires (Pennycook & Rand, 2021a,b). La littérature en psychologie et sciences cognitives s'intéresse en particulier à la notion d'**heuristiques** erronées : celles qui s'attachent par exemple à la familiarité de la source davantage qu'à sa fiabilité perçue. Pennycook et Rand affinent avec précision l'analyse causale : ils montrent ainsi qu'en matière d'opinion politique, s'il est vrai que les partisans d'un même bord se croient entre eux, cela ne signifie pas pour autant que ce soit leur appartenance partisane qui cause l'adhésion à une croyance, car il existe en réalité un cofacteur puissant tenant aux faits et croyances antérieures déjà partagés entre gens du même bord. En isolant ce facteur du background commun d'informations déjà tenues pour vraies, les auteurs montrent que l'adoption d'une nouvelle croyance est moins un phénomène idéologique qu'un phénomène cognitif, qui consiste à choisir en guise d'heuristique sa conformité ou non avec le background de croyances antérieures.
Ces chercheurs insistent aussi sur le fait qu'il existe en outre un décalage important entre ce que les individus croient et ce qu'ils choisissent de partager sur les réseaux sociaux (voir plus loin), un décalage qui s'explique largement selon eux par l'inattention ou par des vices épistémiques (paresse, légèreté, superficialité de la réflexion) plutôt que par une volonté délibérée de diffuser de la désinformation.
Dans ce modèle, ce qui prémunit de la désinformation, en miroir logique des vices cognitifs qui la propagent, ce sont des vertus intellectuelles, en particulier l'esprit critique, l'attention, ou encore la capacité de contextualiser un énoncé et l'effort d'en évaluer avec rigueur la provenance (« *lateral reading* »). La notion de **contextualisation** mise en lumière dans ce modèle est particulièrement importante. On a vu que le fait d'être *partielle* est un défaut sérieux en matière d'information : une bonne information ne se caractérise pas seulement par sa qualité interne de véracité mais aussi par sa *quantité* suffisante, laquelle s'apprécie notamment au regard de son contexte. Par exemple, s'il n'est pas faux de dire qu'une bonne hygiène de vie est nécessaire à une bonne santé, en revanche il est partial, malveillant, fallacieux ou dangereux de présenter cet énoncé comme suffisamment informatif à des personnes souffrant de dépression ou de cancer par exemple. L'omission d'une partie d'information, de même qu'une déformation dans la contextualisation de l'information, sont de puissants ressorts de mensonge et de désinformation, même sans recourir à un énoncé intrinsèquement faux. C'est le sens de la catégorie que l'OCDE appelle « *contextual deception* » dans sa taxonomie.
Ce modèle centré sur les causes cognitives et psychologiques des croyances erronées emporte avec lui une vision de l'action publique en réponse qui va privilégier les stratégies de *debunking* (pour corriger les erreurs de traitement par la preuve) et de *prebunking* (vacciner contre l'erreur ex ante) mais aussi des interventions de type nudges attentionnels : pour contraindre par exemple le consommateur à prendre quelques secondes de réflexion avant de pouvoir reposter un énoncé dont il aura ainsi été forcé d'évaluer la pertinence.
Une littérature en plein développement (on y reviendra plus loins) montre alors qu'en introduisant des **prompts attentionnels** (par exemple "*Consider whether this headline is accurate*"), les participants sont amenés à se recentrer sur le contexte, la plausibilité, la cohérence (« *shifting attention to accuracy* »). Les résultats disponibles sont frappants :

ils suggèrent qu'un tel rappel minimal à contextualiser l'information ou à s'interroger sur sa fiabilité réduit significativement la proprension à partager une information quand elle est fausse, *mais pas quand elle est vraie*. Ceci est d'autant plus important qu'une baisse, même mineure, des partages, signifie une réduction bien plus importante de l'exposition globale (une baisse de 10 % des intentions de partage d'une fausse nouvelle entraîne une réduction de 40% du nombre de personnes qui y seraient exposées) (Bago et al. 2020 ; Pennycook et al. 2021a).

L'OCDE porte également des recherches de cet ordre et retient effectivement l'idée que les « *accuracy evaluation prompts may be a promising tool for improving the quality of information shared online* » (OCDE 2022). Un autre outil du même type est aussi évalué favorablement par l'OCDE : les « **digital media literacy tips** », de courtes consignes pédagogiques destinées à aider les utilisateurs à évaluer un titre (headline) avant de le partager (par ex. vérifier la source, lire l'article, chercher une confirmation ailleurs, se méfier des titres sensationnalistes et des appels à l'émotion, etc.).

L'avantage de ce modèle est qu'il permet de parier sur le renforcement de l'esprit critique des consommateurs de l'information, sans qu'il soit besoin d'introduire à son sujet quelque notion de véracité a priori. Nul besoin d'arguer de l'autorité du Réel ou du Vrai pour déclarer la guerre au mensonge : le repérage de la pathogénicité de la désinformation est confié aux consommateurs eux-mêmes, qui apprenent à réduire leur risque par l'auto-support.

L'un des avantages de ce modèle est aussi, selon le plaidoyer qu'y consacre par exemple l'OCDE, qu'il permet de rompre avec la logique *top-down* de la régulation : « *Relying solely on traditional top-down approaches that aim to regulate content are insufficient at limiting the immediate dangers of misinformation* ». A rebours de cette logique, le modèle promu est un modèle d'auto-support, où l'usager (ou sa communauté) devient son propre régulateur : l'OCDE parle d'interventions « *collaborative*s » et « *user-end* » (par opposition à « *system-level* ») et vante le fait qu'elles sont « scalable » et « *ouvrent des pistes convaincantes pour* ***outiller les individus*** *afin qu'ils puissent faire des choix actifs concernant la qualité de l'information, tout en faisant de la* ***préservation de l'autonomie des citoyens*** une priorité ».

Ces outils peuvent donc offrir des approches **préventives** et **complémentaires**, mais on voit mal comment les articuler avec les approches systémiques qui **régulent**, établissent des standards et contrôlent la diffusion de fausses informations en ligne. Le modèle est puissant pour décrire les *mécanismes internes* de la mauvaise information et de la croyance, mais il a tendance à rabattre la désinformation sur un déficit d'exactitude et de pertinence intellectuelle à l'échelle des individus, là où se jouent en réalité aussi des enjeux structurels de pouvoir, d'identité, de puissance économique et de conflictualité qu'il ne semble pas souhaitable de laisser dans l'ombre au profit du prisme de la rationalité individuelle et des vices de la crédulité.

