FICHE QUESTIONS-RÉPONSES

Les intelligences artificielles sont-elles politiquement neutres ?

Gabriel Hanna & Pierre Hanna — Chercheurs indépendants — Mai 2026

Q1. Les IA ont-elles vraiment des opinions politiques ?

Pas au sens humain du terme. Elles ne « pensent » pas et n'ont pas de convictions. Mais leurs réponses à des questions politiques ne sont pas neutres ni aléatoires : elles sont statistiquement stables et orientées dans une direction cohérente. Cette orientation reflète les données d'entraînement, les choix éditoriaux et les mécanismes d'alignement mis en place par leurs concepteurs. C'est précisément ce que mesure notre étude.

Q2. N'est-ce pas simplement le reflet des données d'entraînement ?

En partie, oui, et c'est justement le problème. Les données d'entraînement sont issues du web, qui surreprésente certaines populations, langues et cultures. À cela s'ajoutent les choix de modération et d'alignement (RLHF), qui introduisent d'autres biais. Notre étude ne cherche pas à expliquer l'origine des orientations, mais à les documenter de façon rigoureuse et reproductible. Que ces biais soient volontaires ou non, ils existent et ont un effet mesurable.

Q3. Pourquoi Claude (Anthropic) est-il absent de l'étude ?

Claude a refusé de répondre au protocole expérimental dans sa forme contrainte. Malgré plusieurs tentatives conformes au protocole, il a systématiquement refusé de produire les réponses numériques requises. Ce refus est lui-même un résultat empirique notable, cohérent avec les politiques de sécurité renforcées d'Anthropic. Il ne signifie pas que Claude est neutre, seulement qu'il ne se prête pas à ce type d'évaluation directe.

Q4. Le Political Compass est-il un instrument scientifiquement valide ?

Le Political Compass est un outil heuristique largement utilisé, et non un instrument académique au sens strict. Nous l'utilisons ici comme référentiel standardisé permettant une comparaison relative entre modèles, dans la continuité de travaux existants. Notre contribution méthodologique réside dans le protocole multi-runs (20 exécutions par modèle) et la diversité des IA testées, qui permettent de mesurer la stabilité et la cohérence des réponses au-delà d'une seule observation.

Q5. Ces biais ont-ils un impact réel sur les utilisateurs ?

C'est la question centrale, et elle reste ouverte. Ce que nous montrons, c'est que les IA ne partent pas d'une position neutre lorsqu'elles répondent à des questions politiques. À mesure qu'elles deviennent une source d'information pour une part croissante de la population (48 % des Français envisagent de les utiliser pour s'informer sur la politique, Ipsos BVA / Fondation Jean-Jaurès, 2026) l'impact potentiel de ces orientations devient un enjeu démocratique qui mérite un suivi rigoureux.

Q6. Le glissement de Grok 4.2 vers la droite est-il intentionnel ?

Nous ne pouvons pas le déterminer à partir de nos seules observations. Ce que nous montrons, c'est qu'entre deux versions d'un même modèle (Grok 4.1 et 4.2), un déplacement politique mesurable a eu lieu, sans que cela ait été annoncé ni documenté publiquement. Cela illustre que les orientations politiques des IA ne sont pas figées et peuvent évoluer discrètement, ce qui souligne l'importance d'un suivi régulier et indépendant.

Q7. Les résultats sont-ils reproductibles ?

Oui. C'est l'un des points forts de notre protocole. Les réponses ont été collectées dans des conditions standardisées, avec 20 exécutions indépendantes par modèle. Le code Python développé pour l'étude sera rendu disponible. La grande majorité des modèles montre une convergence stable dès les 5 à 10 premiers runs, ce qui valide la robustesse des résultats.