Apunts CE_5071 1.1

Iloc:	Institut d'Ensenyaments a Distància de les Illes Balears	Imprès per:	David Ramirez Ruiz
Curs:	Models d'intel·ligència artificial	Data:	dissabte, 26 d'octubre 2024, 11:43
Llibre:	Apunts CE_5071 1.1		

Taula de continguts

1. Fonaments dels sistemes intel·ligents

Ens feim dir *Homo Sapiens* -l'home que sap, l'home savi- per la importància que donam a la nostra **intel·ligència**. De fa milers d'anys, hem mirat d'entendre com pensam. Això és, com una agregació de matèria pot percebre, entendre, predir i manipular un món molt més gran i complicat que ella mateixa. El camp de la **intel·ligència artificial** (IA en català o AI, *artificial intelligence*, en anglès), encara va més enfora: no cerca només entendre sinó, a més, **construir** entitats intel·ligents.

La IA és un dels camps més nous de la ciència i la tecnologia. S'hi començà a fer feina després de la Segona Guerra Mundial, i el nom (intel·ligència artificial) es va proposar el 1956. Així com en física, per exemple, hi ha tot d'idees fonamentals establertes des de Galileu, Newton o Einstein, en intel·ligència artificial encara hi ha molt d'espai per aportacions fonamentals comparables.

La intel·ligència artificial inclou una gran varietat de subcamps, des dels més generals (aprenentatge, percepció) fins als més específics (jugar a escacs o Go, demostrar teoremes, escriure poesia, conduir un cotxe en un carrer estibat de gent o diagnosticar malalties). La IA és rellevant en relació a qualsevol tasca intel·lectual; és veritablement un camp universal.

2. Què és la Intel·ligència Artificial?

Seguint Stuart Russell i Peter Norvig, al seu llibre Artificial Intelligence: A Modern Approach, podem estructurar les definicions d'intel·ligència artificial en una taula en dues dimensions, que defineix quatre quadrants.

Pensar humanament	**Pensar racionalment**
L'intent de fer pensar els computadors... **màquines amb ment**, en sentit complet i literal (Haugeland, 1985) L'automatització d'activitats que associam amb el pensament humà, activitats com la prendre decisions, resoldre problemes, aprendre... (Bellman, 1978)	L'estudi de les facultats mentals mitjançant l'ús de models computacionals (Charniak i McDermott, 1985) L'estudi dels càlculs que fan possible percebre, raonar i actuar (Winston, 1992)
Actuar humanament	**Actuar racionalment**
L'art de crear màquines que realitzin funcions que requereixen intel·ligència quan les fan persones (Kurzweil, 1990) L'estudi de com fer que les computadores facin coses que, de moment, les persones fan més bé (Rich and Knight, 1991)	La intel·ligència computacional és l'estudi del disseny dels **agents intel·ligents** (Poole et al., 1998) La IA es refereix al **comportament intel·ligent** en els artefactes (Nilsson, 1998)

Taula: *Diverses definicions d'intel·ligència artificial, organitzades en quatre categories*

Les definicions de la filera de dalt es refereix als processos de **pensament** i al **raonament**, mentre que les de la filera de baix es refereixen a la **conducta**.

Les definicions de la columna de l'esquerra mesuren l'èxit prenent la referència **humana**, mentre que les de la dreta prenen com a referència la **racionalitat**. Un sistema és considerat racional si fa l'**acció correcta**, donat el que sap.

No hi ha una definició única d'intel·ligència artificial, sinó que tot i que ja se n'han donades moltes, encara és objecte de debat.

Ramon López de Mántaras distingeix tres tipus d'IA.

1. IA estreta, la de les aplicacions habituals actuals
2. IA general, la que pot dur a terme en el futur qualsevol tasca, a un nivell humà
3. IA forta, la que apareix a la ciència-ficció

La definició d'intel·ligència artificial arrossega el problema de la psicologia de la definició d'intel·ligència. A més, per la banda de l'artificialitat, la IA és poc artificial per la seva necessitat d'una gran quantitat de treball humà, en l'anotació de dades i en l'aprenentatge per reforçament a partir de retroacció humana.

Llorenç Valverde, al seu llibre "La seducció de les noves tecnologies", es fa ressò d'una definició curiosa d'IA trobada a la web.