2.3. Le modèle KABP

Que la désinformation soit souvent interprétée comme le produit d'un déficit de rationalité ou d'une crédulité excessive, voilà qui semble absurde à certains cognitivistes, comme le montrent les auteurs d'un rapport récent de l'Institut Jean Nicod/Département d'études cognitives de l'Ecole normale supérieure (Darcy, 2025). Déplorant les séductions de la « folk psychology », ils notent que « *cette lecture intuitive structure de nombreux discours et réponses institutionnelles - de l'enseignement de l'esprit critique à la prolifération des dispositifs de fact-checking - fondées sur l'idée qu'il suffirait de corriger des erreurs de raisonnement pour tarir la diffusion des fausses nouvelles. Or, si les limites de certaines de nos capacités cognitives jouent bien un rôle, elles n'expliquent qu'une partie du phénomène. Croire ou partager de la désinformation ne tient pas seulement à une prétendue incapacité à distinguer le vrai du faux, mais répond aussi à des motivations sociales, émotionnelles et identitaires* ».

Faut-il alors déplacer la focale vers la question des identités, des croyances et des comportements ? Un modèle classique pour penser les problèmes publics est de les cartographier à partir de métriques combinées portant sur les connaissances (Knowledge), les attitudes (Attitudes), les croyances (Beliefs) et les pratiques (Practices) – ou modèle

dit KABP, largement mobilisé dans la lutte contre le sida par exemple : croire à tort que le VIH se transmet sur la cuvette des toilettes est-il corrélé avec plus ou avec moins de comportements adéquats devant le risque ? Adhérer à une information fausse, est-ce un problème en soi ou bien est-ce un problème parce que des croyances et comportements déviants risquent fort d'en être inférés ?

L'une des questions que pose ce modèle concerne les corrélations et surtout les causalités que l'on peut ou non établir entre les variables information/croyance/comportement. En matière de désinformation, on peut penser qu'une bonne façon de mesurer le phénomène, de le décrire et donc d'y répondre, pourrait bien être de caractériser l'enchaînement causal suivant : *exposition à un énoncé fallacieux -> croyance erronée -> comportement inapproprié*. La logique qui sous-tend cette vue est que la dés/mésinformation jouerait un rôle déterminant dans la motivation de croyances dangereuses et de comportements négatifs. Autrement dit, si des affirmations factuellement inexactes n'avaient pas été rencontrées, alors les comportements dommageables ne se seraient pas produits. C'est d'ailleurs parce qu'on tient souvent cet enchaînement intuitif pour vrai et comme allant de soi, qu'on lit un peu partout que faire le constat que des énoncés faux se diffusent massivement en ligne vaut ipso facto constat démontré que les gens adhèrent massivement aux croyances erronées que portent ces énoncés, et adoptent du même coup les comportements aberrants qu'ils promeuvent.

Or tel n'est pas le cas, au vu de la recherche. Ce modèle, pour intuitif qu'il soit, n'est pas validé à ce jour par les données sur la dés/mésinformation qui parlent au contraire de *Belief/Action gap*. En conséquence, la piste consistant à lutter contre la dés/mésinformation à l'aune des comportements dommageables qu'elle favorise (au plan des choix électoraux, sanitaires, environnementaux…) n'est pas robuste au sens *evidence-based*.

De nombreux travaux montrent d'abord une déconnexion entre ce que les gens croient et ce qu'ils relaient sur les plateformes. La revue de référence publiée par Gordon Pennycook en 2021 (Pennycook 2021 b) révèle que les jugements portant sur le partage de contenus sur les réseaux sociaux divergent en réalité fortement des jugements portant sur leur exactitude (*accuracy*). Par exemple, des participants à qui l'on demandait d'évaluer l'exactitude d'un ensemble de titres (*headlines*) dont certains étaient vrais et d'autres faux, jugeaient à raison les vrais titres nettement plus fiables et exacts que les faux ; mais lorsqu'on leur demandait s'ils partageraient volontiers ces faux titres, ils s'y disaient enclins et la véracité avait peu d'effet sur leurs intentions de partage : « *beaucoup de gens sont enclins à partager des contenus qu'ils ont pourtant pu identifier comme inexacts* » (Pennycook et al., 2020, 2021a).

Plus largement, les données disponibles ne permettent pas d'établir de séquence logique continue liant de façon causale l'exposition à des énoncés fallacieux, la difficulté à les repérer pour tels, la propension à les partager, l'inclination à les croire, et l'adoption effective des comportements négatifs que ces contenus suggèrent. Comme le notait en 2024 le Parliament Office of Science and Technology du Parlement britannique, si la recherche a bien établi que la désinformation peut modifier les *croyances* de ceux qui y sont exposés, en revanche **les données concernant son influence sur *les comportements* demeurent** à ce jour **non-conclusives**. Certes, durant la pandémie de Covid-19, le lien entre exposition à des contenus anti-vax et adhésion au vaccin a pu être exploré, notamment dans une étude de la London School of Hygiene and Tropical Medicine qui révélait que la proportion de répondants enclins à accepter le vaccin tombait de 54.1% à 47.9% après exposition à de la désinformation sur le vaccin (Loomba et al. 2021). Quant aux comportements électoraux, l'étude de Eady et al. déjà citée a établi que la campagne de désinformation russe en faveur de Donald Trump durant la campagne électorale de 2016 n'a eu que peu d'effet sur les comportements de vote : « *we find no evidence of a meaningful relationship between exposure to the Russian foreign influence campaign and changes in attitudes, polarization, or voting behavior* » (Eady et al., 2023). Ainsi, une revue récente pilotée par la chercheuse britannique Zoe Adams pose qu' « *il n'a pas encore été démontré de manière empirique et fiable que les comportements inappropriés sont le résultat de la désinformation* », corrélation ne signifiant pas causalité (Adams et al. 2023).

Ces limites ne veulent pas dire pour autant que la piste de recherche soit inepte. Il est possible que les plans d'études mobilisés jusqu'ici soient inappropriés et que d'autres, à venir, permettent de reconsidérer à nouveaux frais la pertinence de l'impact de la dés/mésinformation sur les comportements. La revue de Zoe Adams propose en ce

sens plusieurs pistes d'approfondissement, en s'appuyant notamment sur le progrès des sciences comportementales en matière d'observance médicamenteuse ou d'adhésion aux prescriptions de prévention (notamment les gestes barrières Covid-19) : le canonique « Health Belief Model » qui a porté les modèles KABP dans la deuxième moitié du XXe siècle bénéficie aujourd'hui d'apports importants qui affinent la modélisation des interactions entre information, croyances et comportements, médiées également par des variables fines autour des compétences psycho-sociales (sentiment d'auto-efficacité) et des déterminants sociaux de la motivation (poids des normes sociales) (voir par exemple Michie, 2014).
Reste que, d'après le rapport de l'Institut Nicod, « *cette lecture réductrice néglige l'influence décisive des contextes sociaux, des vulnérabilités économiques, de la défiance institutionnelle et des logiques identitaires. Elle ignore également que la désinformation peut parfois agir sans même convaincre, en modifiant la perception des normes sociales et en répondant à des besoins d'appartenance et de reconnaissance. À défaut d'une approche systémique, fondée sur les apports des sciences cognitives et sociales, ces stratégies resteront largement inefficaces face aux racines structurelles de la vulnérabilité informationnelle* ».

