Séminaire Concepts de la computer ethics

Biais : un concept technosolutionniste ?

par Ambre Davat, post-doctorante TIMC/GRESEC
et membre de la chaire Ethique&IA

Citation recommandée : Ambre Davat, « Biais, intelligence artificielle et technosolutionnisme », partie de l'œuvre collective intitulée *Éthique, politique, religions. 2023 – 1. n° 22. L'éthique de l'intelligence artificielle à travers les dispositifs et les pouvoirs*, à paraître dans la collection « Éthique, politique, religions » (dirigée par Thierry Gontier)

Point de départ

Constat : concept de « biais » mobilisé dans de nombreux travaux en IA
mais dans des contextes extrêmement différents

Exemples classiques :

- classification d'images
  husky + neige => loup
- traitement du langage naturel
  « nurse » => infirmière ? infirmier ? infirmi* ?
- scandales « applications à haut risque »
  voiture à conduite automatisée
  médecine : erreurs de diagnostic
  justice : logiciel COMPAS
  RH : logiciel de recrutement Amazon

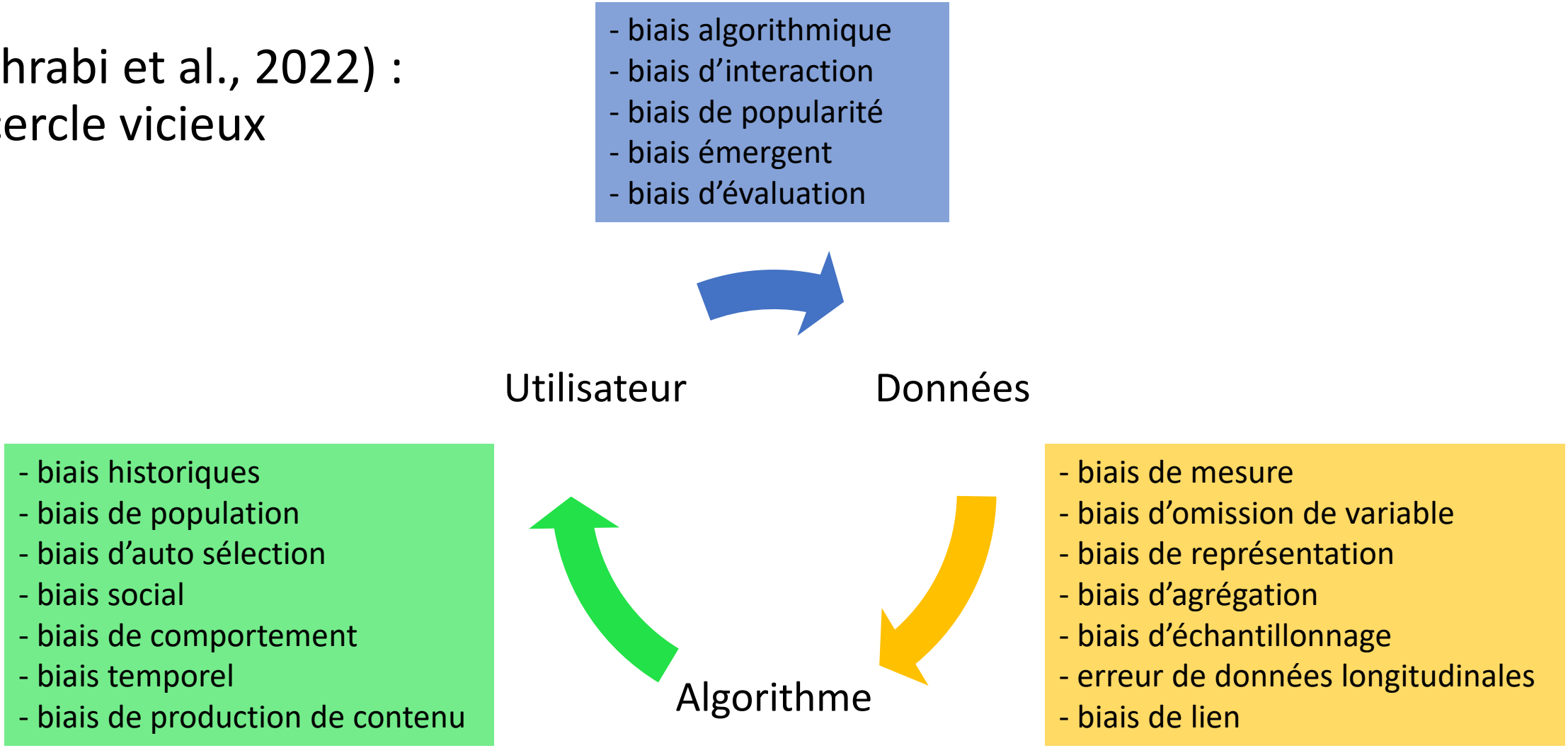
Exemple de typologie des biais en IA 1/2

