International Panel on the Information Environment

Répondre à la désinformation générée par l'IA générative

Résultats d'une méta-analyse des données scientifiques

Résumé à l'intention des décideurs 2026.2

Numéro DOI : 10.61452/QXAF2136

A. Herasimenka, S. Valenzuela, S. Boulianne, F. Esser, L. M. Given, S. Lewandowsky, E. M. Navarro-Lopez, P. N. Howard

Répondre à la désinformation générée par l'IA générative

Résultats d'une méta-analyse des données scientifiques

Résumé à l'intention des décideurs

Référence bibliographique :
International Panel on the Information Environment [A. Herasimenka, S. Valenzuela, S. Boulianne, F. Esser, L. M. Given, S. Lewandowsky, E. M. Navarro-López, P. N. Howard (dir.)], « Responding to Generative AI Misinformation: Results from a Meta-Analysis of Scientific Evidence » (Répondre à la désinformation générée par l'IA générative : Résultats d'une méta-analyse des données scientifiques), Zurich, Suisse : IPIE, 2026. Résumé à l'intention des décideurs, SFP2026.2, doi: 10.61452/QXAF2136.

SYNOPSIS

L'intelligence artificielle générative (IA générative) peut désormais produire rapidement de grands volumes de textes, d'images, de contenus audio et de vidéos trompeurs à l'aide d'outils facilement accessibles. Cette évolution est importante pour les décideurs, car les contenus faux ou trompeurs peuvent être adaptés à des publics précis. Ces contenus peuvent se propager rapidement et imiter des communications authentiques, ce qui complique l'évaluation de leur véracité.

Le présent Résumé à l'intention des décideurs expose les principales conclusions du rapport de synthèse (SR2026.2) de l'IPIE sur les effets de la désinformation produite à l'aide de l'IA générative et sur les mesures les plus susceptibles d'en réduire l'influence.

Cette évaluation repose sur une méta-analyse à grande échelle de données scientifiques expérimentales. Elle s'appuie sur 60 estimations de l'effet issues d'essais contrôlés randomisés, provenant de 24 publications évaluées par les pairs, portant sur 33 801 participants et publiées entre 2018 et 2025.

Le rapport aboutit à quatre grandes conclusions :

1. La désinformation textuelle générée par l'IA présente actuellement des risques de persuasion plus importants que la désinformation visuelle.
2. Les données actuellement disponibles excluent la majeure partie du monde. La recherche se concentre sur les pays anglophones et à revenu élevé, ce qui laisse d'importantes lacunes dans nos connaissances.
3. L'intervention dont l'efficacité est la plus constante consiste à fournir préventivement aux utilisateurs des informations correctives afin qu'ils puissent évaluer eux-mêmes l'exactitude et la crédibilité des contenus.
4. L'étiquetage est efficace lorsqu'il est appliqué de manière cohérente. L'étiquetage des contenus réduit généralement leur crédibilité perçue, mais son impact varie fortement selon la manière dont les entreprises conçoivent et appliquent les étiquettes.

Les implications pour les politiques publiques sont claires. Les décideurs devraient accorder la priorité à la lutte contre la désinformation textuelle générée par l'IA et soutenir la diffusion d'informations préventives et correctives en tant que stratégie fondamentale. L'étiquetage doit être considéré comme une mesure nécessitant une conception et des essais rigoureux. Nous devons considérablement élargir la recherche indépendante au-delà des contextes anglophones et des pays à revenu élevé.

L'accès des chercheurs aux données des plateformes et des modèles est essentiel pour améliorer la compréhension publique de la désinformation générée par l'IA et vérifier l'efficacité réelle des mesures de protection.

INTRODUCTION

Les systèmes d'intelligence artificielle générative (IA générative) peuvent produire des textes, des images, des contenus audio et des vidéos qu'il est parfois difficile de distinguer des contenus créés par des humains. Les contenus générés par l'IA se sont largement répandus et sont apparus dans le cadre d'élections partout dans le monde — souvent en provenance de sources inconnues ou malveillantes —, tout en influençant de plus en plus la perception du public dans des domaines tels que la santé publique, les conflits et les marchés, avec d'importantes répercussions concrètes [1]. Ces systèmes comprennent notamment les grands modèles de langage qui génèrent du texte, les outils de synthèse d'images et de contenus audio, ainsi que les systèmes de création vidéo. Ils peuvent également produire des contenus trompeurs ou préjudiciables, comme de fausses informations, des rumeurs, de la propagande ou des deepfakes. Dans le présent rapport, ces phénomènes sont collectivement désignés par le terme « désinformation ».