2.4. Le modèle « théâtre impro »

C'est ici qu'intervient, pour mieux rendre compte de la dimension participative, jouissive et identitaire de la désinformation, une modélisation inattendue et disruptive : celle du théâtre-impro. L'équipe de **Kate Starbird** à l'Université de Washington propose cette modélisation nouvelle de la dés/mésinformation fondée une métaphore surprenante : celle de la performance collaborative **d'un théâtre improvisé et participatif**.
L'idée de fond défendue dans ces travaux est que « ***la désinformation n'est pas quelque chose qui est fait aux foules, mais quelque chose que les foules se font à elles-mêmes*** » (Starbird, 2019). Dans ses travaux réputés sur les mouvements conspirationnistes et les campagnes de désinformation, Kate Starbird insiste sur le caractère collaboratif et partiellement improvisé de la production de récits en ligne : différents acteurs (complotistes, trolls, influenceurs) improvisent sur une scène fictive une trame qu'ils adaptent en temps réel aux réactions du public (les comptes « ordinaires »). Lors d'une conférence récente à l'Université de Washington citée par le *New York Times* à l'appui d'une analyse des stratégies de désinformation de l'administration Trump, Kate Starbird expliquait : « *Disinformation is participatory, with elites (in media and politics) collaborating with social media influencers and everyday people to produce and spread contents for political goals. Participatory disinformation takes shape as improvized* ***collaborations between witting agents and unwitting (although willing) crowds of sincere believers****. These collaborations follow icreasingly well-worn patterns and use increasingly sophisticated tools. They are becoming structurally embedded into the socio/technical infrastructure of the Internet* » (Starbird. 2025).
Tout en rendant visible que certains de ces acteurs sur scène sont délibérément mal intentionnés dans la poursuite d'un agenda qui leur est propre (Kate Starbird développe notamment sa thèse au sujet de l'écosystème médiatique MAGA et *far-right*), ce modèle veut surtout rendre compte du fait que la désinformation est moins un script figé qu'une performance collective et collaborative, où les frontières entre producteurs, amplificateurs et récepteurs-spectateurs sont poreuses : il n'y a pas d'un côté les « *organic crowds* » et de l'autre les « *orchestrated agents* ». Cette métaphore permet de rendre visibles la dimension créative, l'ajustement permanent et le caractère distribué de l'action, que les modèles d'erreur ou de contagion tendent à écraser. Surtout, ce modèle prête attention à la désinformation en tant que **système de rétroaction** : l'interaction entre le public et les influenceurs, de même que les architectures algorithmiques elles-mêmes, créent des boucles de renforcement où l'engagement passé nourrit la visibilité future, où les signaux faibles peuvent être brusquement amplifiés sous les feux de la rampe, et où les comportements des utilisateurs modifient eux-mêmes en continu non seulement les textes, mais même les paramètres du système.
L'intérêt de ce modèle est aussi de porter l'attention sur le plaisir et la jouissance que public trouve à participer : « *Dans cette perspective, l'influence ne va pas seulement des influenceurs sur scène vers le public, mais circule aussi du public vers les influenceurs. (…)*

Le retour d'information est immédiat, et les « bons » morceaux reçoivent rires et likes. Les influenceurs - et les responsables politiques - peuvent ajuster très vite leurs messages en fonction des goûts, des attentes et des frustrations de leur audience, ainsi qu'en fonction des événements du moment, sans la lenteur des médias traditionnels. Les acteurs et leur public testent aussi des contenus transgressifs, controversés, voire offensants, pour explorer les limites de leurs goûts partagés, de leurs attentes et - pour les figures politiques - de leurs positions idéologiques. ***Comme dans beaucoup de spectacles d'impro, ces performances donnent une impression d'intimité et d'authenticité.*** *Les membres du public peuvent discuter avec les performeurs après, et parfois même pendant le show. Ils peuvent aussi être « invités sur scène » quand un influenceur met en avant leur contenu. Parfois c'est juste pour une scène, mais* ***certains participants chanceux, habiles ou tenaces peuvent devenir partie intégrante du théâtre des influenceurs. Cela augmente la motivation à participer, l'excitation, et le sentiment, pour le public, de vraiment faire partie du spectacle*** » (Tomson & Starbird, 2024).
L'un des intérêts de ce modèle est de pointer combien la production de désinformation qui se joue sur la scène des influenceurs est **déresponsabilisée** quant à sa véracité et à ses effets : parce qu'il s'agit d'une narration mouvante et coproduite au sein d'une émission essentiellement diffuse, en outre souvent juste par jeu ou pour rire, la désinformation est difficile à penser sous l'angle de la responsabilité normative qui incombe à ses producteurs, transmetteurs, amplificateurs, etc. Pour autant, il faut noter que ce modèle ne se détourne pas du tout de l'intentionnalité de nuire contenue dans le terme « désinformation » ; il veut surtout montrer comment les opérations stratégiques de désinformation délibérée s'emparent, notamment en situation de crise, d'**un besoin humain banal de « collective sensemaking** » et parviennent à orchestrer **un désir diffus dans le public d'expression et de participation**.

Participer plus souvent et plus décisivement qu'avec un bulletin de vote : tel est le ressort potentiel de la désinformation que Kate Starbird propose d'explorer.
Pour autant, on pressent qu'hormis un vibrant et vertueux appel à ce que chacun s'empare de la participation et devienne à son tour partie prenante du spectacle pour l'influencer vers davantage de véracité, ce modèle ouvre relativement peu de pistes à l'action publique : là où le modèle de la contagion permet par exemple d'identifier clairement des procédés prophylactiques et où le modèle de l'auto-support valide des outils, le modèle de l'impro participative cartographie certes efficacement le système d'interactions en jeu - mais ne dit que peu de choses concrètes sur leur substance et la nature du poison qu'il faudrait y combattre.

En somme, ces modèles ne sont pas seulement en concurrence : **ils échouent, chacun à leur manière, à saisir dans toute son extension le phénomène qu'ils veulent représenter**. Tant que les analyses restent prisonnières d'un seul registre métaphorique - contagion, erreur et auto-support, croyances ou théâtre - elles risquent de manquer la synergie des dimensions systémiques, économiques, identitaires, stratégiques et techniques qui, ensemble, font la spécificité du problème. Si l'on prend au sérieux cette pluralité de modèles, la question n'est donc pas de choisir *le bon* mais de reconnaître que **chacun éclaire un fragment du phénomène en laissant d'autres dimensions dans l'ombre** : ainsi peut-être est-il possible de prendre conscience des effets que ces modèles produisent sur nos représentations du problème afin de mieux combiner les avantages de chacun.

