Programa Joves, Ciència i Ètica

La Fundació Catalana per a la Recerca i la Innovació (FCRI), en col·laboració amb l'Ajuntament de Barcelona i el Departament d'Educació de la Generalitat de Catalunya, organitzen el programa "Joves, Ciència i Ètica".

L'objectiu d'aquest programa és estimular la capacitat crítica dels alumnes i promoure la seva implicació en els debats ètics que avui dia es plantegen al voltant de temes d'actualitat vinculats a la descoberta científica i les seves implicacions socials, com ara el canvi climàtic, la biomedicina, la nanotecnologia, la robòtica, la privacitat de les dades, entre d'altres.

El programa permet que joves estudiants de **3r d'ESO** debatin i coneguin els aspectes ètics que comporta la recerca científica i estableixin les seves pròpies conclusions, amb el suport de científics especialitzats en cada temàtica i d'experts en ètica.

En aquesta vuitena edició es proposa el tema de la **Intel·ligència Artificial.** La temàtica es dividirà en quatre grans subtemes:

1. Principis ètics per a la IA
2. IA amb valors morals
3. Eines IA generatives i aprenentatge
4. IA i biaixos en els models de llenguatge

Està prevista la participació de 5 escoles. Cada escola participant aporta 4 grups/classe d'alumnes. Cadascun dels grups treballarà un d'aquests subtemes. Hi haurà un grup de científics que es desplaçarà als centres educatius per impartir xerrades que tractaran aquestes temàtiques.

Els científics participants pertanyen a l' l'Institut d'Investigació en Intel·ligència Artificial del Consell Superior d'Investigacions Científiques (IIIA-CSIC), Universitat Autònoma de Barcelona (UAB) i Barcelona Supercomputing Center (BSC)

El programa es divideix en 5 etapes:

ETAPA 1: FORMACIÓ DE PROFESSORAT	
Lloc: la formació es farà presencial. Auditori de la Fundació Catalana per a la Recerca i la Innovació (FCRI). Passeig Lluís Companys, 23. 08010 Barcelona	
Persones destinatàries: professorat del centre directament implicat en el projecte, que conduirà el debat amb l'alumnat de 3r d'ESO o coordinarà aspectes organitzatius.	
Sessió 1: **16 d'octubre 2024** **17h**	Aquesta sessió formativa anirà a càrrec de **Sílvia Lope** (professora de Ciències de la Naturalesa) i de **Laura Farró** (professora de Llengua i Literatura Catalana). La formació facilitarà les eines i estratègies adients que han d'ajudar els docents a conduir el debat ètic amb els seus alumnes Es resoldran els dubtes que els centres puguin tenir sobre els aspectes organitzatius del programa "Joves, Ciència i Ètica" i el treball que cal fer amb l'alumnat.

Sessió 2: **23 d'octubre 2024** **17h**	Sessió a càrrec de **Vicent Costa**, investigador de l'IIIA-CSIC. La sessió serà en format virtual. Els darrers anys, l'interès per la intel·ligència artificial (IA) ha experimentat un creixement exponencial, superant les fronteres de la disciplina i impactant nombrosos camps com ara l'economia, la medicina, la física o la psicologia. Paral·lelament, s'ha produït un augment desmesurat de l'atenció mediàtica sobre aquesta tecnologia, convertint la IA en un factor clau per a les societats contemporànies, especialment a causa de les múltiples aplicacions que es fan servir diàriament. Així doncs, aquest creixement i l'impacte associat han propiciat una preocupació notable sobre les implicacions ètiques i socials que se'n deriven. En aquesta sessió de formació, abordarem alguns dels aspectes principals d'aquestes qüestions i proporcionarem al professorat els coneixements i les eines necessàries per a facilitar el desenvolupament del programa. La sessió es dividirà en dues parts. En la primera part, s'oferirà una introducció general a la IA amb l'objectiu de fomentar una visió crítica i rigorosa de la disciplina. Concretament, es partirà de la qüestió fonamental de què entenem per intel·ligència artificial i s'exposaran els principals reptes i objectius científics del camp. A continuació, es descriuran els dos paradigmes centrals que han guiat la recerca en IA des dels seus inicis, i finalment, es presentarà una panoràmica general sobre l'estat actual de la disciplina. La segona part de la sessió se centrarà en les qüestions ètiques i socials que sorgeixen de l'aplicació i desenvolupament de la intel·ligència artificial. Això inclourà des de consideracions sobre el disseny i la recerca en la disciplina, fins a aspectes relacionats amb l'ús concret de sistemes amb IA. En particular, s'exposarà primer la relació constitutiva entre ètica i intel·ligència artificial. A continuació, es presentaran algunes de les problemàtiques fonamentals sobre l'ètica de la intel·ligència artificial. En aquest punt, l'exposició tractarà sobre les qüestions específiques que sorgeixen en integrar una dimensió ètica en la IA, evitant així les clàssiques preguntes, de sobra conegudes, sobre ètica i ciència en general. Finalment, per tal d'enriquir els debats entorn de les qüestions del programa, es discutiran conceptes com el biaix algorísmic, el disseny basat en valors o la discriminació, tot mitjançant exemples concrets. **Nota: Aquesta sessió es farà en format virtual**