Le présent Résumé à l'intention des décideurs expose les principales conclusions d'une méta-analyse des données expérimentales relatives aux effets, sur les individus, de l'exposition à la désinformation créée par l'IA générative [2]. Il évalue également deux contre-mesures étayées par suffisamment de données expérimentales pour faire l'objet d'un débat sur les politiques publiques : les informations correctives et l'étiquetage des contenus. Le rapport met l'accent sur les essais contrôlés randomisés, car ces études permettent des comparaisons plus fiables entre les expériences et une estimation plus systématique des effets moyens.

Les données révèlent un environnement informationnel en évolution rapide. Le rapport constate que les réactions du public à la désinformation textuelle et visuelle divergent désormais. Il constate également que la portée géographique et les méthodes des données disponibles restent limitées. Parallèlement, il recense deux mesures concrètes susceptibles de réduire la crédibilité perçue de la désinformation générée par l'IA, bien que l'une produise des résultats plus constants que l'autre.

Pour constituer cette base de données, les auteurs de la revue ont effectué des recherches dans deux grandes bases de données bibliographiques, Scopus et Web of Science, intégré les recommandations d'experts, fusionné les résultats, supprimé les doublons et examiné un ensemble final de 6 952 publications. Après vérification des critères d'admissibilité et codage du texte intégral, 87 publications répondaient aux critères d'un examen approfondi. Des contrôles supplémentaires des biais, des vérifications et des examens du protocole suivant les recommandations de Cochrane ont abouti à un sous-ensemble de 24 études particulièrement pertinentes, retenues pour la méta-analyse. Le rapport de synthèse principal recourt à une méta-analyse à effets aléatoires à trois niveaux et évalue le risque de biais à l'aide de l'outil révisé d'évaluation du risque de biais dans les essais randomisés. Veuillez consulter le rapport de synthèse complet [2] pour connaître l'intégralité des détails méthodologiques.

RÉSULTAT 1. LES TEXTES TROMPEURS GÉNÉRÉS PAR L'IA REPRÉSENTENT LA PLUS GRANDE MENACE

PRINCIPALE CONCLUSION

La désinformation textuelle générée par l'IA présente actuellement des risques de persuasion plus importants que la désinformation visuelle.

Le rapport met en évidence une nette différence entre la perception de la fiabilité des contenus textuels et celle des contenus visuels générés par l'IA. Les études les plus récentes montrent que les faux textes générés par l'IA sont souvent perçus comme plus exacts, crédibles ou vraisemblables que les informations véridiques. Dans le même temps, les fausses informations visuelles générées par l'IA, telles que les deepfakes, sont souvent perçues comme moins crédibles que les informations véridiques. Dans les études sur les contenus visuels menées après 2021, l'estimation combinée indique une baisse modérée de la crédibilité perçue.

Cela ne signifie pas que la désinformation visuelle soit sans danger ou sans importance. Les deepfakes présentent toujours des risques, et le rapport souligne que les réactions du public pourraient à nouveau évoluer à mesure que les systèmes visuels s'améliorent. Toutefois, les données actuelles indiquent que la désinformation textuelle doit faire l'objet d'une attention plus urgente de la part des pouvoirs publics. Ce constat est important, car les systèmes axés sur le texte peuvent être peu coûteux, déployables à grande échelle, faciles à personnaliser et de plus en plus intégrés aux moteurs de recherche, aux services de messagerie et aux interfaces conversationnelles. Le rapport indique également que les réponses des IA conversationnelles pourraient être nettement plus persuasives que les messages statiques générés par l'IA, ce qui laisse penser que les estimations actuelles pourraient être prudentes.

La tendance générale des résultats est suffisamment nette pour étayer un message clair à l'intention des décideurs : le risque le plus immédiat et le plus persuasif provient actuellement des textes faux ou trompeurs, et non des seuls deepfakes visuels.

RÉSULTAT 2. LACUNES DES DONNEES A L'ECHELLE MONDIALE

Le rapport souligne clairement que la plupart des publications examinées portent sur des pays anglophones et à revenu élevé, tels que l'Autriche, l'Allemagne, les Pays-Bas, la Corée du Sud, le Royaume-Uni et les États-Unis. Les données actuellement disponibles se limitent aux articles en texte intégral publiés en anglais, et cinq publications examinées dans d'autres langues ne répondaient pas aux critères d'admissibilité. Par conséquent, les connaissances actuelles sur la désinformation générée par l'IA se concentrent sur un petit nombre de pays, de langues et d'environnements médiatiques.