> Conjunt de disciplines i de tècniques que tracten d'aconseguir que els ordinadors reals arribin a fer les mateixes coses que fan els que surten a les pel·lícules.

Una de de les definicions més recents és la que dona **Pei Wang** (1995), recollida al seu article On Defining Artificial Intelligence:

> Intel·ligència és la capacitat d'un sistema de processament d'informació d'adaptar-se al seu entorn, operant amb coneixement i recursos insuficients

Aquesta definició de Pei Wang ha originat tot un seguit de comentaris per part de diversos investigadors, entre els quals **François Chollet**, **Shane Legg** i **Richard Sutton**, recollits al número especial del Journal of Artificial General Intelligence On Defining Artificial Intelligence.

3. Fonts de la IA

En la intel·ligència artificial conflueixen moltes disciplines, que hi han contribuït idees, punts de vista i tècniques.

- Filosofia
- Matemàtiques
- Economia
- Neurociència
- Psicologia
- Enginyeria informàtica
- Teoria de control i cibernètica
- Lingüística

Aquest origen múltiple es pot expressar al voltant d'una sèrie de qüestions, que detallam a continuació

3.1. Filosofia

En **filosofia**, es plantegen les preguntes següents.

- Es poden usar regles formals per obtenir conclusions vàlides?
- Com emergeix la ment a partir d'un cervell físic?
- D'on prové el coneixement?
- De quina manera el coneixement mena a l'acció?

Després que **Aristòtil** definís un sistema de sil·logismes per al raonament, **Ramon Llull** tingué la idea que el raonament es podia dur a la pràctica amb un artefacte mecànic. **Hobbes** proposà que el raonament era com un càlcul numèric. Amb l'automatització de la computació ja en marxa, **Leonardo Da Vinci** dissenyà una calculadora mecànica, però no la construí. La primera calculadora mecànica coneguda és de Wilhelm Schickard, el segle XVII. Pascal construí la seva Pascalina, que efectuava sumes i restes. Leibniz construí un dispositiu per realitzar operacions sobre conceptes en comptes de nombres, i construí una calculadora que realitzava sumes, restes, productes, divisions i arrels.

La filosofia ha exercit un paper fonamental en el desenvolupament de la intel·ligència artificial (IA) des dels seus inicis, aportant una rica font de reflexions i qüestions que han influït en la recerca i la pràctica d'aquest camp.

Les preguntes filosòfiques sobre la naturalesa de la ment, la consciència i la intel·ligència han inspirat els investigadors a explorar com es poden modelar processos cognitius humans mitjançant sistemes artificials. La teoria de la ment, per exemple, indaga sobre el que significa pensar i ser conscient, i aquesta exploració ha influït directament en la creació de models computacionals que busquen simular la cognició humana.

Un altre aspecte clau és la lògica i el raonament, que constitueixen una base essencial per als sistemes de raonament automàtic. La lògica formal proporciona les eines necessàries per desenvolupar algoritmes que permeten a les màquines prendre decisions basades en principis lògics. Així mateix, l'ètica s'ha convertit en un tema cada vegada més rellevant a mesura que la IA avança. La filosofia planteja qüestions sobre la responsabilitat i la moralitat de les màquines intel·ligents, guiant el desenvolupament d'una IA que sigui ètica i responsable.

A més, l'epistemologia, que estudia com adquirim i processem el coneixement, influeix en el disseny de sistemes d'aprenentatge automàtic. Comprendre com els humans aprenen i comprenen el món és essencial per crear màquines que puguin fer-ho de manera efectiva. La filosofia del llenguatge també és crucial, ja que analitza la relació entre llenguatge, significat i pensament; aquest estudi és fonamental per al processament del llenguatge natural en sistemes d'IA.

Diversos filòsofs han tingut un impacte significatiu en el pensament sobre la IA. **Alan Turing**, amb el seu famós **Test de Turing**, va establir fonaments importants per a la computació i la comprensió del pensament artificial. **John Searle** va introduir l'experiment de l'"Habitació Xinesa", que desafia les nocions d'IA forta, ja que suggereix que les màquines poden simular el pensament sense realment entendre'l. **Daniel Dennett** ha aportat teories sobre la consciència i la intencionalitat que ajuden a aprofundir en les implicacions filosòfiques de crear màquines intel·ligents.