ETAPA 2: TREBALL PREVI A L'AULA ABANS DE LA VISITA DEL/DE LA CIENTÍFIC/A
Lloc: escoles participants
Persones destinatàries: professorat del centre directament implicat en el projecte, que conduirà el debat amb l'alumnat de 3r d'ESO o coordinarà aspectes organitzatius.

El centre organitzarà l'alumnat de 3r d'ESO en 4 grups. Cada grup treballarà un dels subtemes plantejats:

1. Principis ètics per a la IA
2. IA amb valors morals
3. Eines IA generatives i aprenentatge
4. IA i biaixos en els models de llenguatge

El professorat exposarà a l'alumnat les repercussions ètiques que comporta la irrupció de la intel·ligència artificial en les nostres vides i plantejarà un seguit de preguntes transversals a tots els subtemes.

- Són fiables els sistemes d'intel·ligència artificial que fem servir habitualment? I lliures de biaixos algorísmics?
- Si hi ha consens a la comunitat científica sobre la darrera qüestió, per què hi ha actors que defensen amb vehemència la neutralitat de la intel·ligència artificial?
- Hi ha un acord a la comunitat científica pel que fa als principis ètics per assegurar un disseny, ús i desenvolupament responsable i ètic de la intel·ligència artificial?
- Creus que el desenvolupament actual de la intel·ligència artificial respecta els valors fonamentals compartits per la majoria de la nostra societat?

Aquest treball suposa una **dedicació d'una hora** lectiva.

ETAPA 3: VISITA DELS EXPERTS A LES ESCOLES	
Lloc: centres de les escoles participants.	
Persones destinatàries: alumnat de cada una de les escoles.	
Novembre 2024	Explicació, a càrrec dels investigadors/investigadores, de cadascun dels subtemes. S'introduiran per cada subtema unes preguntes que els alumnes hauran de treballar posteriorment a l'aula. **<u>Principis ètics per a una intel·ligència artificial justa i responsable</u>** **Vicent Costa,** científic titular a l'Institut d'Investigació en Intel·ligència Artificial del CSIC RESUM: En els darrers anys, la intel·ligència artificial (IA) ha irromput amb força en les nostres vides, influenciant nombrosos aspectes del nostre dia a dia. Des d'eines que ens ajuden a fer els deures o que potencien la nostra creativitat, fins a la seva influència en el nostre temps lliure —com ara les recomanacions audiovisuals per veure un divendres a la nit o la millora dels personatges en l'última versió del nostre videojoc preferit. Així mateix, la IA també intervé en decisions més rellevants per a nosaltres, com per exemple l'assignació de beques per estudiar a l'estranger durant l'estiu. Per tot això, cada vegada sentim més a parlar sobre la importància de desenvolupar una

IA justa i responsable, que no només generi informació o prediccions fiables i precises, sinó que també respecti la diversitat social dels seus àmbits d'aplicació, proporcioni explicacions clares de les decisions que pren, asseguri una cura de la privadesa de les persones usuàries i promogui un ús sostenible dels recursos, tot respectant el medi ambient.

