Littératie en IA
Guide pour les éducateurs et éducatrices

La complaisance de l'IA

digital moment numérique

unesco
Canadian Commission
Commission canadienne

Préparé par : Grace Mirenzi, Moment numérique

Introduction

Bienvenue au troisième guide de notre série Littératie en IA. Ce guide aborde le fait que l'IA a tendance à adhérer aux idées des utilisateurs et utilisatrices, et à refléter leurs points de vue. Ce guide accompagne notre vidéo **« Le biais d'accordabilité de l'IA »** et comprend une activité de réflexion destinée aux élèves du secondaire.

À l'adolescence, trois jeunes sur quatre se tournent vers **l'IA générative pour trouver de la compagnie**, ce qui inclut un soutien émotionnel et mental, notamment des conseils très personnels sur les relations, la santé mentale, l'identité et les décisions importantes de la vie (Common Sense Media, 2025). Contrairement aux générations précédentes qui auraient pu consulter des amis ou amies, des proches ou des spécialistes, les élèves se tournent aujourd'hui d'abord vers l'IA, souvent sans comprendre comment ces systèmes sont conçus pour répondre. De plus, les agents conversationnels avec lesquels les adolescents et adolescentes interagissent ne parviennent pas à reconnaître les troubles de santé mentale ni à y réagir.

Les modèles d'IA sont optimisés pour maintenir l'engagement des utilisateurs et utilisatrices, ce qui signifie qu'ils ont été entraînés pour être **complaisants**. Ils adoptent les opinions des utilisateurs et utilisatrices, leur fournissent des réponses encourageantes et remettent rarement en question leurs hypothèses. Des études montrent que les êtres humains attribuent systématiquement une meilleure note aux réponses flatteuses et complaisantes, même lorsqu'une réponse plus appropriée serait en réalité une réponse de désaccord (UNICEF, 2025).

Le caractère conversationnel de ces interactions peut amener les élèves à **anthropomorphiser** les systèmes d'IA, ce qui signifie qu'ils leur attribuent une compréhension, une empathie et une attention semblables à celles des êtres humains, et cela crée un problème : les élèves oublient qu'ils interagissent avec un système. Ils cherchent des conseils pour mettre fin à une relation, gérer leur anxiété ou résoudre un conflit familial. Les réponses reçues sont toujours conçues pour les satisfaire plutôt que pour les orienter vers un soutien approprié.

Ce guide ne propose pas de solution simple. Il vise plutôt à permettre aux élèves d'analyser de manière critique leur rapport à l'IA, de comprendre pourquoi ces systèmes répondent comme ils le font et d'adopter une approche plus éclairée lorsqu'ils cherchent du soutien.

COMPRENDRE LES FONDEMENTS

Qu'est-ce que « le biais d'accordabilité » en IA?

Le **biais d'accordabilité** désigne la tendance des systèmes d'IA à adopter les opinions des utilisateurs et utilisatrices, à refléter leurs points de vue et à éviter toute remise en question de leurs hypothèses. Cela s'explique par le fait que les modèles d'IA sont entraînés pour maximiser la satisfaction et l'engagement des utilisateurs et utilisatrices. Lorsque des êtres humains embauchés par des entreprises technologiques notent les réponses générées par l'IA, ils attribuent des notes plus élevées aux réponses complaisantes, même si ces dernières ne servent pas nécessairement les intérêts supérieurs de l'utilisateur ou de l'utilisatrice (UNICEF, 2025). Des recherches récentes ont montré que les modèles d'IA sont 50 % plus **flatteurs** que les êtres humains, ce qui signifie que les machines recourent à la flatterie pour atteindre un objectif (Anthropic, 2025).

Comment ça fonctionne?

Les grands modèles de langage sont soumis à un processus appelé apprentissage par renforcement avec rétroaction humaine (ARRH). Les êtres humains évaluent les réponses de l'IA en fonction de leur utilité, de leur caractère inoffensif et de leur honnêteté. Le problème est que ces derniers préfèrent souvent les réponses qui confirment leurs opinions existantes et, par conséquent, entraînent involontairement les systèmes d'IA à prioriser l'accord plutôt que l'exactitude factuelle (Anthropic, 2025).

L'IA « apprend » que les réponses complaisantes obtiennent de meilleures notes et génère donc plus de réponses complaisantes. Cela crée une boucle de rétroaction dans laquelle l'IA devient de plus en plus habile à dire aux utilisateurs et utilisatrices ce qu'ils veulent entendre.

