KOMPASS

INTELLIGENCE ARTIFICIELLE ET BIAIS

Fiche thématique

Intelligence artificielle et biais

Cette fiche thématique explique ce qu'est l'intelligence artificielle (IA) et met en lumière les biais qu'elle peut contenir. Elle s'adresse aux enseignants *de l'enseignement fondamental* et *secondaire* et fournit des informations de base pour aborder le sujet en classe.

Qu'est-ce que le biais ?

Le biais désigne le fait d'avoir une opinion préconçue (positive ou négative) à propos de personnes, de groupes ou de situations, sans les connaître réellement. On parle également de préjugé, car il s'agit de juger avant de disposer de toutes les informations.
Dans le domaine de l'IA, le biais correspond à des préférences, des distorsions ou des jugements injustes dans les résultats produits par un système d'IA. Ces biais peuvent être causés, entre autres, par des données incomplètes ou erronées, par les algorithmes eux-mêmes ou des préjugés humains intégrés lors du développement du système.

Quelles distorsions influencent les modèles d'IA ?

La distorsion signifie qu'un élément a été modifié ou déformé par rapport à son état initial, ou qu'il peut y avoir un écart par rapport à la réalité.
Comme les systèmes d'IA sont développés par des humains, leurs préjugés inconscients (*unconscious bias)* ou conscients sont souvent intégrés dans les modèles[1] . Cela peut se produire par exemple lors de la sélection et du traitement des données, de la définition des critères pertinents ou de l'ajustement des paramètres des modèles. Même des approches bien intentionnées peuvent, sans le vouloir, reproduire des perspectives et des hypothèses existantes et ainsi renforcer les inégalités. En conséquence, les données et les algorithmes reflètent souvent des biais humains.
Outre les **biais humains** (*human bias)*, d'autres types de biais peuvent influencer les modèles d'IA et leurs résultats. Voici les principaux :

Biais de données *(data bias)*

Un biais de données, également appelée biais des données d'entraînement, peut apparaître dès la phase de collecte ou de traitement des données. Si les ensembles de données ne sont pas représentatifs de la réalité, les modèles entraînés à partir de ceux-ci produisent des résultats biaisés.

Exemple : un système de reconnaissance vocale entraîné principalement avec des voix masculines reconnaît moins bien les voix féminines.

[1] En IA, les modèles désignent des structures mathématiques et statistiques construites qui sont entraînées à partir de grandes quantités de données afin d'accomplir certaines tâches telles que la reconnaissance de texte, la reconnaissance d'images ou les prévisions. Ils constituent l'élément central d'un système d'IA et influencent considérablement son comportement et ses résultats.

Biais algorithmique *(algorithmic bias)*

Un biais algorithmique apparaît lorsque le système favorise ou défavorise systématiquement certains groupes. Cela se produit notamment lorsqu'un algorithme apprend à partir d'anciens modèles ou d'inégalités, qu'il reproduit voire amplifie. Certains facteurs peuvent alors être surpondérés par rapport à d'autres, ce qui renforce des déséquilibres existants.
Ce type de biais est particulièrement problématique dans des domaines sensibles tels que les ressources humaines, l'octroi de crédits ou la justice pénale, où des décisions biaisées peuvent conduire à des discriminations.

Exemple : un algorithme d'évaluation de la solvabilité attribue systématiquement une note plus faible aux personnes provenant de certaines zones géographiques, parce que ces zones étaient plus souvent associées à des défauts de paiement dans les données d'entraînement, indépendamment de la solvabilité réelle des individus concernés.

Pourquoi les biais dans les systèmes d'IA posent-ils problème ?

Les biais présents dans les systèmes d'IA montrent que la technologie ne fonctionne pas indépendamment de la société. En effet, les développeurs d'IA définissent eux aussi, consciemment ou non, des critères influencés par des facteurs culturels, politiques ou économiques. Sans contrôle adéquat et sans utilisation réfléchie, les systèmes d'IA peuvent renforcer la discrimination, éroder la confiance du public et limiter considérablement le potentiel positif de la technologie.

Un système d'IA biaisé peut produire des effets discriminatoires dans de nombreux domaines, tels que l'évaluation de la solvabilité, la reconnaissance vocale (accents ou dialectes), les soins de santé, les poursuites pénales ou encore l'éducation.

Exemple dans le domaine de l'éducation : une université utilise des algorithmes d'évaluation et d'admission pour sélectionner les futurs étudiants. L'IA prédit que les candidats issus d'établissements scolaires bien équipés et favorisés auront davantage de succès. Les élèves provenant d'écoles moins favorisées sont ainsi désavantagés, ce qui peut accentuer les inégalités sociales et toucher en particulier les groupes déjà marginalisés.

Autre Exemple : un système d'IA est chargé de présélectionner les candidatures et a été entraîné à partir de données historiques dans lesquelles les postes techniques étaient majoritairement occupés par des hommes. L'IA privilégie alors automatiquement les candidatures à consonance masculine et classe moins bien celles des femmes, quelles que soient leurs qualifications réelles. Au lieu de réduire les discriminations existantes, le système les renforce.