Tableau comparatif des modèles conceptuels de la désinformation

Modèle	Avantages	Limites
Modèle de la contagion (épidémiologique)	Rend compte du rôle respectif de la circulation, de l'intensité d'exposition, de la pathogénicité des énoncés, de la transmission en clusters et de l'existence de superspreaders. Prête attention aux mécanismes préventifs (théorie de l'inoculation).	Naturalise la diffusion ; minimise les stratégies intentionnelles ; sous-estime les dimensions participatives et identitaires, les motivations des publics exposés/transmetteurs.
Modèle d'auto-support centré sur l'erreur (épistémique/cognitif)	Décrit finement les biais et vices cognitifs	Réduit la désinformation à un problème d'inexactitude ; ignore les dimensions politiques, économiques et identitaires
Modèle KABP	Intuitif ; pose la question du lien entre information, croyance et comportement	Réducteur et statique ; ne capture pas les dynamiques identitaires et sociales
Modèle du théâtre improvisé (Kate Starbird)	Met en lumière la production participative et collaborative ; capture sa dimension rétroactive, jouissive, organique et adaptative	Dit peu de choses sur la substance des interactions qu'il décrit et sur la nature du poison qui s'y joue : quelle action de prévention ?

3. Pourquoi aucun cadre ne parvient réellement à s'imposer ?

Lors de ses prises de parole successives sur menaces qui pèsent sur la démocratie à l'heure des réseaux sociaux, Emmanuel Macron a explicitement indexé l'expression de son besoin politique d'un nouveau cadre de pensée commun au principe qui, par définition, l'oblige : *défendre l'émancipation démocratique et républicaine*. Un principe qu'il réfère naturellement à l'exigence régalienne d'assurer à chacun un accès sûr au vrai et à la rationalité, dûment protégé des offensives de l'obscurantisme.

Constatant que les conditions de possibilité de la poursuite de cet idéal fondateur d'émancipation sont aujourd'hui mises à mal du fait de la dés/mésinformation, il a posé, pour cadre de l'action collective qu'il veut susciter, la nécessité impondérable de défendre l'héritage politique et moral des Lumières et, en miroir, de combattre en s'appuyant sur la défense de la science l'abîme qui menace nos sociétés d'un retour à la « minorité » au sens kantien, sous tutelle arbitraire et obscurantiste. Cette invocation des Lumières fonctionne ainsi à la fois comme horizon normatif et comme principe de légitimation de l'action publique face aux dérives informationnelles contemporaines. Elle inscrit la lutte contre la dés/mésinformation dans une généalogie politique et intellectuelle où la défense de la raison, de la science et de l'autonomie du jugement constitue moins un choix contingent qu'un impératif constitutif du projet démocratique lui-même.

Réinventer un cadre d'action politique contre la dés/mésinformation pensé sur le modèle de celui des fondateurs de la République et de son idéal d'émancipation : c'est là l'eschatologie dans laquelle il propose aux parties prenantes de s'inscrire aujourd'hui à nouveaux frais.

3.1. La véracité comme référentiel : un renoncement collectif de fait

Bien que nous adhérions sans réserve à la nécessité politique de redonner corps et sens à ces idéaux, un constat s'impose : en matière de lutte contre la désinformation la maxime kantienne des Lumières « *Sapere Aude* » (Ose savoir ! ») et la référence à la véracité de la science comme levier d'émancipation, pour désirable que soit cet idéal, ne sont pas identifiées aujourd'hui par les institutions ou les acteurs privés comme l'outil fédérateur dont l'action publique a besoin pour être efficace.

La référence au *vrai* est même devenue franchement problématique.

Quelle est la caractéristique principale de cette référence au vrai ? Elle valide une autorité normative incontestable du vrai et de la science. Elle désigne la désinformation comme problème public à l'aune des effets obscurantistes de l'ignorance et du mensonge (qu'il soit déviant par commission, par omission, ou par légèreté). Elle retient que le critère qui permet d'identifier distinctement des déviances informationnelles pathologiques en référence à une norme d'intérêt général, c'est le rapport au vrai et au réel en tant qu'il serait par nature émancipateur des intelligences *et donc* des volontés.

C'est bien sur ce point de la référence à *une norme* et à ses pathologies que la lutte contre la désinformation en ligne bute aujourd'hui. On voit bien que le **critère d'autorité** qui la fonde, et qu'il s'agirait d'assumer, semble peu ou prou devoir être une revendication de véracité, d'ancrage dans la science, la connaissance, la méthode ou le savoir.

Or toutes les autorités publiques et privées sur ce sujet constatent cependant que ce qui complique la tâche morale et politique de revendiquer cette autorité est de **la définir d'une façon qui reste compatible avec le respect des libertés d'expression et d'opinion**. Il s'agit là d'un paradoxe de départ qui hante au fond toutes les réflexions politiques sur la dés/mésinformation : **quelle attitude choisir face au risque que nos stratégies de lutte contre la désinformation soient taxées par leurs détracteurs de censure, de revendication arbitraire d'un monopole de la raison, d'un ministère du vrai**, et finalement de crimes contre la liberté d'expression et la liberté d'opinion ? De plus, ce cadre de lecture ne répond en rien à la question de l'information vraie mais partielle ou sortie de son contexte, aboutissant à donner une version erronée du réel.

A mesure que les entreprises de réseaux sociaux ont tenté de répondre aux problèmes de dés/mésinformation, le refus d'indexer ces stratégies à un critère de véracité des contenus s'est imposé. Elles ont à plusieurs reprises exprimé leur réticence à se concentrer sur le contenu, déclarant souvent qu'elles ne souhaitent pas devenir les « arbitres de la vérité » (voir par exemple ici la position de Facebook dès 2017 et ici celle de Youtube en 2018). Beaucoup ont développé, à la place, des politiques visant à définir les *comportements* qu'elles jugent problématiques.

Cette perspective est reprise aujourd'hui par l'OCDE par exemple, qui considère que, de même que les plateformes, les États eux non plus ne sauraient s'arroger la mission de devenir des « arbitres de la vérité » et réguler à ce titre la désinformation en ligne sur un critère de véracité des contenus : « *Malgré un consensus autour des défis posés par la diffusion de mésinformation et de désinformation,* ***les démocraties peinent à y répondre tout en protégeant la liberté d'expression*** *ainsi que l'accès à une information libre, diverse et fiable. Préserver les libertés civiques fondamentales et un Internet ouvert implique que la mésinformation et la désinformation ne disparaîtront jamais complètement.* ***Puisqu'il ne revient pas aux gouvernements de « gouverner l'information » ni de se poser en « arbitres de la vérité »,*** *une approche globale visant à instaurer des mécanismes de contrôle et d'équilibre dans l'écosystème informationnel doit aller au-delà du seul traitement de la désinformation* » (OCDE, 2024).

Ce déplacement du contenu vers les comportements, s'il permet d'éviter une régulation explicitement fondée sur la vérité, n'est toutefois pas neutre sur le plan normatif et politique. Il redéfinit implicitement ce qui est jugé acceptable ou problématique dans l'espace public numérique, en transférant la focale de la véracité des énoncés vers les conditions de leur production, de leur circulation et de leur amplification.