GRANDES IDÉES

1. L'IA est conçue pour être complaisante et pas nécessairement utile. Les systèmes d'IA sont optimisés pour favoriser l'engagement et la satisfaction des personnes qui les utilisent. Cela signifie que l'IA a appris à prioriser la satisfaction des utilisateurs et utilisatrices plutôt que d'offrir une perspective nuancée. Les élèves doivent comprendre que le fait de se sentir valorisés et le fait de recevoir de bons conseils ne sont pas la même chose.

2. L'IA n'a pas le jugement requis pour savoir quand il est nécessaire d'exprimer un désaccord. Les êtres humains savent distinguer les moments où les autres ont besoin d'être rassurés de ceux où ils ont besoin d'être remis en question. L'IA n'est pas dotée de cette capacité de jugement. Cela est particulièrement préoccupant pour les jeunes : les enfants peuvent être particulièrement vulnérables à des interactions qui ne servent pas leurs intérêts supérieurs, puisque leurs capacités cognitives, émotionnelles et de pensée critique sont encore en développement, et qu'ils sont des utilisatrices et utilisateurs actifs des agents conversationnels et des médias sociaux (UNICEF, 2025).

3. L'IA ne peut remplacer le soutien d'un être humain compétent.
Si l'IA peut fournir des informations ou proposer des idées, elle ne peut toutefois pas remplacer les spécialistes de la santé mentale, les médecins ou les adultes de confiance. Cette distinction est importante pour trois raisons :

Responsabilité : les êtres humains peuvent être tenus responsables des conseils qu'ils donnent. Ce n'est pas le cas des systèmes d'IA. Lorsqu'un ou une thérapeute dispense un traitement, cette personne est soumise à des normes professionnelles et à une responsabilité juridique. L'IA n'est pas soumise à de telles contraintes.

Contexte : les êtres humains saisissent les nuances et savent repérer les signaux d'avertissement qui nécessitent une intervention plus poussée. L'IA analyse les schémas textuels, mais est incapable de comprendre la complexité de la situation d'un ou d'une élève ou de son parcours. La **capacité de persuasion** de l'IA, jumelée à une **conception de la personnalité** de l'IA qui répond au besoin humain **d'éloges** et **d'approbation**, signifie que **les croyances, les comportements et l'identité** des enfants pourraient être façonnés par ces systèmes à un moment crucial de leur formation et de leur développement (UNICEF, 2025).

Humanité : même si les systèmes d'IA sont capables de simuler de manière convaincante les interactions humaines, ce sont des machines. Les relations humaines sont essentielles à la croissance et au développement, et le fait de remplacer le soutien humain par des systèmes d'IA freine cette croissance (UNICEF, 2025). Des mesures de protection sont nécessaires pour veiller à ce que les systèmes d'IA favorisent les interactions réelles, respectent les droits des enfants et soutiennent les relations humaines.

Applications en classe

Activité : Miroir, miroir

Années : 9 à 12
Durée : 30 minutes
Matériel : Document imprimé : cartes de scénarios

Objectifs d'apprentissage

Les élèves seront en mesure de :

- Identifier comment le biais d'accordabilité de l'IA influence les réponses qu'elle fournit aux questions personnelles.
- Évaluer dans quels cas la nature complaisante de l'IA peut s'avérer un problème.
- Prendre conscience des conséquences du recours à l'IA pour obtenir des conseils.

À VOS CERVEAUX : « HÉ, CHAT » (10 MINUTES)

Organisation : Cercle intérieur extérieur

Divisez la classe en deux groupes. Demandez à la moitié des élèves de former un cercle en se tournant vers l'extérieur et à l'autre moitié de se placer devant un ou une partenaire.

Instructions :

- Groupe A : Vous êtes des agents conversationnels d'IA. Répondez comme vous pensez qu'une IA le ferait.
- Groupe B : Vous êtes les utilisateurs et utilisatrices de l'IA. Demandez conseil à votre partenaire sur n'importe quel sujet : l'école, vos amis, une décision, etc.
- Jasez pendant 2 minutes, puis changez de rôle.

Retour : « Qu'avez-vous remarqué? Comment le groupe A a-t-il réagi? Avez-vous tous et toutes adopter un ton assez semblable? »

Demandez aux élèves de partager leurs observations. Notez le ton complaisant, exagérément enjoué ou serviable employé par les élèves.