(Hovy, Prabhumoye, 2021) : biais et étapes de conception d'un SIA

- **biais de conception de la recherche** Pourquoi ce sujet de recherche ?
  « syndrome du lampadaire »
- **biais de sélection** Quelles sont les données utilisées ?
  bases de données inadaptées à l'application visée
- **biais d'annotation** Comment et par qui ont-elles été annotées ?
  experts /crowdsourcing, représentativité / subjectivité des annotateurs
- **biais sémantique** Y a-t-il des corrélations indésirables ou fallacieuses entre les données ?
  « nurse » => « infirmière »
- **biais de suramplification** Quelles sont les dérives possibles du modèle ?
  avenir du SIA après son déploiement

Exemple de typologie des biais en IA 2/2

(Mehrabi et al., 2022) :
un cercle vicieux



Objectifs de cette présentation

- Préciser l'origine de l'emploi du mot « biais » en IA
- Eclaircir ses différentes significations
- Discuter de son caractère « technosolutionniste »

biais = écart à une norme

Biais : origine du concept et champs applicatifs

Etymologie

source : TLFi (Trésor de la langue Française informatisé)

Mot probablement emprunté à l'ancien provençal (Occitan)
et issu du latin *biaxius* « qui a deux axes »

vers 1250 : locution adverbiale « de biais » = en diagonale (*domaine de la couture*)

1563 : adjectif « biais » = oblique (ex : *mur biais*, *porte biaise*)

fin du XVIème :
- substantif « un biais » :
  peut designer un aspect d'une chose
  ou une forme oblique

- utilisé au sens figuré : moyen de résoudre un problème (ex : *par le biais de*)
  (péjoratif : *détour, travers, subterfuge*)

Etymologie

source : Online Etymology Dictionary

En anglais : « bias » emprunté au français vers 1520

Terme technique (*bowling*)
pour décrire des boules mal équilibrées
qui ont tendance à s'incliner dans une
direction donnée

⇒ dès 1570 : utilisé au sens figuré
pour parler d'une tendance de l'esprit

⇒ 1610 : utilisé comme verbe
to bias = donner un biais
biased = « *biaisé* »

English Life in Tudor Times, Roger Hart, NT: Putnam, 1972

Un retour en France par le monde scientifique

biais **(1)**

source : Larousse

nom masculin

(peut-être ancien provençal *biais*, détour, du latin populaire **biaxius*, qui a deux axes)

1. Caractère oblique ; ligne oblique par rapport au plan générateur : Le biais d'un mur.

2. Moyen indirect et habile de résoudre une difficulté : Chercher un biais pour éviter une corvée.
   SYNONYMES :
   détour - faux-fuyant - ruse - subterfuge

Couture

3. Diagonale d'un tissu par rapport à ses deux droits-fils (chaîne et trame).

4. Bande de tissu coupée dans le sens de cette diagonale et utilisée en garniture.

Passementerie

5. Ornement de guipure posé en spirales sur un galon.

Un retour en France par le monde scientifique

source : Larousse

biais **(2)**

nom masculin

(de l'anglais *bias*)

1. Distorsion, déformation systématique d'un échantillon statistique choisi par un procédé défectueux, ou d'une évaluation.