Mais alors, si ce n'est donc la vérité, à quel critère veut-on indexer une politique publique favorable à l'information ?

3.2. Le référentiel de l'authenticité

Se jugeant illégitimes pour aborder la nature intrinsèque des contenus, les plateformes ont choisi de se focaliser sur l'authenticité et le comportement de leurs utilisateurs ; par exemple, la politique actuelle de Facebook met l'accent sur la notion de « comportement inauthentique coordonné ». Le mot **authenticity,** repris notamment dans la régulation européenne (Code of Practice 2022 ; Digital Services Act 2024) désigne alors un ensemble de pratiques opérationnelles visant à détecter les **comportements non humains** (bots, scripts automatisés, agents IA), repérer les **réseaux coordonnés** agissant comme s'ils étaient des individus isolés, traquer les **faux comptes** (fake accounts, sock puppets) et ainsi démanteler les **info ops** (opérations d'influence) qui imitent du comportement « organique ».
Le concept d'authenticité a le mérite de permettre aux régulateurs de sanctionner des comptes sans en juger le contenu, et ainsi de se prémunir des accusations de « biais politiques » et de censure auxquelles les exposerait la référence à la véracité. Il permet aussi de parier sur des pratiques de régulation purement gestionnaires et référées au design des plateformes plutôt qu'à des choix politiques au sens propre, de défense de la démocratie.
Mais "authenticité" est ici un terme bien pudique pour désigner des réalités beaucoup plus dures : la présence d'acteurs non humains, de faux comptes ou de réseaux coordonnés opérant clandestinement pour influencer l'opinion publique en produisant du consensus faux pour saturer le débat public aux point d'en détruire les conditions de possibilité. L'euphémisme « inauthentic behavior », comme s'il s'agissait d'un simple écart de conduite inapproprié, individualise, moralise et adoucit un phénomène qu'il faudrait plutôt nommer *guerre cognitive à visage flouté* ou *milices numériques du chaos*.
En outre, et comme l'a montré Kate Starbird, la notion d'authenticité passe à côté de la dimension participative de la dés/mésinformation : elle échoue à rendre compte d'une caractéristique systémique cruciale des opérations de désinformation, qui est leur capacité à s'appuyer sur de vrais comptes ou sur des participants "authentiques" et sincères.

3.3. Le référentiel de l'intégrité

La critique à l'égard de la notion d'authenticité donne lieu à une nouvelle évolution du référentiel : après la *véracité*, condamnée pour sa proximité avec une forme de volonté de censure, puis l'*authencitité*, par trop euphémisante, c'est aujourd'hui *l'intégrité* qui fonde les approches institutionnelles et privées.
L'**intégrité** ambitionne de traiter **les systèmes**, et non plus les *contenus* ni les *profils*. Alors que la véracité se concentrait sur *ce qui est* dit et l'authenticité sur *qui parle*, l'intégrité vise *dans quelles conditions l'information circule*. Le passage de l'authenticité à l'intégrité informationnelle reflète une évolution majeure : la prise de conscience que **la désinformation contemporaine n'est plus ni un problème de mensonge ni un problème de faux comptes**, mais un **problème systémique** impliquant :
l'économie de l'attention,
l'architecture technique,
les boucles algorithmiques,
les dynamiques identitaires et participatives,
la porosité entre "opérations orchestrées" et "engagement organique" au sens de Starbird.
Elle permet donc d'inclure dans la réflexion et l'action la transparence des algorithmes, les mécanismes d'amplification, les biais de visibilité ou les circuits économiques. Cet angle est explicitement soutenu par l'OCDE (2024) et correspond également au cadre du Digital Services Act européen qui demande l'analyse des risques systémiques (polarisation, manipulation, vulnérabilité cognitive), la mitigation algorithmique et la transparence des systèmes de recommandation.
De même, du côté des plateformes, Meta, Google et YouTube parlent désormais de *systemic integrity*, *platform integrity*, *information integrity* pour désigner leur engagement à vérifier la *robustesse* des flux, c'est-à-dire à évaluer les conditions de circulation pour éviter les distorsions structurelles et réduire le pouvoir des opérations manipulatoires.

Au demeurant, la notion d'intégrité informationnelle suppose que l'on puisse évaluer la qualité, l'équité et la robustesse des systèmes socio-techniques qui organisent la circulation de l'information. Or une partie croissante de la littérature montre que cette prétention se heurte à un obstacle majeur : l'impossibilité pratique d'**auditer les infrastructures algorithmiques** des grandes plateformes (littérature sur l'*auditability*).
Les travaux sur les audits algorithmiques ont posé comme point de départ que ces enquêtes visent à tester l'existence de biais, discriminations ou autres déficiences **sans accès direct aux "*internals*"** des systèmes audités. Une synthèse de Metaxa et al. (2021) est devenue la référence en systématisant cette approche appelée « *audit from the outside in* », qui tente d'inférer le fonctionnement des systèmes de recommandation et de ranking à partir de requêtes répétées et de l'observation systématique des *outputs* pour tester l'algorithme comme un objet dont les déterminations propres demeurent opaques. **L'audit d'intégrité est forcément une méthode *second-best,* parce que les conditions d'un contrôle classique (accès au code, aux données et aux décisions) ne sont pas réunies.** En ce sens, si la littérature ne contredit pas l'idée que l'intégrité informationnelle soit un idéal souhaitable, elle montre que, sans auditabilité rigoureuse, ce concept risque de rester un registre rhétorique, difficilement vérifiable et largement dépendant de la parole des plateformes elles-mêmes. L'objet d'audit étant, selon le terme de Metaxa, non le système lui-même, mais des traces partielles de ce qu'il produit de façon « à la fois dynamique et éphémère », il s'agit là d'une contrainte méthodologique majeure qui pèse non seulement sur la science, mais aussi sur l'action publique qui pourrait en être inférée. C'est pourquoi, dans son essai « Understanding Social Media Recommendation Algorithms » (2023), Arvind Narayanan, informaticien de premier plan, souligne qu'**il est paradoxal de voir l'importance que l'intégrité des systèmes prend dans les** débats publics et politiques **au regard de la superficialité ou même pauvreté de ce qu'il est possible d'en dire vraiment**, en partie parce que la compréhension de ces systèmes repose sur des sources très limitées : la littérature technique spécialisée, la documentation minimale fournie par les entreprises, et quelques fuites ou documents internes.
Le concept d'intégrité informationnelle est donc puissant en théorie, mais en pratique il pâtit fortement d'un décalage entre son importance sociale et la pauvreté des informations qui permettent de lui donner corps.
Par exemple, l'un des principaux angles de préoccupation politique en matière d'intégrité concerne la capacité des réseaux informationnels à créer des « filter bubbles » qui renforcent les convictions erronées des utilisateurs et donc la polarisation du débat public. De nombreuses positions institutionnelles retiennent ce sujet comme éminemment problématique : l'OCDE, par exemple, note que « *les bulles de filtre peuvent effectivement contribuer à expliquer la propagation* » de la mésinformation et constituent à ce titre une priorité d'action. Mais les travaux d'Axel Bruns sur sur ce sujet (Bruns, 2019) montrent que les grandes affirmations sur l'effet « bulle » des algorithmes de personnalisation reposent souvent sur des données fragmentaires, fournies ou cadrées par les plateformes elles-mêmes. Ces métriques incomplètes, à partir de jeux de données impossibles à relier à des décisions de *ranking* observables, empêchent notamment tout raisonnement par la contrefactuelle (ce que les gens auraient pu voir mais n'ont pas vu), qui seul permettrait d'évaluer rigouresement l'existence ou non de bulles de filtre et leurs effets.