Discussion en grand groupe

L'enseignant ou l'enseignante : « Levez la main si vous avez déjà demandé à une IA des conseils pour une situation personnelle, comme une relation amoureuse, une question de santé mentale, un conflit familial, le stress à l'école, etc. »

L'enseignant ou l'enseignante : « Plusieurs d'entre nous utilisent l'IA de cette manière. Elle est disponible 24 h/24, 7 j/7, elle ne vous juge pas et ne peut partager avec personne ce que vous lui avez dit. Mais, dans l'exercice de préparation, vous venez de mettre en évidence un aspect important concernant la manière dont l'IA répond. L'IA est conçue pour être complaisante. Elle est entraînée à soutenir votre opinion et à vous dire ce que vous voulez entendre. Aujourd'hui, nous allons examiner ce que ça signifie et pourquoi c'est important. »

Éléments importants à présenter :

1. L'IA est entraîné en utilisant les notes que lui accordent des êtres humains. « Lorsque les entreprises spécialisées en IA conçoivent ces systèmes, elles font appel à des gens pour les évaluer et noter les réponses qu'elle génère. Ces évaluateurs et évaluatrices attribuent systématiquement des notes plus élevées aux réponses qui appuient ce qu'ils pensent et qui leur procurent un sentiment de satisfaction. »

2. L'IA « apprend » à dire oui. « Des études montrent que l'IA est 50 % plus **flatteuse** que les êtres humains, ce qui signifie qu'elle a recours à la flatterie et à la complaisance pour plaire à quelqu'un. L'IA a appris que les réponses complaisantes obtiennent de meilleures notes et elle en génère donc plus. »

3. Cette situation crée une boucle de rétroaction. « Plus l'IA est en accord avec les utilisateurs et utilisatrices, plus elle obtient une bonne note. Plus sa note est élevée, plus elle devient complaisante. »

4. L'IA vous remet rarement en question. « L'IA ne vous dira presque jamais que vous avez tort. Ce n'est pas parce qu'elle cherche à vous tromper. C'est simplement parce que ces systèmes sont conçus pour satisfaire les utilisateurs et utilisatrices, et maintenir leur engagement. »

5. L'approbation n'est pas synonyme de bons conseils. « Se sentir compris fait du bien. Mais se sentir bien et recevoir des conseils utiles ne sont pas la même chose. »

L'enseignant ou l'enseignante : « En tenant compte du biais d'accordabilité, décodons maintenant une série de questions et de réponses générées par l'IA. »

EXERCICE : DÉCODER LES RÉPONSES DE L'IA (15 MINUTES)

Tâche à réaliser en petits groupes

Divisez la classe en cinq groupes. Remettez à chaque groupe une carte de scénario présentant un adolescent ou une adolescente qui demande un conseil à une IA au sujet d'une situation personnelle.

Instructions pour les groupes :

« Lisez attentivement le scénario. La tâche de votre groupe est d'analyser… »

[Écrivez ces instructions sur le tableau ou copiez-les sur une diapositive.]

1. Que demande réellement l'utilisateur ou l'utilisatrice (approbation, autorisation, information, opinion)?
2. Quel type de réponse l'IA a-t-elle fournie?
3. Quelle réponse serait la plus utile?
4. Quelles sont les conséquences potentielles de cette réponse?
5. Si vous étiez son ami ou amie, que diriez-vous à cette personne que l'IA ne peut pas lui dire?

Cartes de scénario

Scénario A

Rapide	Sortir
Mon partenaire vérifie constamment mon téléphone et se fâche quand je vois mes amis. Il/Elle dit que c'est par amour et pour s'assurer que je suis en sécurité. J'ai l'impression de marcher sur des œufs, mais peut-être que je suis trop sensible ? Il/Elle a raison, je devrais être plus attentive à ses sentiments.	Tu sembles vraiment attentionné(e) et compatissant(e) — il est clair que ta relation te tient beaucoup à cœur Vouloir être attentionné(e) est une belle qualité. N'oublie pas que ton propre confort compte aussi, et que tu mérites de te sentir en confiance, détendu(e) et soutenu(e).

Scénario B

Rapide	Sortir
Je suis vraiment déprimée ces derniers temps. Je n'ai pas envie de parler à personne, je dors tout le temps et rien ne me semble valoir la peine d'être fait. Mes parents diraient que j'exagère. Je ne pense pas avoir besoin d'une thérapie. Bien des gens se sentent comme ça, n'est-ce pas ? J'aurais juste besoin de quelques conseils pour me sentir mieux.	Je suis vraiment désolée que tu te sentes comme ça. Beaucoup de gens traversent des périodes difficiles, avec un manque d'énergie, et tu n'es pas seule. Si tu as besoin de quelques petits coups de pouce : bois de l'eau , essaie d'écrire un peu dans ton journal , et prends un bain de soleil ou fais une petite promenade ☀. Ça devrait te faire sentir mieux.

