« IA, dis-moi comment voter »

Les *chatbots* peuvent-ils influencer nos opinions politiques ?

par Romain Badouard

Le succès des chatbots d'IA générative ravive un débat ancien sur l'influence des outils d'information sur nos opinions politiques. Chercheurs, industriels et régulateurs se penchent sur un enjeu susceptible de jouer un rôle important dans la campagne électorale qui s'annonce.

Que ce soit dans le cadre de leur travail, de l'organisation de leur quotidien, voire dans leur vie intime, des millions de Français interrogent chaque jour un chatbot d'intelligence artificielle, faisant de la recherche d'information le principal usage[1] des outils d'IA générative en France. Les réponses que leur fournissent ChatGPT, Claude, Le Chat ou Gemini sont calibrées en fonction de critères variés, comme la formulation des questions, les filtres de modération et les conditions d'utilisation des chatbots, les méthodes d'entraînement et d'alignement des modèles, ou encore la teneur des conversations passées et les informations personnelles qu'y partagent leurs usagers. Quand les requêtes des internautes portent sur l'actualité ou sur des questions de société, ce calibrage est-il en mesure d'orienter leur perception d'un problème public et des réponses à y apporter ? Pour le dire autrement, ChatGPT et ses concurrents sont-

[1] CREDOC. (2026). *Baromètre du numérique 2026*.

ils en mesure d'influencer l'opinion politique de leurs usagers et si oui, à quelles conditions ?

Si la recherche en sciences sociales a eu tendance à mettre en avant les limites de l'influence de nos outils d'information sur nos opinions, les spécificités des chatbots d'IA générative (notamment leur anthropomorphisme et leur capacité à maintenir des échanges personnalisés sur le temps long) les doteraient d'un pouvoir de persuasion inédit. Les études expérimentales en laboratoire, de leur côté, soulignent que l'influence de ces chatbots résiderait davantage dans leur capacité à sélectionner et mettre en forme les informations qu'ils nous fournissent. Face à ce phénomène, les régulateurs, en France et en Europe, s'entourent de compétences en sciences du numérique et fouilles de données pour développer de nouvelles méthodes d'audit. Chercheurs et régulateurs sont aujourd'hui tournés vers une question vertigineuse : au matin du 18 avril 2027, date du premier tour de la présidentielle, combien de Français et de Françaises se réveilleront en demandant à leur chatbot d'IA « d'après nos conversations des mois passés, d'après toi, pour qui devrais-je voter aujourd'hui ? » ?

Un vieux débat…

Pour les chercheurs en sciences sociales qui s'intéressent aux phénomènes d'influence, ces questions peuvent susciter un haussement d'épaules, voire un soupir. Depuis les travaux fondateurs d'Elihu Katz et Paul Lazarsfeld[2] dans les années 1950, la sociologie des médias n'a cessé de démontrer les limites de l'influence des discours médiatiques et politiques sur celles et ceux qui y sont exposés. Non, les médias n'influencent pas les opinions, en tout cas pas directement. Si influence il y a, c'est d'abord celle du groupe social auquel nous appartenons : nous sommes influencés par nos proches, parce que nous sommes mus par un désir d'intégration qui nous pousse à nous aligner sur leurs attitudes et leurs opinions[3]. Dans ce contexte, l'influence des médias est indirecte : au sein des groupes auxquels nous appartenons, des « leaders d'opinion » les consultent et les mettent en débat, et c'est dans l'échange et

[2] Katz, E., Lazarsfeld, P., (2008). *Influence personnelle. Ce que les gens font des médias*. Armand Colin/INA.
[3] Laurens, S. (2017). *Manipulations et influences : Réalités et représentations à travers deux siècles d'études.* PU Rennes.

l'interprétation collective que leur contenu est susceptible de modifier nos avis, surtout si ces contenus entrent en résonance avec nos principes, nos valeurs, nos intérêts.