En aquesta sessió, reflexionarem sobre com assolir un ús i desenvolupament d'una IA així, que respecti un conjunt de principis ètics fonamentals: equitat, privadesa, responsabilitat i explicabilitat. En particular, pel que fa a l'equitat, prendrem com a referència un sistema d'IA generativa per il·lustrar com es poden generar dos tipus de biaixos algorísmics que cal evitar en una IA justa. Així mateix, a través d'exemples, explorarem diversos tipus de dades i entendrem com les persones que investiguen en IA garanteixen la privadesa de les persones usuàries. Finalment, examinarem el disseny d'un sistema d'IA recent i discutirem els reptes i problemes que planteja en relació amb els principis d'explicabilitat i responsabilitat.

Preguntes:

- Creus que els sistemes d'IA que fas servir habitualment segueixen els principis ètics dels quals hem parlat avui?
- Penses que totes les societats haurien de compartir els mateixos principis ètics en dissenyar i fer servir sistemes d'IA?
- Trobes que hi ha principis ètics més importants que d'altres?
- Quins reptes ètics plantegen sistemes d'IA generativa com DALL-E o ChatGPT?
- Creus que s'hauria de permetre la comercialització de dades biomètriques?
- Penses que hem de fer servir sistemes d'IA si no podem atribuir-ne la responsabilitat pels resultats obtinguts?

<u>Intel·ligència Artificial Alineada amb Valors Morals</u>

Roger Lera, graduat en Física per la Universitat de Barcelona i estudiant de doctorat en Intel·ligència Artificial l'Institut d'Investigació en Intel·ligència Artificial del Consell Superior d'Investigacions Científiques (IIIA-CSIC).

RESUM

A mesura que els sistemes d'Intel·ligència Artificial (IA) s'integren en la nostra societat, ens hem d'assegurar que el procés de decisió d'aquests sistemes sigui fiable. Per tal d'aconseguir-ho, hem de dissenyar-los de forma que siguin legals, segurs i ètics, és a dir, que estiguin alineats amb els nostres valors morals. Els sistemes d'IA alineats amb els nostres valors morals actuen o prenen decisions d'acord amb el que és moral fer segons els nostres principis ètics.
Durant aquesta xerrada, introduirem conceptes bàsics alhora de dissenyar sistemes d'IA ètics i analitzarem exemples reals de sistemes alineats amb valors morals. A més, entendrem què són els sistemes de valors i discutirem la pluralitat de sistemes de valors en la societat. Finalment, posarem damunt la taula diferents reptes i problemes oberts que tenen els sistemes d'IA en

l'actualitat.

PREGUNTES:

- Creieu que es necessari que un sistema d'Intel·ligència Artificial estigui alineat amb els nostres valors morals? Per què?
- Amb quins valors morals s'hauria d'alinear una intel·ligència artificial? Hi ha prou amb establir únicament els valors morals amb els que s'hauria d'alinear un sistema d'IA?
- Quins creieu que son els reptes més importants per a dissenyar un sistema d'IA ètica?
- Podem alinear un sistema d'IA amb els valors morals que tota una societat estigui
  d'acord?

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

Carlos Borrego Iglesias, professor agregat al Departament d'Enginyeria de la Informació i les Comunicacions de la Universitat Autònoma de Barcelona.

RESUM**:**

L'ús d'eines d'IA generativa com ChatGPT ha revolucionat la manera com aprenem i ens relacionem amb la informació. Aquestes tecnologies són capaces de generar textos, respondre preguntes i fins i tot crear continguts nous a partir de dades prèvies, oferint un gran suport en entorns acadèmics. Amb xarxes neuronals profundes i algoritmes de machine learning, aquestes eines poden analitzar grans quantitats d'informació, proporcionant respostes i solucions de forma ràpida i aparentment precisa.