2. Différence entre l'espérance mathématique d'un estimateur et la grandeur à estimer.

biaisé, biaisée

adjectif

(de biais 2)

1. Se dit de quelque chose qui est déformé, fallacieux, détourné de son but originel : Débat biaisé.

2. Se dit d'une observation présentant une distorsion systématique, d'un raisonnement fondé sur une telle observation.

Disciplines scientifiques concernées

source : Web of Science

Multitude d'occurrences et de significations

voltage bias, algorithmic bias,
bias in peer-review, education, history, politics, clinical research...

= objet / domaine d'étude spécifique

small sample bias, confounding bias, selection bias...

= biais statistiqueS
(≠ le biais statistique)

bias elicitation, bias awareness, bias testing, bias correction...

= biais méthodologiques

difficulty bias, attention bias, optimism bias, memory bias,
hinsight bias, ideological bias, confirmation bias...

= biais cognitifs

*(**implicit**) racial bias, gender bias, social class bias...*

= biais socio-historique

Biais statistique

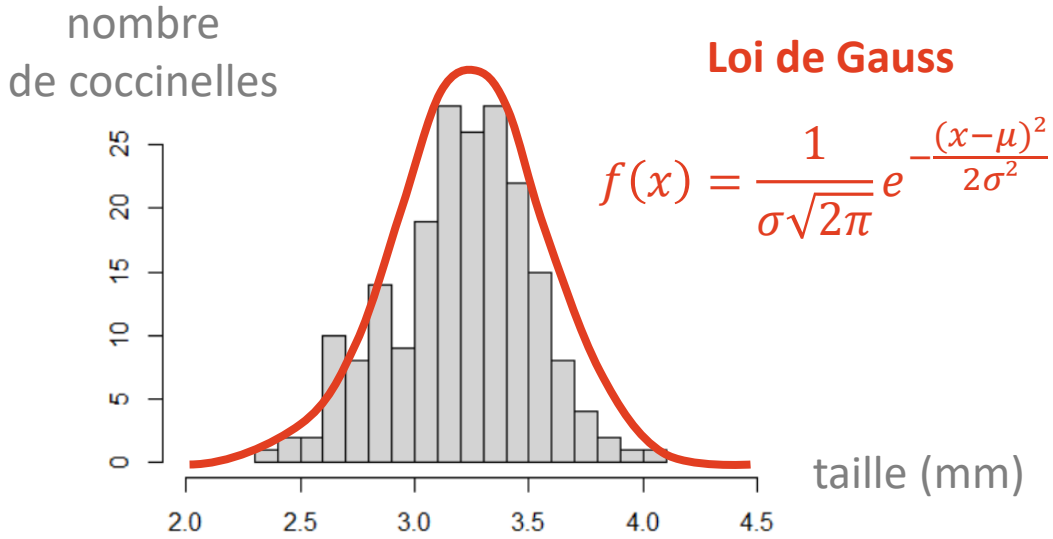
Biais statistique

Concept défini dans le domaine de l'**estimation**

Objectif : décrire un phénomène x à partir d'un modèle mathématique
= fonction f qui dépend de plusieurs paramètres $(\theta, \mu, \sigma, \ldots)$

Exemple : On souhaite connaître la taille moyenne d'une espèce de coccinelle

1) Prélèvement d'un échantillon de N coccinelles
2) Mesure de la taille x_i de chaque individu
2) Représentation sous forme d'histogramme
3) Choix du modèle mathématique
4) Estimation des paramètres du modèle



Biais statistique

Estimateur = fonction mathématique permettant d'approcher les paramètres du modèle
Estimation = résultat de l'estimateur pour un échantillon donné

exemple de la moyenne empirique : $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i$

Plusieurs estimateurs possibles pour chaque paramètre θ

=> Comment choisir le meilleur ?

