Joves, Ciència i Ètica

9è debat: Intel·ligència Artificial

12 de desembre de 2024, Museu de la Ciència CosmoCaixa

ORGANITZEN:

fcri Fundació Catalana per a la Recerca i la Innovació

Generalitat de Catalunya

AMB EL SUPORT DE:

EduCaixa
Fundació "la Caixa"

Ajuntament de Barcelona

IIIA-CSIC
INSTITUT D'INVESTIGACIÓ EN INTEL·LIGÈNCIA ARTIFICIAL

UAB
Universitat Autònoma de Barcelona

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Principis ètics per a una intel·ligència artificial justa i responsable

- Quins reptes ètics plantegen sistemes d'IA generativa com DALL-E o ChatGPT?

Es recomana que la intel·ligència artificial generativa respecti els principis **d'equitat**, **privadesa** i **explicabilitat/responsabilitat**.

ChatGPT "creu" que els respecta.

Principis ètics per a una intel·ligència artificial justa i responsable

- Creieu que els sistemes d'IA que feu servir habitualment segueixen els principis ètics de la nostra societat?

- **Privadesa**: Hi ha molts interessos perquè no es respecti.

- Cal tenir en compte dades sensibles.
- Hem de ser usuaris crítics.

Què en podem fer nosaltres?

- **Explicabilitat/Responsabilitat**:

- Com aprenen de tot arreu i funcionen per probabilitats, s'equivoca.
- Responsabilitat compartida.
- No és explicable l'origen de la resposta i, per tant, no sabem qui és responsable.

- **Equitat**:

- Presenten molts biaixos, que surten dels productes de la nostra societat i, a més, els amplifica.

Intel·ligència Artificial alineada amb valors morals

- Quins reptes morals planteja el disseny de sistemes d'IA?

1. La IA s'ha de dissenyar de manera que respecti una sèrie de valors morals a l'hora de prendre decisions o actuar. Per tant, s'ha de guiar pels següents principis ètics.
- Transparència: d'on obté les dades, quin és el raonament darrera de les seves decisions, etc.
- Equitat: no presenti biaixos en les seves accions o decisions.
- Beneficència: procurar el bé de les persones.
- Seguretat: respectar les nostres dades humanes i reduir al mínim el dany a les persones.
- Autonomia humana: entendre la IA com a una eina que no limiti la presa de decisions de les persones.

2. Educar a la ciutadania en un ús correcte de la IA:
- Conèixer què signem quan donem consentiment.
- Conèixer l'ús concret de cada sistema d'IA.
- Seguir un comportament ètic en l'ús de la IA.

- Podem alinear un sistema d'IA amb els valors morals de manera que tota la societat estigui d'acord?

Obtenir un acord en els valors amb els que s'hauria d'alinear una IA és una tasca difícil degut a la diversitat cultural de la nostra societat. Tot i això, creiem que s'hauria d'arribar a un **consens general** de mínims basats en els principis ètics mencionats.

Com afecten les eines d'IA generatives el nostre procés d'aprenentatge

- Com podem utilitzar la IA de manera responsable en les nostres tasques acadèmiques?

- Cal que ens **focalitzem** més en el **procés d'aprenentatge** i no tant en els resultats.
- Ser **específics** en les preguntes que formulem en una **interacció contínua**.

- Cal ser crítics amb la informació de la IA? Què podem fer per ser-ho?

- Cal entendre bé el **funcionament** de la IA per a poder fer-ne un **bon ús**.
- Els textos amb els quals han estat **entrenades** aquestes eines poden contenir errors, estar desactualitzats i contenir biaixos. Fem servir el nostre **criteri**!

Els models de llenguatge són imparcials? Explorem els biaixos dels models de llenguatge

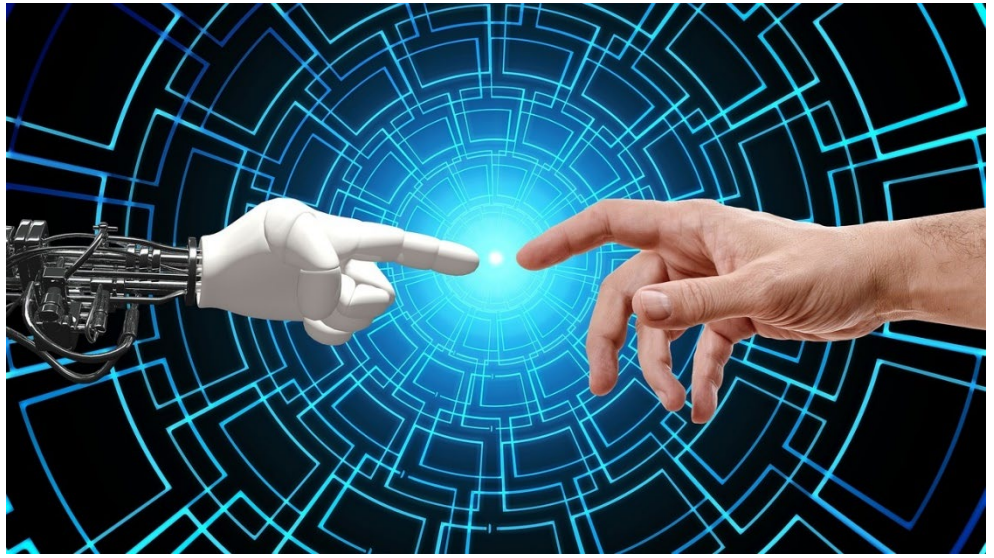
- Creieu que els biaixos que presenten els models de llenguatge són un risc per a la societat?

- Sí, els models de llenguatge presenten riscos com ara discriminació i perpetuació d'estereotips (minories ètniques, lingüístiques, gènere, cultura, ideologia, orientació sexual, diversitat funcional, etc.).
- Els biaixos provenen de les dades amb què entrenem els models de llenguatge, les quals contenen la nostra percepció del món, opinió, cultura. El model reflecteix aquests biaixos.

- Què significa per a vosaltres crear models de llenguatge més justos i inclusius?

- Filtrar les dades que s'utilitzen per entrenar els models.
- Filtrar les respostes del model.
- Crear models més petits, especialitzats en àmbits específics.
- Crear una comissió o grup regulador per avaluar de manera contínua els models i assegurar-nos que siguin justos. Aquest grup ha de ser plural.
- Fer que els models donin respostes múltiples a les nostres preguntes.
- Finalment, **els usuaris hem de tenir un punt de vista crític**.

Moltes gràcies per la vostra atenció!



ORGANITZEN:

fcri Fundació Catalana per a la Recerca i la Innovació

Generalitat de Catalunya

AMB EL SUPORT DE:

EduCaixa Fundació "la Caixa"

Ajuntament de Barcelona

IIIA-CSIC Institut d'Investigació en Intel·ligència Artificial

UAB Universitat Autònoma de Barcelona

BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación