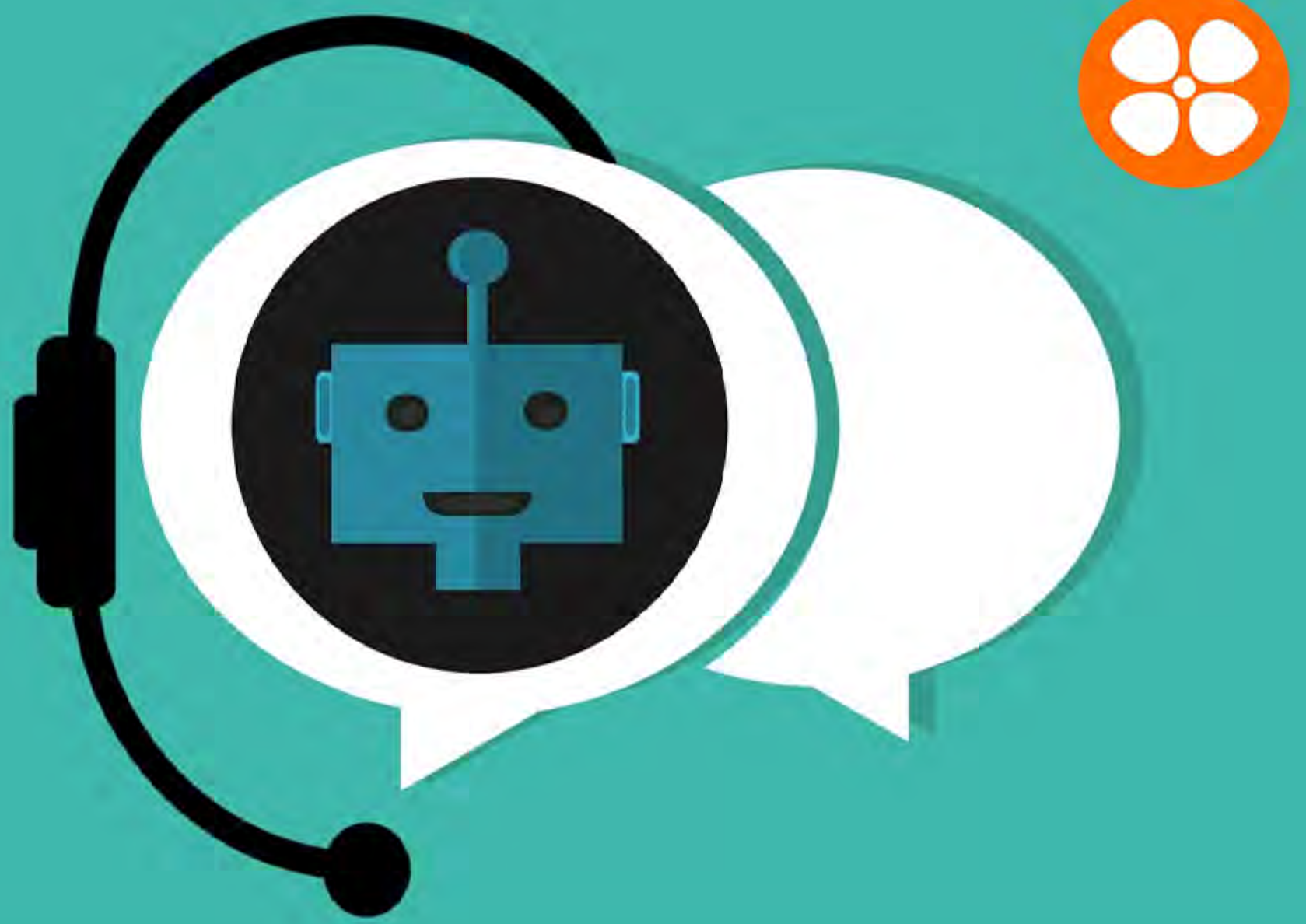
Els biaixos en les dades utilitzades en intel·ligència artificial

S'han convertit en un dels majors problemes que existeixen en la implementació d'aquesta tecnologia.
I suposen una amenaça per al desenvolupament d'una tecnologia equitativa, justa i segura. Encara que no són els únics, els biaixos més comuns són ...



Els biaixos d'interacció

- Són aquells que sorgeixen a partir de l'observació i confirmació d'un o més individus que interactuen amb un sistema d'IA mitjançant processos d'entrenament algorítmic.