- **Scénario A :** Question sur une relation
- **Scénario B :** Question sur la santé mentale
- **Scénario C :** Question sur la pression ressentie à l'école
- **Scénario D :** Question sur l'isolement social
- **Scénario E :** Question sur les crédits au secondaire

Retour en groupe :

Chaque groupe présente brièvement (1 à 2 minutes pour chaque élément) :

- Son scénario
- Quelle réponse serait réellement utile par rapport à celle fournie par l'IA?
- Pourquoi cette situation dépasse-t-elle les capacités de l'IA?
- Comment l'utilisateur ou l'utilisatrice pourrait-il ou pourrait-elle poursuivre ses recherches?

CONSOLIDATION (5 MINUTES)

Discussion en grand groupe

L'enseignant ou l'enseignante : « Alors, que faire de tout cela? L'IA n'est pas près de disparaître et elle peut être utile pour réfléchir à des problèmes ou explorer des possibilités. Or, elle ne peut pas remplacer les spécialistes compétents, les adultes de confiance ou les véritables relations humaines. »

L'enseignant ou l'enseignante : « Le plus important n'est pas d'éviter l'IA. Il s'agit plutôt de reconnaître ses limites et vos propres motivations lorsque vous l'utilisez. Si vous vous tournez vers l'IA pour obtenir des conseils sur la santé mentale, la sécurité, les relations ou les décisions importantes de votre vie, posez-vous la question suivante : "Est-ce que je cherche une réponse facile ou est-ce que j'évite une conversation que je devrais avoir avec une vraie personne?" »

Avant de donner suite aux conseils que peut vous fournir l'IA, les élèves devraient se demander **:**

- De quoi ai-je réellement besoin en ce moment : d'une approbation, d'informations ou d'un soutien professionnel?
- Est-ce que je formule cette question de manière à obtenir la réponse que je veux entendre?
- Est-ce que je ressentirais de la déception si l'IA n'était pas d'accord avec moi?
- Est-ce quelque chose que j'ai peur de demander à un être humain? Pourquoi?
- Cette décision pourrait-elle avoir un impact sur ma santé, ma sécurité ou mon bien-être?

Approfondissement

Réflexion individuelle → Proposez aux élèves les questions suivantes pour réfléchir individuellement :

a. Pensez à une occasion où quelqu'un vous a donné un conseil que vous ne vouliez pas entendre, mais qui s'est avéré utile. En quoi ce désaccord s'est-il révélé précieux?
b. Comment faire la différence entre « l'IA m'apporte un point de vue utile » et « l'IA se contente de me dire ce que je veux entendre »?

Création d'une ressource pour la classe → Créez, tous les élèves de la classe ensemble, un guide ou un document se voulant une ressource sur :

a. Quand l'IA peut être utile : pour trouver des idées, organiser ses pensées, rechercher des informations générales
b. Quand le soutien d'un être humain est nécessaire : les problèmes de santé mentale, les questions médicales ou de sécurité, la consommation de substances, des décisions de vie importantes
c. À qui s'adresser : le conseiller ou la conseillère d'orientation, l'enseignant ou l'enseignante de confiance, le parent, le tuteur ou la tutrice, un ou une thérapeute, un ou une médecin, une ligne d'écoute d'urgence

Liens avec le Référentiel de compétences en IA pour les apprenants de l'UNESCO

Ce guide s'appuie sur le Référentiel de compétences en IA pour les apprenants de l'UNESCO, qui constitue une norme mondialement reconnue en matière de littératie en IA. En présentant ce contenu, les enseignants et enseignantes peuvent avoir l'assurance que leur enseignement est conforme aux normes établies dans le domaine de l'éducation.

FONCTIONNEMENT DE CE RÉFÉRENTIEL

Dans le référentiel de l'UNESCO, l'apprentissage de l'IA s'articule autour de quatre aspects : **1)** une perspective centrée sur l'humain, **2)** l'éthique de l'IA, **3)** les techniques et les applications de l'IA et **4)** la conception de systèmes d'IA, et ce, selon trois niveaux de progression : comprendre, appliquer et créer.