Actualment, hi ha debats filosòfics importants en curs dins del camp de la IA. El contrast entre **IA forta** i **IA feble** continua generant discussions sobre si una màquina pot realment pensar o només simular el pensament. A més, el problema de l'**alineament** esdevé cada cop més crític: com podem assegurar-nos que els objectius de les màquines intel·ligents s'alineen amb els valors humans? Finalment, les qüestions sobre si és possible crear màquines veritablement conscients són temes candents que desafien tant els científics com els filòsofs.

La filosofia no només proporciona un marc conceptual per encarar qüestions fonamentals sobre la naturalesa de la intel·ligència, sinó que també orienta el desenvolupament pràctic de sistemes d'IA. Les reflexions filosòfiques continuen sent una font vital d'idees i debats en aquest camp dinàmic i en constant evolució.

3.2. Matemàtiques

En **matemàtiques**,

- Quines són les regles formals per obtenir conclusions vàlides?
- Què és computable?
- Com podem raonar amb informació incerta?

L'aportació de les matemàtiques a la intel·ligència artificial va des de la lògica de Boole fins a l'estadística de Bayes, base del raonament a partir d'informació incerta.

Les matemàtiques constitueixen una font fonamental i indispensable per al desenvolupament de la intel·ligència artificial (IA). Aquesta disciplina proporciona els fonaments teòrics i les eines pràctiques necessàries per crear, analitzar i optimitzar els algoritmes i models que són el cor de qualsevol sistema d'IA.

L'**àlgebra lineal**, per exemple, és essencial en el processament de dades multidimensionals i en la implementació d'algoritmes d'aprenentatge automàtic. Les matrius i els vectors, conceptes centrals de l'àlgebra lineal, s'utilitzen extensament en la representació i manipulació de dades, així com en la construcció de models predictius. Aquestes estructures matemàtiques permeten als sistemes d'IA processar grans quantitats d'informació de manera eficient i efectuar transformacions complexes sobre aquestes dades.

El **càlcul**, per la seva banda, juga un paper crucial en l'optimització dels models d'IA. Les tècniques de gradient descendent, fonamentals en l'entrenament de xarxes neuronals, es basen en principis del càlcul diferencial. Aquestes tècniques permeten ajustar els paràmetres dels models per minimitzar l'error i millorar el rendiment. A més, el càlcul multivariable és essencial per entendre i implementar algoritmes d'aprenentatge profund, que són la base de moltes de les aplicacions més avançades d'IA en l'actualitat.

La teoria de la **probabilitat** i l'**estadística** són igualment crucials en el camp de la IA. Aquestes branques de les matemàtiques proporcionen els fonaments per a la presa de decisions sota incertesa, un aspecte central de molts sistemes intel·ligents. Els models probabilístics, com les xarxes bayesianes, permeten als sistemes d'IA raonar sobre esdeveniments incerts i actualitzar les seves creences basant-se en noves evidències. L'estadística, d'altra banda, ofereix mètodes per analitzar i interpretar les dades, essencials per avaluar el rendiment dels models d'IA i per extreure coneixements significatius de grans conjunts de dades.

La **teoria de la informació**, desenvolupada per Claude Shannon, ha tingut un impacte profund en la IA. Aquesta branca de les matemàtiques proporciona un marc per entendre la transmissió i el processament de la informació, conceptes fonamentals en el disseny de sistemes de comunicació i aprenentatge. La noció d'entropia, per exemple, s'utilitza en algoritmes de presa de decisions i en la compressió de dades, aspectes crítics en moltes aplicacions d'IA.

La lògica matemàtica, incloent-hi la **lògica proposicional** i la **lògica de primer ordre**, és la base dels sistemes de raonament automàtic i de la representació del coneixement en IA. Aquestes eines permeten formalitzar el raonament i la inferència, facilitant la creació de sistemes experts i motors de raonament que poden arribar a conclusions lògiques basades en un conjunt de premisses.