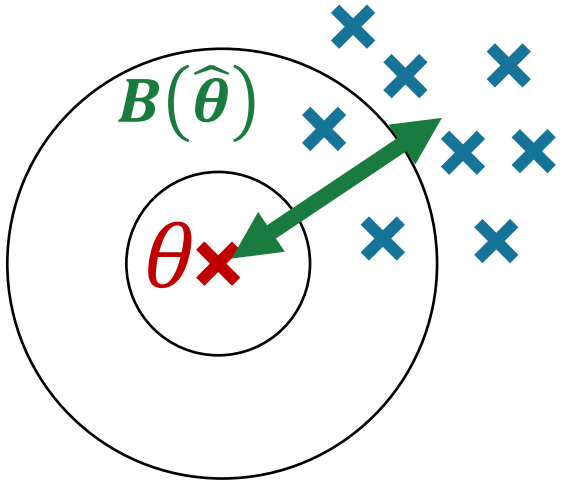
Erreur quadratique moyenne : $MSE(\theta) \stackrel{\text{def}}{=} \mathbb{E}\left[\left(\hat{\theta} - \theta\right)^2\right]$

Deux types d'erreurs : $MSE(\theta) = Biais\left(\hat{\theta}\right)^2 + Var\left(\hat{\theta}\right)$

Biais statistique

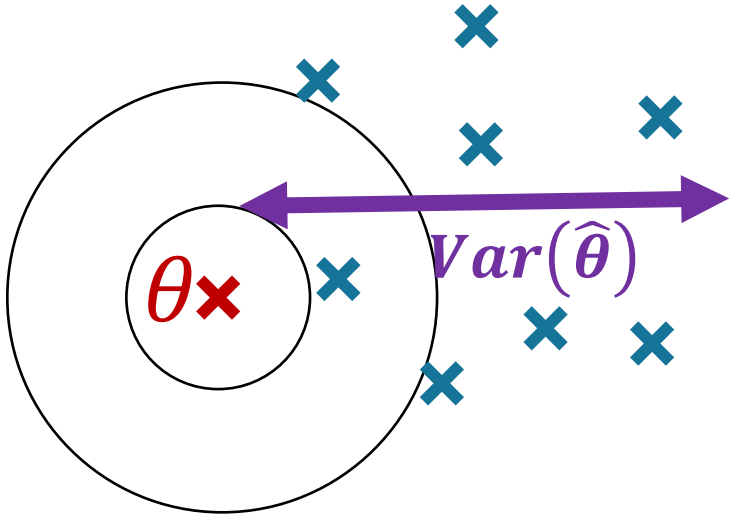
Biais : $Biais(\hat{\theta}) \overset{\text{def}}{=} \underbrace{\mathbb{E}(\hat{\theta})} - \theta$

espérance mathématique
(moyenne des **estimations**)



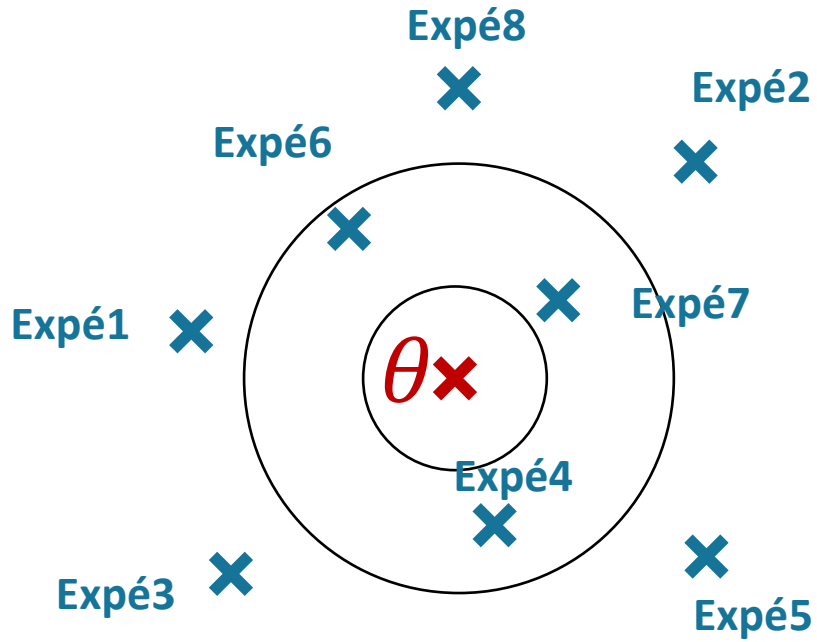
Variance : $Var(\hat{\theta}) \overset{\text{def}}{=} \mathbb{E}\left[(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2\right]$