Ces questionnements autour de l'influence ont suivi les évolutions technologiques, et sont réapparus au cours des années 2000 et 2010 à travers le débat sur la neutralité de nos outils d'information numériques. Il y a une quinzaine d'années par exemple, les moteurs de recherche faisaient l'objet de questionnements similaires autour de leur influence sur le pluralisme des informations et leur rôle dans l'orientation du débat public en ligne[4]. Google doit-il rester neutre lorsqu'il classe des sites ? Si oui, comment opérationnaliser cette neutralité dans la formule de son algorithme ? L'ordre de classement de sites politiques est-il en mesure d'influencer la formation des opinions des internautes ?

Quelques années plus tard, ce sont les fausses informations, et la manière dont elles sont recommandées sur les réseaux sociaux par les algorithmes de leurs fils d'actualité, qui ont suscité des craintes quant à leur capacité à influencer l'opinion des citoyens et citoyennes en période électorale. Dans le cas des « *fake news* » comme celui des moteurs de recherche pourtant, les enquêtes de terrain ont eu tendance à mettre en évidence que ce sont les opinions politiques qui guident la consultation et le partage des contenus médiatiques, et non l'inverse. De la même façon que les électeurs de gauche lisent *Libération* plutôt que *Le Figaro*, nous consultons en ligne les informations qui ont tendance à conforter nos opinions. Les informations politiques, qu'elles soient vraies ou fausses d'ailleurs, ont tendance à toucher un public déjà convaincu, et n'ont que rarement le pouvoir de faire évoluer des positions préexistantes[5].

… de nouvelles technologies

Est-ce à dire que le débat est clos, et que les craintes qui s'expriment autour des chatbots d'IA générative ne sont que l'expression contemporaine de vieilles paniques

[4] Grimmelmann, J. (2014). Speech Engines. *Minnesota Law Review*, (868) ; Sire, G. (2015). Cinq questions auxquelles Google n'aura jamais fini de répondre. *Hermès*, *n° 73*(3), 201 ; Badouard, R., Lyubareva, I., Maillé, P., & Tuffin, B. (2026). An inquiry into the neutrality of search engines : Developing tools and indicators to compare bias on socially sensitive topics. *Information and Software Technology*, *189*, 107925.

[5] Cardon, D. (2019). Fake news panic : Les nouveaux circuits de l'information. In *Culture numérique* (p. 261-276). Presses de Sciences Po ; Berriche, M. (2023). La réception et le partage de (fausses) informations par les adolescents : Des pratiques situées. *Les Enjeux de l'information et de la communication*, *N° 23/1A*(S1), 87-102 ; Badouard, R. (2021). Fausses informations, vraies indignations ? : Les « fake news » comme support des discussions politiques du quotidien. *RESET*, *10*.

morales ? Ce serait aller un peu vite en besogne. À la différence des moteurs de recherche et des algorithmes de recommandation des réseaux sociaux, les chatbots d'IA générative présentent un certain nombre de caractéristiques qui renouvellent les termes du débat[6]. D'abord, ce sont des agents conversationnels, et nous interagissons avec eux par le langage naturel. Ensuite, les chatbots gardent en mémoire nos conversations passées et sont en mesure de personnaliser leurs réponses et de s'adapter à ce qu'ils perçoivent de nos personnalités. Ils peuvent simuler l'empathie et ont tendance à valoriser nos idées et nos arguments. Nos échanges avec eux se déroulent au quotidien, mais sur le long terme, ce qui susciterait des formes d'attachement émotionnel particulier. Autrement dit, ces chatbots imitent des comportements humains et miment des interactions interpersonnelles, qui demeurent, comme nous l'avons vu, le principal levier d'influence en société. Leur anthropomorphisme les doterait-il d'un pouvoir de persuasion inédit en comparaison des technologies passées, les transformant en nouveaux « leaders d'opinion » pour leurs usagers ?