Il s'agit d'une limite importante. Cela signifie que les décideurs, les autorités de réglementation, les chercheurs et les entreprises disposent encore de connaissances limitées sur le fonctionnement de la désinformation générée par l'IA dans de nombreuses autres régions. Le rapport indique explicitement que cette situation crée une importante

lacune dans les connaissances au-delà des quelques pays et langues dans lesquels des travaux sont publiés sur ce sujet. Cette lacune est importante pour les sociétés multilingues, la réglementation transfrontalière et les efforts visant à mettre au point des mesures de protection efficaces dans différentes langues, cultures et différents systèmes politiques.

PRINCIPALE CONCLUSION

Les données actuellement disponibles laissent hors de leur champ la majeure partie du monde, tant les pays que les langues.

En outre, les données sont limitées par le rythme rapide des avancées technologiques. De nombreuses expériences portent sur des systèmes plus anciens, comme GPT-2, ou sur des outils de première génération, comme FaceSwap. Le rapport avertit que la recherche scientifique accuse souvent un retard sur les technologies actuelles. Par conséquent, les politiques publiques reposent parfois sur des données concernant des systèmes moins performants que ceux actuellement accessibles au public. La recherche indépendante portant sur des contextes plus diversifiés constitue donc une nécessité concrète pour garantir que la réglementation et la gouvernance restent en phase avec les technologies qui façonnent le débat public.

RÉSULTAT 3. LES CORRECTIONS PREVENTIVES PEUVENT ETRE UTILES

Les informations correctives préventives comprennent de brefs avertissements pédagogiques, des rappels ou des invites, ainsi que d'autres moyens de prémunir les utilisateurs avant leur exposition à des contenus potentiellement trompeurs. Par exemple, certaines interventions expliquent brièvement comment les deepfakes permettent de manipuler des vidéos de manière réaliste ou rappellent aux utilisateurs que les systèmes d'IA générative peuvent produire des erreurs. Ces approches visent à améliorer la capacité des utilisateurs à évaluer la fiabilité des contenus.

Dans les études postérieures à 2020, l'estimation combinée montre que les informations correctives préventives réduisent la perception de l'exactitude et de la crédibilité de la désinformation générée par l'IA, ainsi que son acceptation, avec une taille d'effet faible à modérée. Ces résultats sont relativement cohérents d'une étude examinée à l'autre : les variations entre les résultats sont faibles, et des effets sont observés auprès de différentes populations et avec différents protocoles expérimentaux dans ce sous-ensemble de données.

PRINCIPALE CONCLUSION

Les informations correctives préventives constituent l'intervention dont l'efficacité est la plus constante.

Ces résultats indiquent que de telles interventions devraient rester efficaces à court terme, même si la plupart des données disponibles reposent sur des études portant sur des générations antérieures de systèmes d'IA générative.

La revue de synthèse souligne également une réserve importante : l'effet positif des informations correctives préventives apparaît lorsque les participants sont invités à évaluer l'exactitude ou la crédibilité des contenus. Les données sont moins cohérentes lorsque les études portent uniquement sur la détection. Des recherches antérieures sur les deepfakes ont montré que les avertissements pouvaient parfois renforcer la méfiance à l'égard des vidéos authentiques comme des

vidéos manipulées. Ce constat rejoint une préoccupation plus générale de la recherche et des politiques publiques relatives à la désinformation : les récits mensongers et la propagande cherchent souvent à amener les individus à se méfier de toute information publique. Les interventions doivent donc être conçues de manière à améliorer la capacité des individus à distinguer les contenus fiables des contenus trompeurs.

Les informations correctives préventives comptent parmi les mesures les plus largement testées pouvant être déployées en ligne relativement rapidement, en particulier sur les plateformes de réseaux sociaux. Il peut notamment s'agir de notes indiquant que l'IA générative peut produire des erreurs, d'explications sur la manière dont les deepfakes permettent de manipuler des vidéos de façon réaliste et de brefs supports pédagogiques consacrés à des phénomènes tels que les hallucinations de l'IA. Selon la discipline, ces mesures sont désignées comme des conseils, des techniques d'inoculation, de pré-réfutation, d'amorçage ou des étiquettes d'avertissement.

RÉSULTAT 4. L'ETIQUETAGE EST EFFICACE LORSQU'IL EST COHERENT

PRINCIPALE CONCLUSION

L'étiquetage des contenus reste prometteur, mais uniquement lorsqu'il est clair et cohérent.

