LA "RAT RACE" ACTUAL DE LA #IA FUNCIONA AL CONTRARI DEL CERCLE VIRTUÓS JA QUE TENDEIX: 1) A MÉS ERRORS SOCIOTÈCNICS (TIPUS #GOOGLEBARD), 2) EN QUÈ VEIEM UNA IA MENYS FIABLE, I 3) EN QUÈ DISMINUEIXI LA NOSTRA CONFIANÇA EN LA IA. ESTEU D'ACORD?

- Una majoria (56%) esteu d'acord que la "rat race" actual de la #IA funciona al contrari d'un cercle virtuós, mentre que la resta no ho sabeu (20%), us mostreu ambivalents (16%) o no esteu acord amb aquesta afirmació (8%).

- Quan la indústria d'IA es preocupa més per competir i fer salts tecnològics endavant sense resoldre problemes estructurals (com de dades i dissenys) i maximitzant la monetització dels serveis per sobre del benestar dels ciutadans, els errors sociotècnics estan garantits.

- D'aquí que calgui invertir molt més en una comprensió més completa dels biaixos en les dades i dissenys i, per tant, cal tenir en compte com els biaixos humans i els biaixos sistèmics perjudiquen determinats grups socials, especialment quan aquests biaixos es combinen.

- No fer-ho és especialment greu ja que la indústria d'IA ho sap, però no ho aborda seriosament com un aspecte cabdal de la innovació. Si volem desenvolupar sistemes d'IA fiables, hem de tenir en compte tots els factors que poden reduir la confiança del públic en la IA.

- I molts d'aquests factors van més enllà de la tecnologia en si. De fet, un dels problemes és que les organitzacions sovint només utilitzen solucions tecnològiques per a problemes de biaix d'IA quan això no significa capturar adequadament el seu impacte social.

- L'expansió de la IA en molts aspectes de la vida pública requereix ampliar la nostra visió per considerar la IA dins del sistema social més ampli en què opera. Això significa que és important incorporar experts de diversos camps i alhora escoltar afectats sobre l'impacte de la IA.

- Però l'actual "rat race" ho dificulta enormement ja que imposa una narrativa en la que el futur de la IA serà capaç de resoldre no tan sols els problemes coneguts sinó també els desconeguts. Malauradament, això pot derivar en què veiem una IA menys fiable.

- I, com a conseqüència, en què disminueixi la nostra confiança en la IA. No és un fet menor, especialment si tenim en compte que els humans som poc indulgents amb les organitzacions i sistemes d'IA després de múltiples errors i, és clar, que és difícil recuperar la confiança.

- Aquesta és la conclusió d'un nou estudi de Esterwood i Robert Jr (2023) que demostra que, quan es produeixen errors, els humans sovint veuen els robots o sistemes d'IA menys fiables cosa que, en última instància, disminueix la seva confiança en ells.

- Tenint en compte aquesta situació, el mateix estudi de la Universitat de Michigan va examinar quatre estratègies per veure si hi ha la possibilitat de revertir la situació reparant o mitigant els impactes negatius d'aquestes violacions de confiança per part de robots o sistemes d'IA.

- Aquestes estratègies de reparació per millorar la confiança eren les disculpes, les negacions, les explicacions i les promeses de major fiabilitat. Els resultats van indicar que, després de tres errors, cap de les estratègies de reparació podrien reparar completament la fiabilitat.

- Això vol dir que si no posem fil a l'agulla en els problemes sociotècnics, ens arrisquem a perdre la confiança en la IA i això és un problema greu, ja que aquesta no es repara del tot (com indica la recerca) a través de disculpes, denegacions, explicacions o promeses.

- I fins aquí per avui. Això si, amb una lectura sobre el tema.

- Esterwood, C. & Robert Jr., L. P. (2023). Three Strikes and you are out!: The impacts of multiple human-robot trust violations and repairs on robot trustworthiness, *Computers in Human Behavior 142*, May 2023, 107658.