mesure de la dispersion
des **estimations**

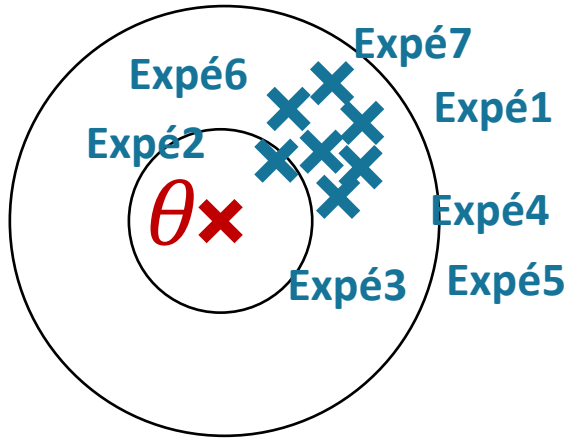


Biais statistique

Choix de l'estimateur = compromis biais / variance



estimateur sans biais
mais avec une grande variance

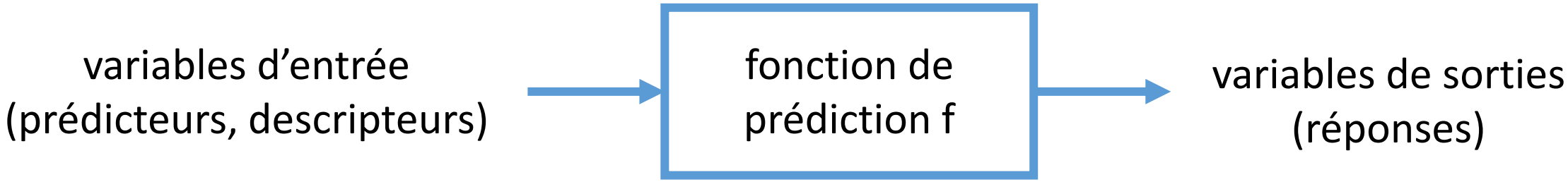


estimateur biaisé
avec une petite variance

Le biais de l'IA ?

modèle d'IA = fonction de prédiction

dont les paramètres sont calculés à partir d'un ensemble d'apprentissage



2 approches :

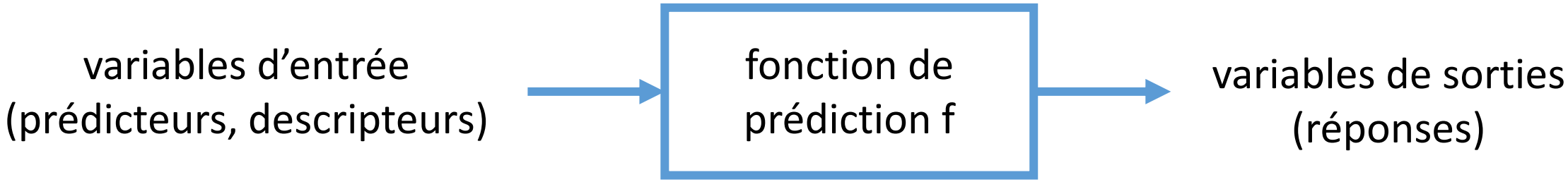
- choix a priori d'un modèle statistique => interprétable, mais moins performant
- modèle « boîte-noire » => très performant, mais pas interprétable

Le biais de l'IA ? NON dans la majorité des cas

modèle d'IA = fonction de prédiction

dont les paramètres sont calculés à partir d'un ensemble d'apprentissage

sans signification physique ou biologique



2 approches :

- choix a priori d'un modèle statistique => interprétable, mais moins performant
- modèle « boîte-noire » => très performant, mais pas interprétable

Biais méthodologiques

Biais méthodologiques

Mauvaises pratiques de recherche
et/ou utilisation malheureuse des statistiques

(≠ fraude scientifique)

Exemples :

- biais de mesure = l'outil utilisé ne permet pas de réaliser une mesure fiable
- biais d'échantillonnage = échantillon pas représentatif de la population étudiée
- *Cherry picking* = sélection des données (suppression des valeurs aberrantes)
- biais de publication = résultats statistiquement significatifs favorisés

Biais méthodologiques

Biais = écart entre la mesure / l'observation / le modèle et la réalité que l'on cherche à étudier

Exemple en IA : (Shah, Schwartz & Hovy, 2019 :

« **écart** entre a) la distribution « réelle » ou attendue (des utilisateurs, étiquettes ou résultats) et b) la distribution **utilisée** ou **produite** par un modèle »

(Campbell, 1957)

⇒ **validité interne** : Est-ce que les conclusions de l'étude sont valables ?