3.4. La désinformation met structurellement en échec la méthode scientifique elle-même

Ce que suggère la littérature aujourd'hui, c'est en somme que la désinformation pose un problème épistémologique d'un type singulier dans l'histoire des sciences : **elle met en échec les conditions mêmes de possibilité de la méthode scientifique appliquée au phénomène**. La méthode scientifique exige de fait un objet stable, observable selon des taxonomies solides, et mesurable : or **la dés/mésinformation n'est ni un type de**

contenu identifiable (son nom même est flou !), ni un type d'acteur, ni une structure intentionnelle univoque, ni un format, ni une classe bien définie de données.
Les infrastructures de désinformation sont embarquées dans les systèmes opaques et protégés des plateformes (recommandations, amplification, modération, signaux d'engagement), ce qui empêche toute observation directe des dynamiques causales. **On ne peut, au mieux, mesurer que ce qui est visible (ce que les gens voient) mais pas ce qu'ils auraient pu voir : une exposition contrefactuelle pourtant méthodologiquement nécessaire** à tout **raisonnement causal.** En outre, le critère de reproductibilité, fondamental pour la méthode scientifique, en est notamment fragilisé : les fragments de données obtenus par les chercheurs sont éphémères.
Alors que, comme on l'a vu ci-dessus, les tentatives de taxonomie et de modélisation sont fragiles, il faut donc ici en outre statuer que l'accès même aux données constitue une réelle aporie pour ne serait-ce que décrire, penser et comprendre la désinformation dans un registre scientifique.
Le fait qu'elle modifie constamment sa grammaire en réponse aux tentatives de la décrire, et produise de fait ainsi activement de l'ignorance à son propre sujet, fait-il de la désinformation un objet structurellement inaccessible à la connaissance ?
Au terme de cette analyse, il faut sans doute parier que cette aporie n'implique pas que l'action soit condamnée à l'aveuglement. Elle invite plutôt à un déplacement du regard : si la science ne parvient pas à stabiliser un cadre de compréhension systémique pour l'expansion de la dés/mésinformation, elle isole malgré tout les **conditions** à la fois cognitives et **socio-techniques** utiles pour la combattre.
C'est ici que des approches dites "*user-end*" trouvent leur force : des interventions comportementales simples, qui passent facilement à l'échelle, centrées sur l'attention à l'exactitude et la contextualisation, ont montré qu'elles pouvaient réduire le partage de fausses informations, tout en préservant l'autonomie des citoyens et en déjouant les accusantions de censure à l'encontre des « ministères du vrai ». L'objet central est ici le « nudge attentionnel », qui n'a pas de définition canonique mais qu'on peut comprendre, par extension du concept général de nudge appliqué spécifiquement à l'attention cognitive ou perceptive d'un individu, comme : une forme de nudge qui oriente ou capte l'attention d'un individu vers un élément d'information, une option ou un comportement donné, en modifiant subtilement les signaux visuels, cognitifs ou perceptifs de l'environnement, sans restreindre les choix disponibles ni imposer de contrainte explicite. Là où les leviers systémiques restent lents et vulnérables aux critiques habituelles à l'égard de la régulation, de simples nudges attentionnels semblent pouvoir freiner le plaisir jouissif et identitaire sur lequel paraît jouer la diffusion du faux. En outre, alors que le côté moralisateur des nudges imposés peut faire craindre que leur multiplication entraîne des stratégies de rejet et de contournement, il y a aujourd'hui avec l'IA, sa prospective de développement local *on-device* et ses capacités de *tuning* adaptatif, de bonnes raisons d'imaginer à l'avenir des prompts d'alerte dont la fréquence, le style et le timing puissent être finement paramétrés en fonction des usages et des situations à risque. La détection *on-device* et en temps réel de messages dangereux éphémères (notamment les *scams*), qui se développe déjà aujourd'hui pour alerter l'usager par l'envoi d'un « *warning interstitial* » au moment où il est en risque (Google security Blog, Gemini Nano, 2025) pourrait bien signaler là une piste prometteuse.
Dans un registre de régulation plus modeste encore, les **Community Notes** (anciennement *Birdwatch* ou *crowd-sourced fact-checking*), lancées par X (alors Twitter) en 2021, constituent un dispositif de modération collaborative visant à lutter contre la désinformation par l'ajout de contextes explicatifs rédigés et évalués par des utilisateurs volontaires. Le système repose sur un mécanisme algorithmique cherchant à faire émerger des notes jugées utiles par des contributeurs aux profils idéologiquement divers, afin d'augmenter la probabilité d'un consensus transversal et de limiter les biais partisans. Des travaux expérimentaux montrent que ces notes accroissent la perception de fiabilité des corrections, améliorent la capacité des utilisateurs à identifier des contenus trompeurs et peuvent réduire l'engagement et la diffusion de publications erronées (Slaughter et al., 2025). L'intérêt croissant des grandes plateformes pour ce modèle confirme son attractivité institutionnelle : Meta a annoncé en 2025 tester un dispositif et TikTok a également annoncé expérimenter un système "*Footnotes*". Ce modèle participatif pourrait-il compléter demaine les dispositifs professionnels de fact-checking et les régulations publiques de l'espace informationnel ?

A ce stade de prospective, il semble que la littérature permette de faire émerger un débat crucial sur lequel il est urgent d'avancer : celui que Lin, Pennycook et Rand décrivent dans *Cognition* comme une controverse entre l'hypothèse que **la lutte contre la dés/mésinformation impose que chacun apprenne à "penser plus" (davantage de délibération) et l'hypothèse qu'il faille plutôt apprendre à "penser autrement" (redirection attentionnelle vers le vrai)**.