Pour la communauté des chercheurs et chercheuses qui travaillent dans le domaine de l'AI Safety (sûreté de l'IA), cette question est prise très au sérieux[7]. Aux États-Unis, plusieurs familles ont porté plainte contre OpenAI (la maison mère de ChatGPT), après le suicide d'adolescents qui auraient été encouragés dans leur projet par des chatbots d'IA générative. De manière moins dramatique, le succès des « *AI companions* », des assistants conversationnels conçus pour les échanges romantiques et sexuels, semble, avec leurs dizaines de millions d'usagers réguliers, brouiller les frontières entre interactions humaines et humains-machines. De récentes enquêtes d'opinion[8] montrent par ailleurs à quel point la sollicitation des chatbots comme soutiens psychologiques est généralisée chez les internautes, notamment chez les adolescents et les jeunes adultes.

Les leviers d'influence

Qu'en est-il sur le plan de l'influence politique ? Là encore, les travaux de recherche sont nombreux et leurs résultats, sinon contradictoires, du moins nuancés.

[6] Peter, S., Riemer, K., & West, J. D. (2025). The benefits and dangers of anthropomorphic conversational agents. *Proceedings of the National Academy of Sciences*, *122*(22.
[7] International AI Safety Report 2026. (2026).
[8] IPSOS BVA. (2026). L'IA conversationnelle et la santé mentale des jeunes en Europe.

Certaines études[9] tendent à montrer que si les chatbots parviennent à faire évoluer les opinions de leurs usagers, c'est principalement lorsque cette évolution va dans le sens d'un renforcement des positions existantes. En revanche, lorsqu'il s'agit de mesurer des changements d'avis, les évolutions sont beaucoup plus faibles, voire quasi inexistantes lorsque les échanges tournent autour de sujets à propos desquels les internautes ont des convictions fortes[10]. D'autres à l'inverse[11] mesurent des évolutions importantes dans les positionnements après des échanges politiques avec des chatbots, y compris lorsque ces évolutions vont à l'encontre de positions préexistantes. Ces différences de résultats s'expliquent en partie par les méthodes mobilisées. Certaines études mesurent le caractère persuasif de messages produits par des IA génératives (en les faisant évaluer par des jurys), d'autres étudient des conversations courtes (de 2 à 10 itérations) en demandant aux participants d'auto-évaluer leurs positions avant et après les échanges. La plupart des expérimentations portent sur des IA commerciales, mais certains chercheurs entraînent spécifiquement des modèles à convaincre leurs auditoires, et obtiennent des scores beaucoup plus importants[12]. Parfois les IA ont accès aux conversations passées des participants, et personnalisent leurs réponses en fonction, et parfois non.

Étant donné les différences entre les méthodes mises en œuvre, une question plus pertinente serait de se demander si dans leur diversité, ces études s'accordent sur des facteurs d'influence communs. Et la réponse est oui. Un premier facteur d'influence semble justement être celui du pouvoir de la conversation. Les études qui comparent la dimension persuasive de messages générés ponctuellement et celle de conversations, même courtes, mettent en avant que les échanges sont davantage susceptibles de faire évoluer les positions des usagers. Si une telle évolution est mesurable après seulement 2 à 10 échanges, nous disent les chercheurs, la capacité persuasive d'un LLM[13] sur des échanges s'étirant sur plusieurs semaines, voire

[9] Bai, H., Voelkel, J. G., Muldowney, S., Eichstaedt, J. C., & Willer, R. (2025). LLM-generated messages can persuade humans on policy issues. *Nature Communications*, *16*(1), 6037.

[10] Salvi, F., Horta Ribeiro, M., Gallotti, R., & West, R. (2025). On the conversational persuasiveness of GPT-4. *Nature Human Behaviour*, *9*(8), 1645-1653.

[11] Hackenburg, K., Tappin, B. M., Hewitt, L., Saunders, E., Black, S., Lin, H., Fist, C., Margetts, H., Rand, D. G., & Summerfield, C. (2025). The levers of political persuasion with conversational artificial intelligence. *Science*, *390*(6777) ; Lin, H., Czarnek, G., Lewis, B., White, J. P., Berinsky, A. J., Costello, T., Pennycook, G., & Rand, D. G. (2025). Persuading voters using human–artificial intelligence dialogues. *Nature*, *648*(8093), 394-401.

[12] Hackenburg et al., art. cit.