L'étiquetage des contenus désigne l'utilisation de courtes mentions visuelles, textuelles ou multimodales associées à une information. Ces étiquettes indiquent que l'information peut avoir été créée à l'aide de l'IA générative, avoir été modifiée ou être trompeuse. Dans l'analyse combinée, les étiquettes sont associées à une diminution faible, mais statistiquement significative, de la crédibilité perçue des informations trompeuses générées par l'IA. Ce résultat est encourageant, mais il ne se vérifie pas dans tous les cas. Le rapport constate que les données relatives à l'étiquetage présentent des variations nettement plus importantes que celles portant sur les informations correctives préventives.

En raison du nombre limité de données, les effets varient considérablement, et certaines études ne constatent aucun effet. Les étiquettes peuvent réduire la crédibilité dans certaines situations, mais pas dans d'autres. Cette évaluation met en évidence plusieurs facteurs modérateurs probables, notamment la conception, la formulation et la présentation des étiquettes, le type de contenu, le cadre expérimental et la source de l'étiquette. Autrement dit, la manière dont les étiquettes sont conçues et mises en œuvre importe au moins autant que leur utilisation elle-même.

Les données disponibles sur l'étiquetage sont également limitées. Les études de cet échantillon relativement restreint portent presque exclusivement sur des participants aux États-Unis. Il est donc difficile de savoir comment les étiquettes fonctionneront dans différentes langues, cultures ou différents systèmes juridiques. Il ne faut donc pas en conclure que l'étiquetage est inefficace. L'étiquetage peut au contraire être utile, mais uniquement si les entreprises réfléchissent à la conception de ces mentions et les testent dans différents contextes, au lieu de supposer qu'elles fonctionnent partout par défaut.

CONCLUSION

Le présent Résumé à l'intention des décideurs met en évidence une évolution de la nature des risques. La désinformation créée par l'IA générative textuelle représente une menace persuasive plus importante que celle créée par l'IA générative visuelle. Cela n'élimine toutefois pas les dangers liés aux deepfakes ou aux autres médias synthétiques. Cela signifie que les décideurs ne doivent pas laisser la forte visibilité de la tromperie visuelle détourner leur attention du pouvoir de persuasion croissant des faux textes générés par l'IA.

Le présent rapport met en avant les résultats des précédentes évaluations de l'IPIE [3], [4] et propose deux réponses concrètes. Les informations correctives préventives constituent l'intervention la plus constante et la mieux étayée par les données actuellement disponibles, en particulier lorsqu'elles sont fournies avant l'exposition et qu'elles améliorent la capacité d'évaluation. L'étiquetage des contenus peut également être utile, mais son efficacité varie et dépend du contexte. Les étiquettes sont plus efficaces lorsqu'elles sont soigneusement conçues et appliquées de manière claire et cohérente.

Le dernier point concerne les données elles-mêmes. La majeure partie du monde reste en dehors du champ des recherches actuelles. Cela limite la capacité des décideurs, des plateformes et des développeurs de modèles d'IA à usage général à comprendre comment ces risques se propagent dans différents environnements linguistiques, culturels et médiatiques. La recherche indépendante, la synthèse continue des données et l'amélioration de l'accès des chercheurs aux données des plateformes et des modèles sont donc essentielles. Sans ces éléments, les politiques publiques resteront réactives, incomplètes et trop étroitement ciblées dans un environnement informationnel en évolution rapide.

RÉFÉRENCES

[1] International Panel on the Information Environment [I. Trauthig, P. N. Howard, S. Valenzuela (eds.)], "The Role of Generative AI Use in 2024 Elections Worldwide", Zurich, Switzerland: IPIE, 2025. Document technique, TP2025.2, doi : 10.61452/HZUE9853.

[2] International Panel on the Information Environment [A. Herasimenka, S. Valenzuela, S. Boulianne, F. Esser, L. M. Given, S. Lewandowsky, E. M. Navarro-López, P. N. Howard (eds.)], « Effects of Misinformation Produced with Generative AI and its Countermeasures: A Meta-Analysis of Experimental Scientific Evidence », Zurich, Suisse : IPIE, 2026. Synthesis Report, SR2026.2, doi: 10.61452/UGTR3022.