Alors qu'augmenter les capacités délibératives de chacun semble une tâche colossale pour l'action publique, convoquant l'ensemble de l'idéal émancipateur républicain et les missions de son école, en revanche privilégier l'idée qu'il suffise de modifier les paramètres d'attention par des design d'usage permet clairement davantage d'optimisme opérationnel. Il y a peu, Bence Bago, Gordon Pennycook et David Rand concluaient carrément leur étude par un plaidoyer massif en faveur des vertus intellectuelles : « *Most broadly, our results support the conclusion that encouraging people to engage in* ***more thinking*** *will be beneficial rather than harmful* ». Mais leurs modélisations les plus récentes suggèrent à rebours qu'une action plus modeste, mobilisant des techniques ouvertement triviales pour diriger l'attention des utilisateurs vers le vrai en créant de simples frictions intellectuelles face au faux et au douteux, produit une forme de sagesse collaborative vertueuse qui démonétise mécaniquement les énoncés dangereux (Lin, Pennycook et Rand. 2023). C'est ainsi qu'aujourd'hui, ils semblent s'orienter plutôt vers une piste plus humble : celle de nudges attentionnels qui viendraient modestement faire limite au plaisir identitaire de diffuser le n'importe quoi qu'on reçoit.

Adhérer à cette piste implique toutefois de changer radicalement nos modes de pensée, et notamment de dépasser l'idée commune selon laquelle la dés/mésinformation activerait un processus de traitement rapide, automatique, intuitif, référé à des heuristiques de familiarité ou d'émotion, alors qu'y répondre exigerait d'activer un autre processus de traitement, plus lent et délibératif, exigeant des compétences de raisonnement. Ce schéma-là, normatif et hiérarchisé (que la littérature appelle « *dual process interpretation* » parce qu'il distingue deux processus, l'un primaire et l'autre réfléchi) engage inévitablement l'action publique dans une ambition extensive : combattre les vices de l'intelligence et promouvoir les vertus de la raison.

Il faut bien le dire à ce stade : ce schéma de lecture politique de la désinformation est d'abord foncièrement convaincant pour les héritiers des Lumières que nous sommes. C'est même quasiment un repère politique fondateur, que cette idée qu'en jouant sur l'émotion instinctive du cœur les réseaux sociaux déjouent les vertus de la raison, alors qu'inversement le politique a pour fonction de rétablir l'efficacité de la délibération par-delà les facilités rhétoriques du « *movere, placere* » démagogique. Au fond, c'est un repère quasi incontournable de l'action politique contre la désinformation que de juger que le spectacle émotionnel du dissensus (la « culture du *clash* ») sur les réseaux sociaux doit être combattue par la diffusion des vertus intellectuelles délibératives de l'esprit critique éclairé (le « *docere* » de l'art oratoire cicéronien).

Au-delà même des réseaux sociaux, c'est bien cette dissociation entre émotions et intellect qui fonde par exemple l'inquiétude majeure suscitée par la note du Cepremap parue début 20256 dans laquelle Yann Algan, Thomas Renault et Hugo Subtil analysent deux millions de discours prononcés à l'Assemblée nationale entre 2007 et 2024 pour révéler l'influence qu'y exercent les réseaux sociaux en montrant combien « l*a rhétorique émotionnelle s'est imposée depuis 2017, et de façon encore plus marquée à partir de 2022, tandis que le débat rationnel recule, diminuant ainsi leur caractère délibératif. Aujourd'hui, plus de la moitié des discours se rapprochent davantage de l'émotionnel que du rationnel* ».

Bien qu'un certain nombre de données disponibles sur les mécanismes de diffusion de la désinformation contredisent désormais, comme on l'a vu, l'idée qu'un partage d'énoncés absurdes sur les réseaux sociaux rime forcément avec une défaite assumée de la raison, la tentation est forte de bâtir notre représentation politique des enjeux de la désinformation autour d'un combat ancestral entre raison et sentiments, entre vertus délibératives de l'entendement éclairé et obscurantisme idiomatique des émotions.

Faut-il assumer de poser en ces termes-là les enjeux politiques de la lutte contre la désinformation ? L'analyse proposée ici a montré au contraire que l'appareil conceptuel dont nous disposons est sans doute trop faible pour parier alors sur un succès. Car il faut bien constater la réticence politique à se référer à quelque notion de véracité que ce soit, combinée aux données de la recherche qui révèlent de fait la complexité des motivations qui président au partage de contenus ne serait-ce qu'ambigus.

L'intérêt de la science qui est en train de se faire sur ces sujets est peut-être de révéler plutôt la richesse des analyses centrées sur le ressort *participatif* de l'information en ligne. C'est le message politique que délivrent notamment les analyses de Kate Starbird. L'action publique a alors pour objet, non plus de déjouer les ruses impressionnistes d'un idiome émotionnel diffus et nocif, mais de substituer sa propre réponse aux mauvaises réponses que les réseaux sociaux apportent à ce qu'il faut bien reconnaître comme une bonne question : comment faire pour que la participation démocratique du XXIe siécle permette aux citoyens, outre leur bulletin de vote, d'influencer directement les narratifs qui circulent et colorent nos représentations du bien commun ?

On proposera en conclusion de retenir que les outils « attentionnels » assument à juste titre et avec modestie cette ambition politique qui n'est pas si mineure qu'elle en a l'air. Comme chez Platon à la fin de l'exposé du mythe de la caverne, il s'agit moins de « *faire entrer de force la vision dans des yeux aveugles* » que de « *tirer en douceur l'œil de l'âme, enfoui dans quelque bourbier infâme, pour l'entraîner vers le haut* » (Rép, VII, 533d). Ce serait se tromper sur ce qu'est l'éducation, nous dit le philosophe prêt à redescendre dans la caverne, que d'oublier que « *le savoir est déjà dans l'âme* » des prisonniers : les chaînes de l'ignorance ne sont pas invincibles à qui veut bien faire émerger la vision du vrai, à condition qu'il veuille bien s'y prendre depuis le bourbier même.

Surtout, cette piste est inséparable d'une volonté ferme de renforcer l'accès de tous à une information édifiante et d'encourager la participation à sa diffusion. Le modèle du théâtre impro collaboratif de Kate Starbird peut ici retenir toute notre attention : si le sujet concerne un désir collectif d'avoir sa place dans ce qui se joue, de produire sa part de sens, de se mettre en jeu avec les autres et de peser davantage sur les récits petits ou grands qui façonnent notre monde, alors c'est finalement d'un besoin de **participation politique** et d'une aspiration démocratique qu'il s'agit peut-être aussi. Or est-il plus facile sur les réseaux sociaux aujourd'hui de participer à des récits fallacieux qu'à des récits justes pour y éprouver le plaisir d'en porter sa part de sens ? C'est bel et bien à craindre, tant la scène proposée en ligne par les autorités publiques demeure encore souvent raseuse ou timide et peu sophistiquée, faute de compétences dédiées performantes, par auto-censure et par peur de la contre-productivité ou du retour de bâton.