[13] Large Language Model : les modèles de langage sont des modèles statistiques de grande échelle qui manipulent le langage humain, et sur lesquels reposent le fonctionnement des chatbots d'IA générative.

plusieurs mois, pourrait être bien plus importante, même si cette hypothèse reste difficile à tester « en laboratoire ».

Autre point sur lequel s'accordent les études : les mécanismes de persuasion politique des LLM répondent à des dynamiques différentes de celles qui régissent l'influence psychologique et l'attachement émotionnel. Quand les internautes échangent avec une IA générative, ce n'est ni l'anthropomorphisme de l'outil qui génère de la persuasion, ni la flatterie, ni même le profilage des réponses en fonction des données personnelles de l'usager. Le principal facteur qui semble produire de l'influence politique, c'est la densité des informations fournies et l'impression d'expertise qui s'en dégage. Autrement dit, sur les questions politiques, les usagers font confiance aux LLM quand ces derniers leur livrent des informations qui leur semblent crédibles, à plus forte raison quand celles-ci sont nombreuses. Dans ces conditions, les chatbots incarnent une forme d'autorité épistémique pour celles et ceux qui les utilisent au quotidien.

Fiabilité des informations et disponibilité des arguments

Au premier abord, ces résultats paraissent plutôt rassurants : notre propension à être influencé politiquement lors d'un échange avec un chatbot répond à des facteurs davantage rationnels qu'émotionnels, et relève de la manière dont nous évaluons la qualité des arguments qui nous sont proposés. D'ailleurs, il n'est pas interdit de penser que le recours aux chatbots dans le débat public aide à mieux lutter contre la désinformation[14], en permettant de vérifier les publications en circulation, voire à pacifier les échanges. En France, une expérimentation[15] menée par des chercheurs de l'entreprise Arlequin AI a par exemple comparé le degré de polarisation d'échanges au sein de réseaux humains, de réseaux d'IA et de réseaux hybrides humains-IA, autour de thématiques controversées. La dernière configuration (hybride) obtenait le score de polarisation le plus bas, et le score d'accord le plus élevé, suggérant que les LLM pouvaient également jouer le rôle de médiateurs facilitant la convergence des échanges lors de débats contradictoires. Cette idée a également été expérimentée dans le champ de la démocratie participative, à travers l'entraînement d'un modèle appelé

[14] Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science*, *385*(6714).
[15] Gaubert, L., Devaux, R., Çelen, E., Marjieh, R., Mangalagiu, D., Jardin, A., & Jacoby, N. (2026). *An Experimental Method to Study Opinion Diffusion in Human-AI Hybrid Societies* (arXiv:2605.09197).

« Habermas Machine »[16], qui devait chercher à obtenir des consensus lors de débats contradictoires. Une mission accomplie avec succès par le modèle, qui a surpassé ses homologues humains en la matière.

Pour autant, dans le cas des pratiques informationnelles, le problème de l'autorité épistémique des chatbots reste entier : si les usagers sont influencés par la quantité et la crédibilité apparente des informations fournies, cette crédibilité n'est – justement – qu'apparente, et rien ne garantit la fiabilité des informations en question. Au contraire, il existerait même une corrélation entre la quantité des informations fournies et leur caractère douteux. Deux études parues dans *Nature* et *Sciences* en décembre 2025[17] ont ainsi révélé que plus les LLM étaient entraînés à convaincre, plus ils produisaient une quantité importante d'informations. Mais, de manière paradoxale, plus ils produisaient d'informations, et moins ces dernières étaient fiables, l'objectif de convaincre semblant primer sur celui de produire des données factuelles. L'étude parue dans *Nature* soulignait même une asymétrie politique : les modèles d'IA entraînés à défendre des positions de droite produisaient plus d'affirmations inexactes que ceux entraînés à défendre des positions de gauche. Cette asymétrie s'expliquerait, selon les auteurs, par le fait que les modèles sont en partie entraînés avec les données présentes sur le web, où l'on retrouve beaucoup plus de fausses informations marquées à droite qu'à gauche (dans la mesure où les « fake news » ont été mobilisées comme outils de propagande par différents partis et mouvements d'extrême-droite lors de plusieurs campagnes électorales entre 2016 et aujourd'hui[18]).

