Biaix algorítmic a la IA: conceptes clau, implicacions i solucions

Aquest OER examina la IA i el biaix algorítmic, les seves causes i les seves repercussions al món real. Destaca la importància de la supervisió humana i les estratègies per mitigar el biaix i aconseguir sistemes d'IA més justos.

La intel·ligència artificial (IA): fa referència als sistemes que mostren un comportament intel·ligent analitzant el seu entorn i emprenent accions -amb cert grau d'autonomia- per assolir objectius específics.

El biaix algorísmic: és un error sistemàtic o un resultat injust produït per algorismes d'aprenentatge automàtic, que sovint reflecteix o reforça biaixos socials, racials o de gènere existents, causats per suposicions prejudicioses en el disseny o per l'entrenament amb dades incompletes o distorsionades.

Causes:

Dades d'entrenament defectuosos

La IA prediu en funció de les dades recollides; si no hi ha dades, podeu al·lucinar

La IA no sap, recopila i pronostica

Prejudicis dels desenvolupadors

La IA és un programa que segueix ordres (el vostre codi i després les indicacions)

Els desenvolupadors són persones amb prejudicis humans

Prejudicis en acció: Exemples i impactes

Sanitat: Els classificadors de lesions cutànies de la CNN, entrenats principalment en pacients blancs, mostren una precisió -50% menor en pacients negres.

Àmbit jurídic: Les eines de risc com COMPAS solen etiquetar determinades ètnies com d'alt risc.

Recursos humans: Els conjunts de dades esbiaixades amplifiquen les diferències salarials o exclouen candidats. Exemple: L'algorisme d'Amazon del 2018 va rebaixar els currículums amb «de dona».

Vies de mitigació

- Dades diverses i representatives
- Auditories periòdiques d'exactitud i imparcialitat
- Directrius reglamentàries i ètiques
- Transparència i explicabilitat de les decisions sobre IA
- Ús de la IA per identificar possibles biaixos i disparitats

Coses a tenir en compte:

La IA no percep el llenguatge gramaticalment, sinó que prediu estadísticament

Podeu confondre temes similars repetint patrons o barrejant conceptes relacionats

Penseu en això com una predicció de text altament qualificada amb una il·lusió de creació de coneixement i aportació de significat.

TU crees coneixement i dónes sentit quan TU llegeixes críticament.

La IA no segueix la lògica humana, per la qual cosa els seus resultats requereixen una revisió humana.

Audiències: 📚 Personal de biblioteques, 👩‍🏫 Professorat, 👨‍🎓 Estudiantat

Referències

Bandara, R. J., K. Biswas, S. Akter, S. Shafique, and M. Rahman. “Addressing Algorithmic Bias in AI-Driven HRM Systems: Implications for Strategic HRM Effectiveness.” Human Resource Management Journal, May 2025. https://doi.org/10.1111/1748-8583.12609

Jonker, Alexandra, and Julie Rogers. “What Is Algorithmic Bias?” IBM Think, September 20, 2024. https://www.ibm.com/think/topics/algorithmic-bias

Kalantzis, Mary, and Bill Cope. “Literacy in the Time of Artificial Intelligence.” Reading Research Quarterly 60, no. 1 (November 23, 2024). https://doi.org/10.1002/rrq.591

Norori, Natalia, Qiyang Hu, Florence Marcelle Aellen, Francesca Dalia Faraci, and Athina Tzovara. “Addressing Bias in Big Data and AI for Health Care: A Call for Open Science.” Patterns 2, no. 10 (2021): 100347. https://doi.org/10.1016/j.patter.2021.100347

Sheikh, Haroon, Corien Prins, and Erik Schrijvers. “Artificial Intelligence: Definition and Background.” Research for Policy, January 2023, 15–41. https://doi.org/10.1007/978-3-031-21448-6_2.

Context del projecte GEDIS

Aquest recurs forma part del projecte europeu GEDIS (Gender Diversity in Information Science), que promou eines educatives obertes per abordar les desigualtats de gènere en l'educació superior, amb un èmfasi especial en les disciplines relacionades amb la informació i la documentació. Aquest material educatiu s’ha desenvolupat en el marc de la Summer School Barcelona, dins del projecte GEDIS.

GEDIS - Gender Diversity in Information Science: Challenges in Higher Education. https://ub.edu/gedis

Citació: Karmil, Kenza, Luka Boskovic and Denisa Kartašová. 2025. *Algorithmic Bias in AI: Ke Concepts, Implications, and Solutions.* DOI: 10.5281/zenodo.17067091. *Biaix algorítmic en la IA: conceptes clau, implicacions i solucions.* Traduït por Claudia San-José i Juan-José Boté-Vericad. DOI:10.5281/zenodo.17306646

Cofinançat per la Unió Europea. Les opinions i punts de vista expressats són únicament responsabilitat del/de la/dels/les autor(s/es) i no reflecteixen necessàriament els de la Unió Europea ni els del Servei Espanyol per a la Internacionalització de l'Educació (SEPIE). Ni la Unió Europea ni l'autoritat concedent no se'n poden considerar responsables.

Cofinançat per la Unió Europea