Quelles sont les données utilisées pour l'apprentissage ?

⇒ **validité externe** : Est-ce que les résultats sont généralisables ?

Quelles sont les données produites par le modèle ?

Biais méthodologiques

Biais = écart entre la mesure / l'observation / le modèle et la réalité que l'on cherche à étudier

Difficultés :

- Pas mesurable dans l'absolu (donc pas de modèle sans biais)
- Question fondamentale en philosophie des sciences & épistémologie
- **En IA : croyance en l'objectivité des données « brutes » et aux capacités du Big Data d'englober toute la réalité**

  ex : Chris Anderson prophétisant en 2010 la fin de la théorie scientifique

(Iliadis & Russos, 2016)
(Zacklad & Rouvroy, 2021)
(Bender et al., 2021)

Biais cognitifs

Biais cognitifs

Concept central au modèle de la pensée basé sur 2 modes de raisonnement
proposé par les psychologues **Daniel Kahneman** et **Amos Tversky** (1972) :

Système 1 : **« heuristiques » =** simplifications du monde, raccourcis de pensées qui permettent un raisonnement **intuitif**, **rapide**, **inconscient**, mais **faillible** (biaisé)

Système 2 : logique, **rationnel**, mais plus long et plus coûteux

« Biais » = Tendance à prendre en compte des éléments non pertinents
dans une prise de décision (« failles » de la rationalité)

=> Erreurs de jugement **systématiques** et **prévisibles**
qui peuvent être étudiées expérimentalement
(ex : biais d'ancrage, biais d'aversion aux pertes)

Biais cognitifs & IA

Existence de définitions similaires dans le domaine de l'IA :

Jean-Michel Loubes (SMAC, 2022) :
« information **non pertinente**, mais qui influence malgré tout le résultat de l'algorithme »

(Shah, Schwartz & Hovy, 2019) :
« Les biais sont une propriété inhérente des systèmes NLP (et plus largement de tout modèle statistique) mais ce n'est pas en soi négatif. En substance, les biais sont des **connaissances a priori** (*priors*) qui informent nos décisions »

Rapport sur l'IA « digne de confiance » (2019) :
« **inclination au préjugé** envers ou contre une personne, un objet ou un point de vue. »
« positif ou négatif, intentionnel ou involontaire » (≠ « biais injustes »)

Biais cognitifs & technosolutionnisme

2 discours contradictoires coexistent à l'heure actuelle

- **Mise en garde contre les biais des SIA**

- **Eloge de la « rationalité » automatisée**

  (Kahneman et al. 2016) : algorithmes « biaisés », mais « sans bruit »

  (Thaler et Sunstein, 2003) : « paternalisme libertarien » & nudges

  (Haselton & Buss, 2000) : psychologie évolutionniste & naturalité des biais cognitifs
  => les biais cognitifs seraient **innés** et dû à une inadéquation entre nos conditions de vie actuelles et celles qui ont vu apparaître l'espèce humaine

Biais cognitifs & ses critiques

- Critiques politiques et philosophiques
  => « Naturalité » des biais cognitifs (Stiegler, 2022)
  = **justification de l'idéologie néo-libérale : technocratique et autoritaire**
  => Nudging = mélange des genres politique / marketing aux effets délétères (Zacklad, 2022)

- Critiques de l'économie comportementale
  => expérimentations non écologiques et aux interprétations contestés (Gigerenzer, 2018)
  => définition restreinte de la rationalité (= maximisation des bénéfices) (Bergeron et al., 2018)
  => paradigme réductionniste : (Servet, 2018)
  analyse sous l'angle de la **déviance psychologique individuelle**, (Jatteau, 2021)
  au mépris des approches sociologiques et historiques

- Critique de l'ethnocentrisme de la psychologie : « Most people are not **WEIRD** »
  (Henrich, Heine et Norenzayan, 2010)