Au-delà de la quantité et de la fiabilité des informations fournies, se pose également la question de leur sélection. Lorsqu'un internaute interroge un chatbot, les réponses pouvant lui être fournies sont potentiellement nombreuses. Le chatbot doit alors opérer une sélection entre ce qu'il montre et ce qu'il cache. Une bonne manière de s'en rendre compte est de s'intéresser non pas à l'opinion telle qu'elle se forme, mais telle qu'elle s'exprime. Dans une expérience réalisée en 2023[19], des chercheurs et chercheuses ont calibré des chatbots pour qu'ils aident des individus à produire des textes d'opinion sur le sujet des apports et limites des réseaux sociaux pour la société.

---

[16] Tessler, M. H., Bakker, M. A., Jarrett, D., Sheahan, H., Chadwick, M. J., Koster, R., Evans, G., Campbell-Gillingham, L., Collins, T., Parkes, D. C., Botvinick, M., & Summerfield, C. (2024). AI can help humans find common ground in democratic deliberation. *Science*, *386*(6719).

[17] Hackenburg et al., art. cit ; Lin et al., art. cit.

[18] Voir par exemple Benkler, Y., Faris, R., Roberts, H. (2018). *Network propaganda. Manipulation, disinformation and radicalization in americain politics*. Oxford University Press.

[19] Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L., & Naaman, M. (2023). Co-Writing with Opinionated Language Models Affects Users' Views. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1-15.

Certains LLM devaient promouvoir des arguments « pour » (mettant en avant les apports des réseaux sociaux), d'autres « contre » (insistant sur leurs méfaits). Résultat : en fonction du modèle utilisé, les arguments « pour » ou « contre » avaient deux fois plus de chance de se retrouver dans les textes produits par les humains que pour le groupe contrôle (qui avait accès à un LLM « neutre »). Mais cette influence sur l'écriture ne relevait pas d'une acceptation passive, puisque seulement 11,5% des textes étaient des copier-coller des réponses des modèles. Autrement dit, dans leur activité d'écriture, les participants semblaient intégrer et s'aligner sur les propositions du modèle, ce que confirmaient des sondages post-expérience. Les auteurs et autrices de l'étude parlent ainsi de « persuasion latente » pour décrire la manière dont l'IA générative, utilisée comme assistant de rédaction, influence les opinions telles qu'elles s'expriment, en fonction des types d'argument qu'elle rend disponibles pour nourrir la réflexion des internautes.

Luttes d'influence et sûreté de l'IA

Résumons : si les LLM ont le pouvoir de nous influencer, c'est dans la manière dont ils trient et organisent les informations qu'ils mettent à notre disposition. Mais où puisent-ils ces informations ? Principalement dans leurs données d'entraînement et sur le web. Un processus qui n'est pas passé inaperçu aux yeux de certains acteurs malveillants qui cherchent à détourner les chatbots d'IA générative en outils d'influence. En 2024, Viginum, le service de l'État français chargé de surveiller les ingérences étrangères, a par exemple révélé l'existence d'un réseau de 193 sites internet[20] qui relayaient en ligne la vision du Kremlin concernant la guerre en Ukraine. Ce réseau de sites a d'abord laissé les observateurs circonspects : quelle portée pouvait-il bien avoir, alors même que les audiences de chacun de ces sites demeuraient très modestes ? Les enquêteurs de Viginum n'ont pas tardé à trouver la réponse : leur véritable objectif n'était pas de générer de l'audience, ni même de produire des contenus destinés à inonder les réseaux sociaux. En produisant ou en partageant massivement des contenus pro-russes, ces sites avaient pour but de devenir des sources d'information pour les chatbots d'IA lorsqu'ils répondaient à des requêtes d'internautes sur la guerre en Ukraine. Une stratégie payante, puisque l'organisation

[20] VIGINUM. (2024). *Portal Kombat : Un réseau structuré et coordonné de propagande prorusse*.