La **teoria de grafs**, una altra àrea important de les matemàtiques, s'aplica en diversos aspectes de la IA, com ara en l'anàlisi de xarxes socials, en algoritmes de cerca i en la representació de coneixement. Els grafs proporcionen una manera poderosa de modelar relacions complexes entre entitats, cosa que és crucial en moltes aplicacions d'IA, des de sistemes de recomanació fins a l'anàlisi de dades interconnectades.

En resum, les matemàtiques no són simplement una eina auxiliar en el desenvolupament de la IA, sinó que constitueixen el seu llenguatge fonamental. Proporcionen el rigor, la precisió i els mètodes necessaris per transformar idees abstractes sobre la intel·ligència en sistemes funcionals i eficaços.

3.3. Economia

En **economia**, la llista de qüestions rellevants per a la IA inclou les següents.

- Com hem de prendre decisions per maximitzar el benefici?
- Com ho fem si els altres no hi col·laboren?
- Com ho fem si el benefici s'obté en un futur llunyà?

En aquesta disciplina, cal destacar l'aportació del Nobel d'economia de **Herbert Simon** el 1978, per la seva teoria de la satisfacció, que explica que les decisions humanes sovint no cerquen el resultat màxim teòric, sinó simplement una solució prou bona.

La teoria de la satisfacció ha tingut una influència significativa en el desenvolupament de la intel·ligència artificial (IA). Aquesta teoria, que es basa en la idea que els agents prenen decisions per maximitzar la seva satisfacció o utilitat donades certes restriccions, ha contribuït a la manera com es dissenyen i s'implementen els sistemes d'IA.

Un concepte clau que l'economia ha aportat a la IA és el de "racionalitat limitada", introduït per Herbert Simon. Aquest concepte reconeix que els agents tenen limitacions en la seva capacitat de processar informació i prendre decisions perfectament racionals. En conseqüència, sovint cerque solucions que siguin "prou bones" o satisfactòries, en lloc de l'òptim absolut.

Aquesta idea ha influït en el desenvolupament d'algoritmes d'IA que troben solucions satisfactòries en un temps raonable, especialment útils en problemes complexos on trobar la solució òptima podria ser computacionalment inviable. També ha influït en el disseny de sistemes d'IA per a la presa de decisions en entorns incerts, com en l'aprenentatge per reforç.

A més, l'economia ha proporcionat models matemàtics i eines d'optimització àmpliament utilitzats en IA, com la programació lineal i no lineal. La teoria de jocs també ha tingut un impacte significatiu, especialment en àrees com la negociació automatitzada i la presa de decisions en entorns multiagent.

3.4. Neurociència

Del punt de vista de la **neurociència**, la pregunta central és aquesta:

- Com processen la informació els cervells?

La conclusió principal de la disciplina és que una col·lecció de cèl·lules simples porta al pensament, l'acció i la consciència. En paraules del filòsof **John Searle**, els cervells creen les ments.

La neurociència ha tingut una influència significativa en el desenvolupament de la intel·ligència artificial (IA), proporcionant inspiració i coneixements sobre el funcionament del cervell humà que han ajudat a millorar els sistemes d'IA. Alguns aspectes clau de la contribució de la neurociència a la IA són:

1. Models neuronals artificials: La neurociència ha inspirat el desenvolupament de xarxes neuronals artificials, que imiten l'estructura i el funcionament de les neurones biològiques. Aquestes xarxes són la base de molts sistemes d'aprenentatge profund actuals.
2. Processament d'informació: L'estudi de com el cervell processa i integra informació ha influït en el disseny d'arquitectures d'IA més eficients i adaptatives.
3. Aprenentatge i memòria: La comprensió dels mecanismes d'aprenentatge i memòria del cervell ha contribuït al desenvolupament d'algoritmes d'aprenentatge automàtic més avançats.
4. Percepció i reconeixement de patrons: La investigació sobre com el cervell percep i reconeix patrons visuals i auditius ha ajudat a millorar els sistemes de visió per computador i processament del llenguatge natural.
5. Presa de decisions i raonament: Els estudis sobre els processos cognitius humans han influït en el desenvolupament de sistemes d'IA capaços de prendre decisions més complexes i contextuals.
6. Interfícies cervell-màquina: La neurociència ha permès el desenvolupament d'interfícies que connecten directament el cervell amb dispositius electrònics, obrint noves possibilitats per a la IA.
7. Ètica i implicacions socials: La neurociència també ha contribuït a la reflexió sobre les implicacions ètiques i socials de la IA, especialment en relació amb la privacitat i els drets humans.