Western, Educated
Industrialized,
Rich and Democratic

Biais socio-historique

Biais socio-historique

Biais = écart entre la décision prise
et le résultat juste / désirable du point de vue éthique

Exemples :

(Shah, Schwartz & Hovy, 2019) :

« écart entre a) la distribution « réelle » ou **attendue** (des utilisateurs, étiquettes ou résultats)
et b) la distribution utilisée ou produite par un modèle »

Rapport sur l'IA « digne de confiance » (2019) :

« **biais injustes** » = « susceptibles d'entraîner des résultats discriminatoires et/ou injustes »

(Mehrabi et al., 2022) :

« **source d'iniquité** » (*unfairness*)

=> Discours technosolutionnistes :

biais dans les jeux de données car produits par une société inégalitaire
mais possibilité de « débiaiser » les algorithmes

(Chayes, 2017)

Biais socio-historique

Problèmes :

➢ Plusieurs définitions différentes et **incompatibles** de la *fairness* (Binns, 2018)

(Mehrabi et al., 2022) : 10 définitions mathématiques, simplifiables en 2 catégories principales

- équité individuelle (principe méritocratique)
- équité de groupe (principe paritaire)

➢ Quels « attributs protégés » ?

(Carvalho et al., 2022)

	G1	G2
♀	**50%**	**0%**
♂	**0%**	**50%**

- Enjeux d'**intersectionnalité** :
  risque d'amplifier les inégalités si attributs protégés considérés comme indépendants

- Enjeux de **vie privée** : « attributs protégés » = motifs de discriminations

Biais socio-historique

➢ Risque d'imposer une **vision du monde hégémonique**

- Qui définit la norme sociale implémentée dans le SIA ?
  Exemples :
  - ➢ GPT-3
    93% des données d'entraînement en langue anglaise
    prévalence de valeurs américaines

    (Johnson et al., 2022)
  - ➢ Modération de contenu
    Comment un algorithme (ou un annotateur peu informé) peut-il distinguer un contenu insultant, d'une réappropriation de l'insulte par la communauté visée ?

- SIA « rebiaisés » pour être plus équitables par nature conservateurs,
  car mise à jour et évaluation des modèles très coûteuse financièrement et écologiquement

(Bender et al., 2021)

Conclusion

Synthèse

Biais = un mot fourre-tout, aux multiples significations

Au moins 4 approches distinctes :

- Biais statistique => **abstraction mathématique** (concept bien défini, mais inopérant)
- Biais méthodologique => **réalité / vérité** (question philosophique et épistémologique)
- Biais cognitif => « **rationalité** » (norme utilitariste)
- Biais socio-historique => **société idéale** (question politique)

Synthèse

Biais en IA : un concept technosolutionniste ?

- Pour certains, résoudre les biais des SIA = résoudre les problèmes de la société

- Tout « rebiaisage » consiste à imposer une norme
  et relève de la « petite éthique » / « éthique externe »

  => véritables enjeux = Quel est le bienfondé d'utiliser le SIA ?
  Dans quelle société souhaite-t-on vivre ?

(Hunyadi, 2015)
(Zacklad et Rouvroy, 2021)

Assumer que les biais en IA ne constituent pas un problème technique (biais vs. performances) mais une question politique :
(D'Ignazio et Klein, 2020) proposent de remplacer le mot « biais » par « oppression »

Grenoble Alpes

ÉTHIQUE & IA

Université Grenoble Alpes

Merci de votre attention !

ambre.davat@univ-grenoble-alpes.fr

Références

Références

Bender, Emily M., et al. « On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜 ». Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, 2021, p. 610-23. ACM Digital Library, https://doi.org/10.1145/3442188.3445922.

Bergeron, Henri, et al. Le biais comportementaliste. Presses de Sciences Po, 2018, https://www.cairn.info/le-biais-comportementaliste--9782724622409.htm. Cairn.info.

Binns, Reuben. « Fairness in Machine Learning: Lessons from Political Philosophy ». Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR, 2018, p. 149-59. proceedings.mlr.press, https://proceedings.mlr.press/v81/binns18a.html.

