LA IA GENERATIVA HA MILLORAT DRÀSTICAMENT I ARA ES PODEN PRODUIR RESULTATS DE TEXT TAN REALISTES I CONVINCENTS QUE SÓN DIFÍCILS DE DISTINGIR DEL CONTINGUT ESCRIT PER HUMANS. AIXÒ OFEREIX UNA POSSIBILITAT MAJOR DE DESINFORMACIÓ I MANIPULACIÓ?

- La gran majoria (81%) considereu que la IA generativa ofereix una possibilitat major de desinformació i manipulació a través de text convincent, mentre que la resta sou ambivalents (10%), no ho considereu així (5%) o no ho sabeu (3%).

- **Així doncs, els "enginyers ràpids" són persones que escriuen prosa més que no pas codi per posar a prova els resultats dels sistemes d'IA generativa, i per detectar que les seves respostes siguin reproduïbles i que segueixin protocols de seguretat.**

- **Per tant, si bé aquest perfil de professional respon a la necessitat d'avaluar els models d'interacció home-màquina, amb l'arribada del Chat GPT es pot dir que tothom s'ha convertit en una espècie de "enginyer ràpid", però no pas per avaluar els sistemes d'IA generativa.**

- **De fet, més que avaluar el que estem fent és entrenant-los amb preguntes cada vegada més específiques, personals i críptiques, un fet que comporta riscos com la generació de contingut esbiaixat, de desinformació i també, és clar, de contingut maliciós.**

- **Davant d'aquesta situació, ja podem intuir un doble problema que cal abordar. Per una banda, calen més persones que facin la veritable feina d'avaluar els models d'interacció home-màquina. Per una altra banda, calen més indicacions sobre els seus abusos i mals usos.**

- **Això és especialment rellevant en àmbits com el de la salut o l'educació, on els consells i continguts personalitzats sempre apareixen com més atractius per un creixement i aprenentatge individualitzat de les persones. Però al marge d'aquests beneficis també hi ha inconvenients.**

- **Però això, per què succeeix?**

- **Primer, perquè nosaltres no escrivim totes les ordres amb les quals s'executa el sistema d'IA i, per tant, mai podem estar segurs que no ens sortirà per pataneres o falsedats.**

- **Segon, estem davant d'una IA generativa que es basa en la predicció de paraules de text. El resultat sempre és text plausible que el model dona basant-se en patrons lingüístics absorbits d'Internet. Malgrat la seva impressionant habilitat lingüística en termes predictius, aquests models no entenen la realitat ja que si bé saben predir no saben llegir i, això, té conseqüències importants.**

- **Com apunta @GaryMarcus, "quan llegim un text, construïm un model cognitiu del significat que diu el text" però la IA generativa això no ho fa. Simplement prediu unes paraules.**

- **Per tant, encara que els chatbots donin textos versemblants i sonin com si estiguessin molt segurs de les seves respostes, són molts els exemples que demostren la seva capacitat de desinformació i manipulació de manera sistemàtica.**

- **La raó hauria de ser fàcil d'entendre. No és el mateix un model de llenguatge on el "coneixement" es limita a patrons lingüístics que un model del món físic descrit pel llenguatge, que és el fem servir els humans per escriure i el que no fan servir les màquines.**

- **Malgrat que sistemes més complexos, com els d'aprenentatge de reforç basat en la retroalimentació humana (RLHF), on s'afinen els models per complir objectius específics i fer veure que els chatbots tenen la capacitat de simular la realitat, el problema principal persisteix.**

- **I és que la resposta sempre passa per un model de llenguatge i no pas per la realitat. O sigui, sempre hi ha un model que parteix de la informació d'Internet (que no és la realitat) i d'aquí escriu una resposta predictiva, cosa que ens remet de nou a l'al·lucinació.**

- **Així doncs, es pot dir que la diversió actual de la IA generativa a través de chatbots que "al·lucinen" com el #ChatGPT pot ser temporal ja que, com expressen moltes persones, es tracta d'una situació que comporta el perill real de contaminar Internet amb desinformació.**

- I fins aquí per avui. Us recomanem algunes lectures sobre el tema:

- Shanahan, M. (2022). Talking About Large Language Models, *arxiv.org*

- Alouani, N. (2023). Artificial Disinformation: Can Chatbots Destroy Trust on the Internet?, *medium.com*

- Marcus, G., & Davis, E. (2019). *Rebooting AI: Building artificial intelligence we can trust.* Vintage.

- Chomsky, N. (2023). La falsa promesa del ChatGPT, *ara.cat*