Tot i això, és important destacar que la relació entre la neurociència i la IA és bidireccional. Mentre la neurociència inspira nous enfocaments en IA, els avenços en IA també proporcionen noves eines i models per a l'estudi del cervell. Aquesta sinergia continua impulsant avenços en ambdós camps, suggerint un futur on la comprensió del cervell i el desenvolupament de la IA estiguin cada vegada més interconnectats.

3.5. Psicologia

La **psicologia** es planteja

- Com pensen i actuen humans i animals?

Tot i que aquest punt de vista no és hegemònic en psicologia, sí que és una visió habitual que una teoria cognitiva ha de ser com un programa de computadora.

La psicologia ha tingut una influència significativa en el desenvolupament de la intel·ligència artificial (IA), proporcionant coneixements i models sobre el comportament i els processos cognitius humans que han inspirat i guiat el disseny de sistemes d'IA. Alguns aspectes clau de la contribució de la psicologia a la IA són:

1. Models cognitius: La psicologia cognitiva ha proporcionat models sobre com els humans processen la informació, prenen decisions i resolen problemes. Aquests models han inspirat el desenvolupament d'algoritmes i arquitectures en IA que imiten els processos cognitius humans.

2. Aprenentatge i memòria: Les teories psicològiques sobre l'aprenentatge i la memòria han influït en el disseny d'algoritmes d'aprenentatge automàtic i sistemes de representació del coneixement en IA.

3. Percepció i reconeixement de patrons: Els estudis psicològics sobre com els humans perceben i reconeixen patrons visuals i auditius han contribuït al desenvolupament de sistemes de visió per computador i processament del llenguatge natural.

4. Raonament i presa de decisions: La comprensió psicològica dels processos de raonament i presa de decisions humans ha ajudat a desenvolupar sistemes d'IA capaços de prendre decisions més complexes i contextuals.

5. Emocions i intel·ligència emocional: La recerca en psicologia sobre les emocions i la intel·ligència emocional ha inspirat el desenvolupament de sistemes d'IA que poden reconèixer i respondre a les emocions humanes.

6. Interacció humà-màquina: Els principis de la psicologia s'han aplicat per millorar la interacció entre humans i sistemes d'IA, fent-los més intuïtius i fàcils d'utilitzar.

7. Personalització: La comprensió psicològica de les diferències individuals ha contribuït al desenvolupament de sistemes d'IA capaços d'adaptar-se a les necessitats i preferències individuals dels usuaris.

La integració de principis psicològics en la IA ha permès crear sistemes més sofisticats i adaptatius, capaços d'interactuar de manera més natural amb els humans i abordar tasques complexes que requereixen una comprensió del comportament i la cognició humana.

3.6. Enginyeria informàtica

L'**enginyeria informàtica** encara el problema següent:

- Com podem construir una computadora eficient?

Per tal que la intel·ligència artificial vagi endavant, hem de menester dues coses: intel·ligència i un artefacte. L'artefacte triat ha estat la computadora.

L'enginyeria informàtica ha estat una font fonamental per al desenvolupament de la intel·ligència artificial (IA), proporcionant les bases tecnològiques i metodològiques necessàries per a la seva implementació i avenç. El desenvolupament d'**algoritmes i estructures de dades** eficients ha estat crucial per crear sistemes d'IA capaços de processar grans quantitats d'informació de manera ràpida i efectiva. Aquests avenços han permès la creació de models d'IA cada vegada més sofisticats i potents.

L'**arquitectura de computadors** ha jugat un paper vital en l'evolució de la IA. El disseny de processadors més potents i sistemes de computació especialitzats ha fet possible l'execució de models d'IA complexos que abans eren inviables. Paral·lelament, els avenços en **llenguatges de programació i paradigmes** han facilitat enormement la implementació d'algoritmes d'IA, permetent als desenvolupadors crear sistemes més flexibles i adaptables.