Campbell, Donald T. « Factors relevant to the validity of experiments in social settings ». Psychological Bulletin, vol. 54, no 4, 1957, p. 297-312. APA PsycNet, https://doi.org/10.1037/h0040950.

Carvalho, Jean-Paul, et al. Affirmative Action with Multidimensional Identities. 4070930, 30 mars 2022. Social Science Research Network, https://doi.org/10.2139/ssrn.4070930.

D'Ignazio, Catherine, et Lauren F. Klein. Data Feminism. MIT Press, 2020.

Gigerenzer, Gerd. « The bias bias in behavioral economics ». Review of Behavioral Economics, vol. 5, no 3-4, 2018, p. 303-36.

Groupe d'experts de haut niveau sur l'intelligence artificielle. Lignes directrices en matière d'éthique pour une IA digne de confiance. Office des publications de l'Union européenne, 2019. Office des publications de l'Union européenne, https://data.europa.eu/doi/10.2759/74304.

Références

Haselton, Martie G., et David M. Buss. « Error management theory: A new perspective on biases in cross-sex mind reading ». Journal of Personality and Social Psychology, vol. 78, no 1, 2000, p. 81-91. APA PsycNet, https://doi.org/10.1037/0022-3514.78.1.81.

Henrich, Joseph, et al. « Most People Are Not WEIRD ». Nature, vol. 466, no 7302, 7302, juillet 2010, p. 29-29. www.nature.com, https://doi.org/10.1038/466029a.

Hunyadi, Mark. La tyrannie des modes de vie: sur le paradoxe moral de notre temps / Mark Hunyadi. le Bord de l'eau, 2015.

Iliadis, Andrew, et Federica Russo. « Critical Data Studies: An Introduction ». Big Data & Society, vol. 3, no 2, décembre 2016, p. 2053951716674238. SAGE Journals, https://doi.org/10.1177/2053951716674238.

Jatteau, Arthur. « Faire preuve par le chiffre ? : Le cas des expérimentations aléatoires en économie ». Faire preuve par le chiffre ? : Le cas des expérimentations aléatoires en économie, Institut de la gestion publique et du développement économique, 2021. OpenEdition Books, http://books.openedition.org/igpde/12229.

Johnson, Rebecca L., et al. « The Ghost in the Machine has an American accent: value conflict in GPT-3 ». arXiv preprint arXiv:2203.07785, 2022.

Kahneman, Daniel, et al. « Noise: How to Overcome the High, Hidden Cost of Inconsistent Decision Making ». Harvard Business Review, octobre 2016, p. 9.

Mehrabi, Ninareh, et al. A Survey on Bias and Fairness in Machine Learning. arXiv:1908.09635, arXiv, 25 janvier 2022. arXiv.org, https://doi.org/10.48550/arXiv.1908.09635.

Références

Mitchell, Margaret, et al. « Diversity and Inclusion Metrics in Subset Selection ». Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, Association for Computing Machinery, 2020, p. 117-23. ACM Digital Library, https://doi.org/10.1145/3375627.3375832.

Shah, Deven Santosh, et al. « Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview ». Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, p. 5248-64. DOI.org (Crossref), https://doi.org/10.18653/v1/2020.acl-main.468.

Stiegler, Barbara. L'idéologie des biais cognitifs. https://www.youtube.com/watch?v=Z71oV00aqxk. Bibliothèques de Bordeaux.

Thaler, Richard H., et Cass R. Sunstein. « Libertarian paternalism ». American economic review, vol. 93, no 2, 2003, p. 175-79.

Vallet, Guillaume. « Jean-Michel Servet, L'économie comportementale en question, Paris, Éditions Charles-Léopold-Mayer, 2018, 208 p. » Revue Interventions économiques. Papers in Political Economy, no 60, 60, décembre 2018. journals-openedition-org.gaelnomade-1.grenet.fr, http://journals.openedition.org/interventionseconomiques/4534.

Zacklad, Manuel. Communication instituante et confusion des institutions signifiantes: approche critique de la communication en santé pendant la crise du Covid. 2022, p. 6.

Zacklad, Manuel, et Antoinette Rouvroy. « Enjeux éthiques situés de l'IA ». Sociétés et espaces en mouvement, 2021, p. 15.