[3] International Panel on the Information Environment, « Countermeasures for Mitigating Digital Misinformation: A Systematic Review », IPIE, Zurich, Suisse, SR2023.1, juillet 2023. [En ligne]. Disponible sur : https://www.ipie.info/research/sr2023-1

[4] International Panel on the Information Environment, « Platform Responses to Misinformation: A Meta-Analysis of Data », IPIE, Zurich, Suisse, SR2023.2, juillet 2023. [En ligne]. Disponible sur : https://www.ipie.info/research/sr2023-2

REMERCIEMENTS

Contributeurs

Auteurs rédacteurs : Aliaksandr Herasimenka (chercheur consultant, Royaume-Uni), Sebastián Valenzuela (directeur scientifique de l'IPIE et président du comité scientifique et méthodologique, Chili), Shelley Boulianne (membre du comité scientifique et méthodologique de l'IPIE, Canada), Frank Esser (membre du comité scientifique et méthodologique de l'IPIE, Suisse), Lisa M. Given (membre du Comité scientifique et méthodologique de l'IPIE, Canada/Australie), Stephan Lewandowsky (membre du Comité scientifique et méthodologique de l'IPIE, Australie/Royaume-Uni), Eva M. Navarro-López (membre du Comité scientifique et méthodologique de l'IPIE, Espagne/Royaume-Uni/Mexique), Philip Howard (président-directeur général de l'IPIE, Canada/Royaume-Uni). Assistants de recherche : Anna George et Xianlingchen Wang. Examens généraux indépendants : George Georgarakis et Mathias Harrer. Conception : Domenico Di Donna. Révision : Beverley Sykes. Nous remercions le secrétariat de l'IPIE pour son soutien : Lola Gimferrer, Jessica Gold, Wiktoria Schulz, Donna Seymour, Anna Staender et Alex Young.

Financeurs

L'International Panel on the Information Environment (IPIE, Panel international sur l'environnement de l'information) remercie ses organismes de financement pour leur soutien. Pour consulter la liste complète des partenaires financiers, rendez-vous sur www.ipie.info. Les opinions, constatations, conclusions ou recommandations exprimées dans le présent rapport sont celles de l'IPIE et ne reflètent pas nécessairement les points de vue des bailleurs de fonds.

Déclaration d'intérêts

Les rapports de l'IPIE sont élaborés et révisés par un réseau mondial de chercheurs affiliés et de scientifiques consultants qui constituent des groupes scientifiques spécialisés et des équipes de collaborateurs. Tous les contributeurs et réviseurs remplissent des déclarations d'intérêts, qui sont examinées par l'IPIE aux stades appropriés du travail.

Citation préférée

Le *Résumé pour les décideurs politiques* de l'IPIE fournit un aperçu de haut niveau de l'état des connaissances et est rédigé pour un large public. Un *rapport de synthèse* de l'IPIE utilise des techniques de méta-analyse scientifique, d'étude systématique et d'autres outils d'agrégation des preuves, de généralisation des connaissances et d'établissement d'un consensus scientifique. Il est rédigé à l'intention d'un public d'experts. Une *note technique* de l'IPIE aborde des questions méthodologiques spécifiques ou propose une analyse politique portant sur un problème réglementaire précis. Tous les rapports sont disponibles sur le site web de l'IPIE (www.IPIE.info).

Ce document doit être cité comme suit :

International Panel on the Information Environment [A. Herasimenka, S. Valenzuela, S. Boulianne, F. Esser, L. M. Given, S. Lewandowsky, E. M. Navarro-López, P. N. Howard (dir.)], « Responding to Generative AI Misinformation: Results from a Scientific Meta-Analysis », Zurich, Suisse : IPIE, 2026. Résumé à l'intention des décideurs, SFP2026.2, doi : 10.61452/QXAF2136.

Informations sur les droits d'auteur

À PROPOS DE L'IPIE

L'International Panel on the Information Environment (IPIE) est une organisation scientifique indépendante et mondiale qui s'engage à fournir les connaissances scientifiques les plus exploitables sur les menaces qui pèsent sur l'environnement de l'information dans le monde. Basé en Suisse, l'IPIE a pour mission de fournir aux décideurs politiques, à l'industrie et à la société civile des évaluations scientifiques indépendantes sur l'environnement mondial de l'information en organisant, en évaluant et en valorisant la recherche, dans le but général d'améliorer l'environnement mondial de l'information. Des centaines de chercheurs du monde entier contribuent aux rapports de l'IPIE.

Pour plus d'informations, veuillez contacter l'International Panel on the Information Environment (IPIE), secretariat@IPIE.info. Seefeldstrasse 123, B.P. 8034 Zurich, Suisse.

International Panel on
the Information
Environment

Seefeldstrasse 123
P.O. Box 8034 Zurich
Suisse