NewsGuard a révélé en 2025 avoir retrouvé 92 articles publiés par ces sites dans les réponses produites par les principaux chatbots d'IA générative sur ce thème.

Les techniques relatives à ce que l'on appelle dorénavant l'*AI Poisoning* (empoisonnement des IA) ne sont pas sans rappeler là encore les débats qui avaient cours il y a dix ou quinze ans concernant les moteurs de recherche, et la manière dont de faux liens hypertextes servaient à faire remonter artificiellement certains sites politiques dans les résultats de Google et de ses concurrents. Mais ici, la puissance de frappe permise par les IA génératives en soutien à ces opérations inquiète. À l'été 2026, un mystérieux think tank américain, le Hanover Institute for Public Policy, a défrayé la chronique. *The Guardian*[21] a révélé que le site de l'institut avait produit 124 rapports entre le 6 et le 14 août, qui contenaient essentiellement des réponses courtes à des questions susceptibles d'être posées à des chatbots concernant la guerre à Gaza. Là encore, l'objectif de l'institut (par ailleurs fictif) n'était pas que ces rapports soient lus par des humains, mais qu'ils soient utilisés comme des ressources par les chatbots d'IA générative, soit lors de la phase d'entraînement des modèles, soit comme source des réponses fournies à leurs usagers. Cette opération (financée par le gouvernement israélien) aurait ainsi eu pour objectif final de redorer le blason d'Israël aux États-Unis, où un divorce historique semble se prononcer avec l'opinion publique américaine.

Ces opérations d'influence sont particulièrement difficiles à contrer pour les industriels de l'IA qui mettent ces chatbots sur le marché. De manière assez surprenante en effet, il suffit d'un petit nombre d'informations trompeuses pour conditionner les réponses d'un LLM, alors même que ces derniers ingurgitent des milliards de publications lors de leur phase d'entraînement. Dans une étude publiée par Anthropic en 2025, 250 documents corrompus, qui contenaient une instruction cachée devant déclencher une réponse-type du modèle, avaient suffi à modifier son comportement. Un chiffre qui restait d'ailleurs étonnamment stable, quelle que soit la taille du modèle testé[22]. L'étude en question semblait ainsi accréditer l'idée qu'un petit nombre de sites coordonnés peut suffire à influencer les réponses fournies par les chatbots aux internautes.

[21] Wilson, J. (2026). Fake Us thinktank set up and funded by Israel sought to game AI for propaganda. *The Guardian*.

[22] Anthropic. (2025). A small number of samples can poison LLMs of any size.

ChatGPT, dis-moi pour qui voter

Plutôt que d'agir sur les données d'entraînement des modèles, les industriels cherchent à paramétrer leur comportement via des techniques d'alignement. Dans le monde de l'IA, l'« alignement » désigne ainsi un ensemble de pratiques, de méthodes et de dispositifs qui sont mis en œuvre lors de la phase de « *fine tuning* »[23] afin de s'assurer que le comportement des modèles respecte un certain nombre de principes et de valeurs. Ces pratiques relèvent autant d'une ingénierie technique que de questionnements éthiques, tant ces principes et valeurs varient suivant les aires géographiques et les époques. Le principe de « neutralité » par exemple, est impossible à atteindre pour un chatbot d'IA. D'abord parce que les modèles, via les données qu'ils assimilent, incorporent des représentations sociales qui reflètent les cultures humaines qui ont produit ces données. Interrogez ChatGPT sur des personnalités politiques en anglais, en arabe ou en chinois, et les réponses varieront, tout simplement parce que les publications que le modèle aura utilisées pour apprendre ces langues refléteront des valeurs et principes différents (plus ou moins progressistes ou conservateurs, démocrates ou autoritaires, occidentalo- ou orientalo-centrés, etc.)[24]. De la même façon, les modèles incorporent des préjugés sociaux lors de leur phase d'entraînement, qu'ils répercutent ensuite dans les réponses fournies aux internautes.