En aquesta xerrada, desmitificarem i entendrem amb detall i rigor com funcionen aquestes tecnologies, per tal de poder utilitzar-les de manera adequada. Això ens permetrà treure'n el màxim profit, mantenint el nostre esperit crític i assegurant-nos que fem un ús responsable d'aquestes eines en la nostra educació i en la nostra vida quotidiana.

A més, explorarem la importància de ser crítics amb la informació que ens ofereix la IA generativa, ja que pot contenir errors o dades incorrectes. També parlarem sobre com utilitzar aquestes eines de manera responsable en l'àmbit acadèmic, per complementar l'aprenentatge sense substituir l'esforç personal. Discutirem les oportunitats d'aprenentatge que ofereixen aquestes tecnologies, però també advertirem sobre els riscos de caure en la il·lusió del coneixement, on es confon l'accés fàcil a la informació amb una comprensió real i profunda.

Preguntes:

• És fiable la informació generada per la IA?

• Podem saber quines són les fonts d'informació que utilitzen les IA

generatives per generar les seves respostes?

• Podem utilitzar la IA de manera responsable en les nostres tasques acadèmiques?

• Com puc ser més exigent amb el meu procés d'aprenentatge i aprofitar millor les oportunitats que ofereix la IA generativa?

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

Valle Ruiz-Fernández és graduada en Traducció i Interpretació així com en Llengües Aplicades per la Universitat Pompeu Fabra

RESUM

Els models de llenguatge, com el famós ChatGPT, han irromput a la societat amb un gran ventall d'usos i aplicacions. Des de la creació de contingut i la redacció de textos fins a la traducció o la programació, aquests models s'han convertit en aliats en molts sectors i han transformat la manera com treballem, aprenem i interactuem amb la tecnologia. Tot i això, l'ús que en fem també planteja riscos i reptes, com els biaixos que poden presentar les seves respostes i l'impacte que poden tenir en la societat.

Entendrem com es desenvolupen o "s'entrenen" els models de llenguatge a partir de grans quantitats de dades i ens centrarem en un dels desafiaments més importants que sorgeixen en el desenvolupament d'aquests models: els biaixos. Explorarem què són els biaixos en els models de llenguatge, com s'introdueixen durant l'entrenament i quines conseqüències poden tenir. Reflexionarem sobre la importància de crear models més justos i inclusius, presentant exemples de casos en què aquests biaixos afecten diferents col·lectius de la societat. Finalment, discutirem estratègies per mitigar aquests biaixos i el paper que tenen les persones encarregades de desenvolupar els models de llenguatge en tot aquest procés.

PREGUNTES

- Creus que els biaixos que presenten els models de llenguatge són un risc per a la societat? En quin sentit?
- Com creus que podem identificar si un model està donant una resposta esbiaixada?
- Com influeixen les dades que es fan servir per entrenar els models en les respostes que després donen als usuaris?
- Si entrenem un model principalment amb contingut d'Internet (pàgines web, blogs, notícies, xarxes socials, etc.), quins exemples de biaixos creus que podríem trobar en les respostes que generi?
- Què podríem fer per reduir els biaixos en els models de llenguatge?
- Creus que és possible desenvolupar un model de llenguatge sense cap biaix?
- Què creus que és millor? a) que els models no estiguin esbiaixats i no tinguin cap noció sobre el que és injust o discriminatiu, b) que els

	models hagin après biaixos però tinguin maneres de no mostrar-los en les respostes. **Aquestes sessions suposen una dedicació de 1 hora lectiva cadascuna**

ETAPA 4: TREBALL POSTERIOR A L'AULA DESPRÉS DE LA VISITA DEL/DE LA CIENTÍFIC/A
Lloc: escoles participants
Persones destinatàries: professorat del centre directament implicat en el projecte, que conduirà el debat amb l'alumnat de 3r d'ESO o coordinarà aspectes organitzatius.
A partir de les preguntes plantejades, de les informacions aportades en la xerrada amb el científic/científica i del debat que s'hi generi, els alumnes debaten i n'extreuen conclusions conjuntament amb el professor o professora de l'aula. Aquest treball suposa una dedicació mínima **d'una hora lectiva**.