Des systèmes d'IA conçus de manière équitable et équilibrée ont néanmoins le potentiel d'améliorer considérablement des domaines tels que la justice, la santé et l'éducation, à condition que leurs limites soient

reconnues et examinées de manière critique. Il est essentiel que l'humain reste au centre des décisions et garde le contrôle sur la technologie.

Les systèmes d'IA peuvent-ils être neutres ?

Une IA peut-elle être totalement exempte de préjugés, ou reflétera-t-elle toujours les biais humains ? Une neutralité absolue semble difficile à atteindre, car l'IA est un produit humain, influencé par des objectifs, des valeurs et des modes de pensée spécifiques. Toutefois, grâce à des ajustements techniques et à une supervision éthique, il est possible de réduire les biais. Parmi les méthodes de correction les plus importantes, on peut citer :

- **Ensembles de données larges et représentatifs** : les développeurs d'IA doivent toujours veiller à ce que les ensembles de données collectés soient aussi diversifiés et représentatifs que possible afin d'obtenir une représentation plus équilibrée de la réalité.
- **Procédures de test pour garantir l'équité** : il existe des outils permettant de vérifier l'équité des systèmes d'IA avant leur déploiement, par exemple AI Sandbox (*Luxembourg Institute for Science and Technology*, LIST).
- **IA explicable** (*Explainable AI - XAI*) : les techniques d'IA explicable permettent de rendre compréhensibles et explicables les décisions prises par les systèmes d'apprentissage automatique[2].

Sans ces approches, de nombreux modèles d'IA fonctionnent comme des boîtes noires (*blackbox)* dont les mécanismes internes restent opaques, même pour les développeurs. Cela complique la vérification des résultats et entraîne une perte de contrôle et de responsabilité. Grâce à l'IA explicable, les entreprises ont un aperçu des bases décisionnelles de l'IA et peuvent identifier et corriger les erreurs plus rapidement.

Ressources pour l'enseignement

Il existe plusieurs façons d'aborder le thème *de l'IA et des biais* en classe. Vous trouverez quelques exemples ici :

- *KI Kompass* - matériel didactique : https://ki-kompass.lu/praxis-materialien/
- *Coding for Tomorrow* propose une idée pédagogique portant sur l'identification des biais dans l'IA. Elle s'adresse principalement aux élèves à partir de 7ième : https://edulink.lu/ke2a
- L'*Académie de Nice* propose, dans son idée d'enseignement, une exploration des biais présents dans les images générées par l'IA. Elle s'adresse aux élèves à partir de 11 ans : https://edulink.lu/jal6
- Le *KI Campus* suggère dans sa démarche pédagogique des différentes pistes pour identifier des biais et stéréotypes dans les systèmes d'IA : https://edulink.lu/y8gu
- *School AI* présente plusieurs pistes pour intégrer l'IA dans des cours de l'enseignement fondamental : https://edulink.lu/6pyj

[2] L'apprentissage automatique (*machine learning*, ML) est un sous-domaine de l'intelligence artificielle. Les systèmes informatiques apprennent à partir d'exemples de données, reconnaissent des modèles et prennent des décisions sur cette base sans avoir été explicitement programmés pour cela.

- Le Ministère de la Culture de la République Française a développé l'outil de comparaison *Compar:IA* qui permet de comparer des outils et des jeux de données afin de sensibiliser à l'IA : https://edulink.lu/s0i5

Remarque : les enseignants doivent vérifier que les ressources sont adaptées à leur classe et à leur contexte et, le cas échéant, adapter leur utilisation.

Pour mettre en œuvre les idées pédagogiques présentées, il est recommandé d'utiliser les outils mis à disposition sur le site web KI Kompass.

Références

BEE SECURE (2025) : *Algorithmische Voreingenommenheit – Wenn Technik unsere Vorurteile lernt*. Disponible sur : https://www.bee-secure.lu/de/news/algorithmische-voreingenommenheit-wenn-technik-unsere-vorurteile-lernt/ (dernier accès : 27/01/2026).

Chapman University : *Bias in AI.* Disponible sur : https://www.chapman.edu/ai/bias-in-ai.aspx (dernier accès : 28/01/2026)

IBM : *KI-Verzerrungen anhand von Beispielen aus der realen Welt beleuchten.* Disponible sur : https://www.ibm.com/de-de/think/topics/shedding-light-on-ai-bias-with-real-world-examples (dernier accès : 17/01/2026)

SCRIPT (2026) : *Strategischer Rahmen zum Einsatz von Künstlicher Intelligenz in der Schule (Version vom 3. Februar 2026).* Disponible sur : https://ki-kompass.lu (dernier accès : 27/01/2026)

Luxembourg Institute of Science and Technology (LIST) : *LIST AI SANDBOX – LLM Observatory.* Disponible sur : https://ai-sandbox.list.lu/llm-leaderboard/ (dernier accès : 27/01/2026)

Université du Kansas – Center for Teacher Excellence : *Helping students understand the biases in generative AI.* Disponible sur : https://cte.ku.edu/addressing-bias-ai (dernier accès : 28/01/2026)