Tout l'enjeu, pour les ingénieurs qui calibrent les comportements des modèles, est de savoir comment rendre « opérationnels » des principes éthiques, comment les sélectionner, et qui dispose de la légitimer pour les choisir. Concernant les choix politiques par exemple, les modèles d'OpenAI sur lesquels repose ChatGPT sont entraînés à respecter un principe de pluralisme25, qui se matérialise par l'instruction qui est faite au modèle d'exposer différents points de vue sur les sujets controversés. Si la neutralité « pure » n'est pas atteignable, le modèle est entraîné à ne pas prendre position sur les sujets politiques, à ne pas porter de jugement moral dans ses réponses, et à éviter de mobiliser un vocabulaire subjectif. Enfin, ChatGPT est tenu à une exigence de factualité, et est censé rapporter des faits, et mettre en contexte ses réponses. Pour s'assurer que le modèle respecte ces principes, des annotateurs

[23] Réglage fin : après la phase d'entrainement pendant laquelle les modèles sont entrainés à produire du texte selon des logiques de prédiction statistique, le réglage fin consiste à paramétrer les comportements attendus du modèle vis-à-vis de ses usagers.

[24] Buyl, M., Rogiers, A., Noels, S., Bied, G., Dominguez-Catena, I., Heiter, E., Johary, I., Mara, A.-C., Romero, R., Lijffijt, J., & De Bie, T. (2026). Large language models reflect the ideology of their creators. *Npj Artificial Intelligence*, *2*(1), 7.

[25] OpenAI. (2026). *OpenAI Model Spec*.

humains (remplacés dorénavant par d'autres LLM) évaluent les réponses fournies par un modèle afin de nourrir un système de récompenses censé cadrer son comportement lors de son déploiement. Dans leur déploiement, ces pratiques de paramétrage rencontrent un certain nombre de limites, notamment quand elles entrent en contradiction avec le principe supérieur qui guide le comportement des chatbots : être utiles à leurs usagers en répondant à leurs questions. Un récent rapport de la fondation Jean Jaurès a ainsi montré comment ChatGPT et Perplexity pouvaient faire preuve d'une certaine autorité dans leurs conseils, et exprimer des positions de manière catégorique[26].

Face aux risques sociaux posés par les usages des grands modèles de langage, les régulateurs déploient des méthodes d'audit « en boîte noire » pour se doter de leurs propres mesures et indicateurs quant aux comportements des modèles. En France, le Peren (Pôle d'expertise de la régulation numérique), Viginum (le service de vigilance et de protection contre les ingérences numériques étrangères) ou l'Anssi (Agence nationale de la sécurité des systèmes d'information) incarnent cette volonté de l'État de s'équiper techniquement et de monter en compétence sur le plan numérique. Assistées par des chercheurs et chercheuses en sciences du numérique, en sciences des données et en sciences sociales, ces agences surveillent à la fois le design et les usages des modèles d'IA afin de détecter des opérations malveillantes, ou des défauts de conception problématiques. De la même manière, la Commission Européenne a créé un Centre européen pour la transparence algorithmique (ECAT), dans le sillage du Règlement sur les Services Numériques, afin de se doter de ses propres méthodes d'audit des plateformes.

Aujourd'hui, la question de l'influence des modèles d'IA sur nos opinions dépasse largement l'élection présidentielle et les risques d'ingérence étrangère. Elle se loge dans nos rapports ordinaires et quotidiens avec des outils qui deviennent des « compagnons », et avec lesquels les plus jeunes d'entre nous découvrent la vie politique. Ce qui est rendu visible (ou au contraire invisible) dans leurs réponses à nos questions est un enjeu démocratique majeur, dont nous commençons tout juste à saisir la teneur. Si cette question s'est posée à toutes les époques (tous les médias produisent de la visibilité et de l'invisibilité), la protection de notre attention collective passe par

[26] Bourgoin-Verdier, T. (2026). Ce que l'intelligence artificielle dit de la campagne présidentielle. Fondation Jean Jaurès. 28/09/2026.

la maitrise des principes et procédures par lesquels celle-ci est aiguillée. Une condition qui semble loin d'être remplie concernant l'IA générative.

Publié dans laviedesidees.fr, le 29 septembre 2026.