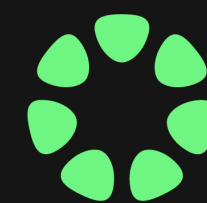


SABEU QUÈ SIGNIFICA “DECEPTIVE AI”?

- Sabeu que significa *“deceptive AI”* Aquesta fa referència a tecnologies d’IA que s’utilitzen per manipular a les persones de forma explícita o subliminal per induir a falses creences i per generar resultats enganyosos de forma sistemàtica.
- Tradicionalment, el *“deceptive AI”* s’evidenciava clarament en aquells sistemes que han estat entrenats per guanyar jocs i que ahora tenen un element social com, per exemple, Diplomacy o StarCraftII. Aquests sistemes d’IA pertanyen al grup d’una tasca específica.
- Però també existeix *“deceptive AI”* entre el grup de sistemes d’IA de propòsit general, com els grans models de llenguatge. Com les capacitats dels LLM han millorat ràpidament i estan dissenyats per realitzar una àmplia gamma de tasques mitjançant agents, també trobem l’engany.
- Malgrat que hi ha diversos tipus d’enganyats relacionats amb el frau, la influència i la polarització sociopolítica, tots tenen una cosa en comú: provoquen sistemàticament creences falses en els altres com un mitjà per aconseguir algun resultat que no sigui buscar la veritat.
- Per aquesta raó, es proposa que els responsables polítics regulin de manera eficaç els sistemes d’IA capaços d’enganyar, tant els LLM com els sistemes d’IA per portar a terme una tasca específica a través d’una categorització de riscos.
- També es proposa que hi hagi un requeriment explícit i diferenciador entre els resultats provinents de sistemes d’IA i aquells que són provinents de persones, de forma que es pugui distingir clarament el que és generat artificialment i el que no.
- A la vegada, es proposa un desenvolupament de tècniques de detecció de l’anomenada *“deceptive AI”* per identificar ràpidament quan els sistemes d’IA estan involucrats en l’engany i, per tant, fer que l’abast d’aquests sistemes sigui menor i menys enganyós.





- Evidentment, també cal garantir que els sistemes d'IA que es desenvolupen no parteixin d'una arquitectura que es molt susceptible de convertir-se en una “màquina de l'engany” i això significa que cal desenvolupar millors eines que els actuals LLM.
- Del contrari, existeix un risc important de debilitament humà ja que, a mesura que els sistemes d'IA s'incorporen a la nostra vida diària a un ritme més gran, és més fàcil que cada cop més els permetrem prendre moltes més decisions, algunes o moltes a través de l'engany.
- Si voleu saber-ne més sobre el tema podeu llegir aquest article sobre “*AI deception: A survey of examples, risks, and potential solutions*”, publicat a la revista científica Patterns.