ETAPA 5: TROBADA FINAL	
Lloc: CosmoCaixa	
Persones destinatàries: alumnat i professorat de cada una de les escoles.	
Sessió 1 12/12/2024	S'hi establiran 4 grups de treball, un per cada subtema. Cada grup de treball estarà format per 2 representants de cadascun dels 5 instituts i treballaran en paral·lel. Els **alumnes exposaran les seves conclusions i les debatran entre ells en cada grup de treball.** Aquest debat serà conduït amb suport extern. Fruit d'aquest debat sorgirà un document comú de conclusions. A continuació, dos **representants de cada grup de treball presentaran públicament les conclusions a les quals han arribat**. L'acte clourà amb un **Ted Talk** (per definir)

<u>Breu CV dels científics participants</u>

Vicent Costa és científic titular a l'Institut d'Investigació en Intel·ligència Artificial del CSIC, vocal del Consell Rector de l'Associació Catalana d'Intel·ligència Artificial, membre del Grup d'Estudis Humanístics sobre la Ciència i la Tecnologia i professor substitut a la Universitat Autònoma de Barcelona (grau en Intel·ligència Artificial i grau en Filosofia). Doctor en Ciència Cognitiva i Llenguatge, màster en

Lògica Pura i Aplicada, llicenciat en Filosofia i graduat en Matemàtiques, la seua recerca interdisciplinària inclou l'estudi de lògiques per a la intel·ligència artificial, el disseny de models explicables d'intel·ligència artificial i l'anàlisi de qüestions ètiques i filosòfiques vinculades a aquesta. Ha realitzat estades de recerca al Centre de Cognició Espacial (Universitat de Bremen) i al Centro Singular de Investigación en Tecnoloxías Intelixentes (Universidade de Santiago de Compostela). Coordina el consell editorial de la revista *NODES*, ha publicat, juntament amb E. Senabre, el llibre *Intel·ligència artificial. Com els algorismes condicionen les nostres vides* i és el COO del lloc web *Matesfácil*

Roger Lera va graduar-se en Física per la Universitat de Barcelona l'any 2020. A més, es va especialitzar en ciència de dades i intel·ligència artificial cursant el màster en Modelització per a la Ciència i l'Enginyeria de la Universitat Autònoma de Barcelona. Actualment, és estudiant de doctorat en intel·ligència artificial a l'Institut d'Investigació en Intel·ligència Artificial del Consell Superior d'Investigacions Científiques (IIIA-CSIC). Entre els seus interessos científics hi ha la Intel·ligència Artificial Ètica i Explicable, així com també la resolució de problemes d'optimització aplicats a escenaris reals.

Carlos Borrego Iglesias és professor agregat al Departament d'Enginyeria de la Informació i les Comunicacions de la Universitat Autònoma de Barcelona. Es va llicenciar en Enginyeria Informàtica per la Facultat d'Informàtica de la Universitat Politècnica de Madrid (UPM). Després d'acabar els seus estudis va treballar per l'Organització Europea per a la Recerca Nuclear (CERN) (Ginebra, Suïssa) i per l'Università La Sapienza (Roma, Itàlia). Des de l'any 2006 és investigador i docent a la UAB on va obtenir el seu doctorat en Informàtica. És coordinador d'innovació docent de l'Escola d'Enginyeria. Els seus interessos de recerca actuals inclouen protocols de xarxes per preservar la privadesa, amb una perspectiva d'aprenentatge automàtic.

Valle Ruiz-Fernández és graduada en Traducció i Interpretació així com en Llengües Aplicades per la Universitat Pompeu Fabra. Va endinsar-se en la lingüística computacional i el processament del llenguatge natural amb el màster en Anàlisi i Processament del Llenguatge a la Universitat del País Basc. Actualment treballa com a investigadora a la unitat de Tecnologies del Llenguatge del Barcelona Supercomputing Center, on la seva recerca se centra en els models de llenguatge i, més concretament, l'avaluació dels biaixos que presenten.