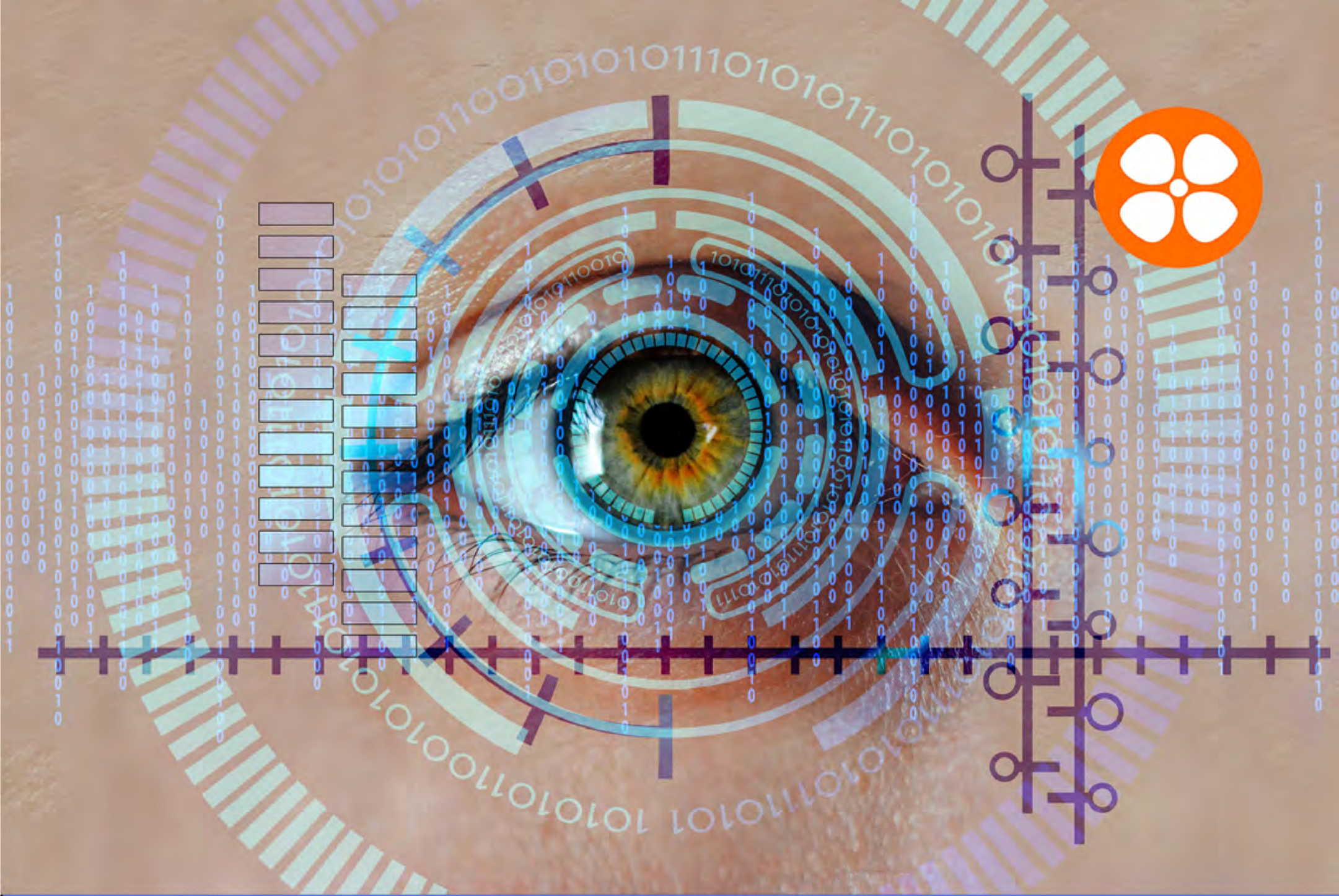
- Aquest biaix se'n deriva d'allò que s'espera veure o es vol veure per part d'un o més individus amb els seus prejudicis conscients i inconscients en un sistema d'IA.

- Trobem diversos exemples, com els xatbots entrenats d'una manera que reforcen estereotips, o imatges de persones etiquetades amb una visió etnocèntrica.

Els biaixos latents

- Són aquells que no es poden observar directament però són latents i estan continguts en conjunts de dades, i reprodueixen els prejudicis implícits històrics en una societat.

- Com per exemple que un sistema d'IA reconegui que un metge és home i no dona a causa de la utilització de dades històrics que revelen que els metges han estat principalment homes.



Els biaixos de selecció

- Sorgeixen quan es selecciona un conjunt de dades que no reflecteix la realitat de l'entorn en què s'executarà un sistema d'IA.

- Aquest biaix de selecció fa que el sistema d'IA només representi un conjunt específic d'elements però no pas tota la realitat o població que hauria de reflectir.

- Per exemple és quan un sistema de reconeixement facial es entrenat exclusivament a partir de cares caucàsiques.