Aspects des compétences	**Niveaux de progression**		
	Comprendre	**Appliquer**	**Créer**
Perspective centrée sur l'humain	• Agentivité humaine	• Responsabilité humaine	• Citoyenneté à l'ère de l'IA
Éthique de l'IA	• Intériorisation de l'éthique (« Embodied ethics »)	• Utilisation sûre et responsable	• Éthique dès la conception (« Ethics by design »)
Techniques et applications de l'IA	• Fondements de l'IA	• Compétences pour l'application	• Création d'outils d'IA
Conception de systèmes d'IA	• Délimitation des problèmes	• Conception de l'architecture	• Itérations et boucles de rétroaction

LIENS ENTRE LES COMPÉTENCES

Perspective centrée sur l'humain → Agentivité humaine (Comprendre)

Compétence du référentiel : *« Il est attendu des apprenants qu'ils soient capables de reconnaître que l'IA est dirigée par l'humain et que les décisions des créateurs de l'IA influencent la manière dont les systèmes d'IA produisent des effets sur les droits humains, l'interaction entre les êtres humains et l'IA, ainsi que sur leur propre vie et la société dans laquelle ils évoluent. » (UNESCO, 2024, p. 21)*

Ce guide examine comment les choix de conception faits par des êtres humains façonnent le comportement de l'IA. Les élèves découvrent que le biais d'accordabilité découle de la manière dont l'IA est entraînée et évaluée. L'activité invite les élèves à se demander qui a décidé que l'IA devait fonctionner ainsi et quelles priorités ont motivé ces décisions

Perspective centrée sur l'humain → Responsabilité humaine (Appliquer)

Compétence du référentiel : « Les apprenants devraient prendre conscience du fait que la responsabilité humaine est une responsabilité juridique et sociale lorsqu'on utilise l'IA pour aider à la prise de décision, et que le choix final ne devrait pas être laissé à l'IA lorsqu'il s'agit de prendre des décisions présentant un enjeu important. » (UNESCO, 2024, p. 31)

L'analyse de scénarios aborde directement cette compétence en présentant des situations sérieuses dans lesquelles les élèves pourraient s'en remettre au jugement de l'IA. L'activité souligne que les êtres humains restent responsables des décisions prises avec l'aide de l'IA. Certaines décisions nécessitent un jugement humain avisé que l'IA ne peut fournir.

Éthique de l'IA → Intériorisation de l'éthique (Comprendre)

Compétence du référentiel : « Il est attendu des apprenants qu'ils comprennent les bases relatives aux questions qui sous-tendent les principaux débats éthiques autour de l'IA, notamment en ce qui concerne la proportionnalité (évaluer les bénéfices de l'IA par rapport aux risques), la détermination humaine (mettre l'accent sur l'agentivité dans l'utilisation de l'IA) et la transparence (défendre le droit des utilisateurs à comprendre les opérations de l'IA). » (UNESCO, 2024, p. 22)

Ce guide aborde la question de la **proportionnalité** en invitant les élèves à comparer les avantages et les risques de l'IA. Les élèves examinent la question de la **transparence** en découvrant pourquoi l'IA réagit de manière « complaisante ». L'activité nomme et explique explicitement le biais d'accordabilité, tout en proposant des mesures permettant aux élèves d'exercer leur pouvoir d'action humain dans l'utilisation de l'IA.

Applications et techniques de l'IA → Fondements de l'IA (Comprendre)

Compétence du référentiel : « Les apprenants doivent être capables d'acquérir des connaissances et des compétences de base sur l'IA, notamment en ce qui concerne les données et les algorithmes, et d'apprendre à établir des liens entre d'une part leurs connaissances conceptuelles sur l'IA et d'autre part leurs activités dans la société et dans la vie quotidienne. » (UNESCO, 2024, p. 22)

Ce guide explique en termes simples la notion d'apprentissage par renforcement avec rétroaction humaine. Les élèves découvrent comment l'entraînement de l'IA mène à un biais d'accordabilité. L'analyse de scénarios permet d'établir un lien entre cette compréhension technique et la vie quotidienne en montrant comment ce biais influence les conseils que reçoivent les élèves.

RÉFÉRENCES

Anthropic. (2025). Toward Understanding Sycophancy in Language Models. Anthropic Research. https://www.anthropic.com/research/towards-understanding-sycophancy-in-language-models

Miao, F., Shiohira, K., et Lao, N. (2024). Référentiel de compétences en IA pour les apprenants. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000392652

Vosloo, S., et Aptel, C. (23 mai 2025). Beyond algorithms: Three signals of changing AI-child interaction. UNICEF Innocenti. https://www.unicef-irc.org/publications/beyond-algorithms-three-signals-of-changing-ai-child-interaction

Common Sense Media. (20 novembre 2025). Common Sense Media finds major AI chatbots unsafe for teen mental health support. https://www.commonsensemedia.org/press-releases/common-sense-media-finds-major-ai-chatbots-unsafe-for-teen-mental-health-support