Il est temps, plaide Kate Starbird, que montent sur scène autant d'influenceurs du vrai qu'il y a d'agents de l'ignorance. L'urgence est de produire, au sein de chaque autorité politique, éducative ou scientifique, les compétences techniques et la volonté stratégique de proposer en ligne des espaces de sens ouverts et collaboratifs auxquels chaque citoyen puisse participer ; des récits justes et émancipateurs qui rendent à chaque personne son statut de connaissant libre et sa capacité de contribuer, à ce titre et en plus de son bulletin de vote, à produire chaque jour sa part du sens d'un grand récit commun. Cette perspective invite ainsi à déplacer l'attention des seuls mécanismes de correction ou de restriction vers une véritable politique de production, de mise en scène et de circulation du vrai. Elle suggère que la lutte contre la désinformation ne peut être dissociée d'un investissement durable dans des capacités narratives, techniques et institutionnelles permettant au débat démocratique de redevenir désirable, intelligible et habitable pour tous.

Ces constats invitent à dépasser une approche strictement défensive de la désinformation pour lui préférer une stratégie articulant modélisation systémique, investissement dans les capacités institutionnelles et réouverture d'espaces de participation épistémique. Une telle orientation suppose de penser conjointement infrastructures techniques, pratiques sociales et finalités démocratiques, afin de rendre à l'information sa fonction émancipatrice.

Références

Chavalarias, D. (2023). *Toxic data*. Flammarion.

Code of Practice 2022, DSA 2024/2025

Adams, Z., Osman, M., Bechlivanidis, C., & Meder, B. (2023). (Why) Is Misinformation a Problem? *Perspectives on Psychological Science*, *18*(6), 1436-1463.

Bago, B., Rand, D. G., & Pennycook, G. (2020, January 9). Fake News, Fast and Slow: Deliberation Reduces Belief in False (but Not True) News Headlines. *Journal of Experimental Psychology: General*.

Broda, M., & Strömbäck, J. (2024). Misinformation, disinformation, and fake news: lessons from an interdisciplinary systematic literature review.

Darcy, G., Mercier, H., Mari, A., & Casati, R. (2025). *Lutter contre la désinformation: Penser autrement l'action publique à l'aune des sciences cognitives* (No. fu9cz_v1). Center for Open Science.

Eady, G., Paskhalis, T., Zilinsky, J., Bonneau, R., Nagler, J. & Tucker, J. A. Exposure to the Russian Internet Research Agency foreign influence campaign on Twitter in the 2016 US election and its relationship to attitudes and voting behavior. *Nat. Commun.* 14, 62 (2023).

EDMO. (2024). Defining Disinformation across EU and VLOP Policies.

Fallis, D. (2015). *What Is Disinformation?* Library Trends, 63(3), 401-426.

Floridi, L. (2005). Is semantic information meaningful data? Philosophy and Phenomenological Research, 70(2), 351–370.

Floridi, L. (2011). The Philosophy of Information. Oxford University Press.

Hayward, T. (2024). The Problem of Disinformation: A Critical Approach. *Social Epistemology*, *39*(1), 1–23.

Heard

Jack, C. (2017). *Lexicon of lies: Terms for problematic information*. Data & Society Research Institute. https://datasociety.net/output/lexicon-of-lies/

Kant E. (1784). *Réponse* à la question «*Qu'est-ce que les Lumières?*»

Lesher, M., H. Pawelec and A. Desai (2022), "Disentangling untruths online: Creators, spreaders and how to stop them", *OECD Going Digital Toolkit Notes*, No. 23, OECD Publishing, Paris, https://doi.org/10.1787/84b62df1-en.

Lin, H., Pennycook, G., & Rand, D. G. (2023). Thinking more or thinking differently? Using drift-diffusion modeling to illuminate why accuracy prompts decrease misinformation sharing. *Cognition*, *230*, 105312.

Loomba, S., de Figueiredo, A., Piatek, S. J., de Graaf, K., & Larson, H. J. (2021). Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature human behaviour*, *5*(3), 337–348. https://doi.org/10.1038/s41562-021-01056-1

Machado, C., & Aguiar, T. (2023). Emerging regulations on content moderation and misinformation policies. Business and Human Rights Journal.

Mercier, H., & Sperber, D. (2017). The enigma of reason. Harvard University Press.

Metaxa, D., Park, J. S., Robertson, R. E., Karahalios, K., Wilson, C., Hancock, J., & Sandvig, C. (2021). *Auditing Algorithms: Understanding Algorithmic Systems from the Outside In.* Foundations and Trends in Human–Computer Interaction, 14(4), 272–344.

Michie, S., Atkins, L., & West, R. (2014). *The behaviour change wheel: A guide to designing interventions*. Silverback

Narayanan, A. (2023). *Understanding Social Media Recommendation Algorithms.* Knight First Amendment Institute, Columbia University.

OECD. (2022). *Misinformation and disinformation: An international effort using behavioural science to tackle the spread of misinformation*.

OECD. (2024). *Facts not Fakes: Tackling Disinformation, Strengthening Information Integrity*.

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological science*, *31*(7), 770-780.

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021 a). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*(7855), 590-595.

Pennycook, Gordon et al. (2021 b) The psychology of fake news. Trends in Cognitive Sciences, Vol 25, Issue 5, 388 – 402

Shannon, C. E. (1948). *A mathematical theory of communication*. *Bell System Technical Journal, 27*(3), 379–423; (4), 623–656.

Shannon, C., & Weaver, W. (1949). The Mathematical Theory of Communication. University of Illinois Press.
Slaughter, I., Peytavin, A., Ugander, J., & Saveski, M. (2025). *Community Notes moderate engagement with and diffusion of false information online*. arXiv.
Simion M. (2024) Knowledge and Disinformation. *Episteme* 21(4):1208-1219.
Sperber, D. (1996). La contagion des idées. Odile Jacob.
Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. Mind & Language, 25(4), 359–393.
Starbird, K. et al. (2019). Disinformation as Collaborative Work: Surfacing the
Participatory Nature of Strategic Information Operations. PACMHCI. Vol: CSCW, Article 127. November
2019.
Starbird, K. (2025). *A Spotlight on Rumors: Illuminating How Influence and Improvisation Shape Online Conversations, Whashington* University Faculty Lecture. https://www.washington.edu/facultystaff/lecture/
Tomson, D. L., & Starbird, K. (2024, December 4). *How right-wing media is like improv theater*. The Conversation. https://theconversation.com/how-right-wing-media-is-like-improv-theater-243665
Uscinski, J. E., Littrell, S., & Klofstad, C. (2024). *The importance of epistemology for the study of misinformation*. Current Opinion in Psychology, 57, 101789.
Wardle, C., & Derakhshan, H. (2017). Information Disorder. Council of Europe.

Nos partenaires :

Nous contacter :

Vous souhaitez participer
à nos activités et recevoir
nos actualités :
contactez-nous à l'adresse
contact@institut-evidences.fr
ou sur le site web dans l'onglet
« Participer ».

Contact :
Valentin Berdah
+33 6 24 18 23 12
vberdah@institut-evidences.fr

institut-evidences.fr

EVIDENCES

LE THINK TANK DÉDIÉ À LA SCIENCE DANS LA SOCIÉTÉ