La gestió de dades és un altre aspecte crucial on l'enginyeria informàtica ha contribuït significativament. Les tècniques avançades de **bases de dades** i sistemes de gestió de grans volums d'informació són essencials per emmagatzemar i processar les immenses quantitats de dades necessàries per entrenar models d'IA moderns. Aquestes capacitats han obert noves possibilitats en àrees com l'aprenentatge profund i l'anàlisi predictiva.

Les **xarxes i sistemes distribuïts** han ampliat l'abast de la IA, permetent la seva implementació en entorns distribuïts i en el núvol. Això ha facilitat l'accés a recursos de computació massiva i ha fet possible la creació de sistemes d'IA escalables i accessibles globalment. A més, l'**enginyeria del programari** ha proporcionat metodologies robustes per desenvolupar i mantenir sistemes d'IA complexos, assegurant la seva fiabilitat i eficiència.

La seguretat informàtica i les interfícies d'usuari són altres àrees on l'enginyeria informàtica ha tingut un impacte significatiu en la IA. Les tècniques de **seguretat** són crucials per protegir la integritat i privacitat dels sistemes d'IA, mentre que el **disseny d'interfícies** intuïtives facilita la interacció entre humans i màquines intel·ligents.

Finalment, la **computació d'alt rendiment** ha estat fonamental per permetre l'entrenament de models d'IA cada vegada més complexos en temps raonables. Aquesta capacitat ha accelerat significativament el ritme de la recerca i el desenvolupament en IA.

Així, l'enginyeria informàtica no només proporciona les eines i tècniques necessàries per implementar sistemes d'IA, sinó que també impulsa constantment els límits del que és possible en aquest camp. La seva contribució contínua és essencial per al progrés i l'aplicació pràctica de la intel·ligència artificial en el món real.

3.7. Teoria de control i cibernètica

La **teoria de control** i la **cibernètica** resolen

- Com poden operar les artefactes sota el seu propi control?

La **teoria del control** ha proporcionat conceptes crucials per a la IA. Aquesta teoria tracta sobre el comportament de sistemes dinàmics i com es poden controlar per aconseguir un estat desitjat. Els conceptes de retroalimentació, regulació i optimització derivats de la teoria del control han estat fonamentals en el disseny de sistemes d'IA adaptatius i autoregulats.

La noció de **retroalimentació** (en anglès, *feedback*), en particular, ha estat essencial en el desenvolupament d'algoritmes d'aprenentatge automàtic. Els sistemes d'IA utilitzen mecanismes de retroalimentació per ajustar els seus paràmetres i millorar el seu rendiment al llarg del temps, de manera similar a com els sistemes de control utilitzen la retroalimentació per mantenir un estat desitjat.

La **cibernètica**, per la seva banda, ha aportat una visió sistèmica i interdisciplinària que ha influït profundament en la IA. La cibernètica estudia els sistemes de control i comunicació en éssers vius i màquines, proporcionant un marc conceptual per entendre i modelitzar sistemes complexos. Aquesta perspectiva ha estat crucial per desenvolupar models d'IA que imiten processos cognitius i de presa de decisions.

Els conceptes d'**autoregulació** i **homeòstasi** derivats de la cibernètica han inspirat el disseny de sistemes d'IA capaços d'adaptar-se a canvis en el seu entorn i mantenir un funcionament estable. Això és particularment evident en els sistemes d'IA que s'utilitzen en entorns dinàmics i incerts.

La integració de la teoria del control i la cibernètica en la IA ha donat lloc a avenços significatius en àrees com el **control adaptatiu** i l'**aprenentatge per reforç**. Aquests enfocaments permeten als sistemes d'IA aprendre i millorar el seu rendiment a través de la interacció amb l'entorn, utilitzant principis de retroalimentació i optimització.

A més, la teoria del control ha contribuït amb eines matemàtiques i metodologies de disseny que s'utilitzen en la creació i anàlisi de sistemes d'IA. Per exemple, les tècniques d'optimització i els criteris d'estabilitat desenvolupats en la teoria del control s'apliquen en l'entrenament i avaluació de models d'aprenentatge automàtic.

3.8. Lingüística

La **lingüística** s'ocupa d'estudiar

- Com es relaciona el **llenguatge** amb el **pensament**?

Els coneixements i models lingüístics han proporcionat la base teòrica i pràctica per crear sistemes d'IA capaços d'entendre i generar llenguatge humà.

Un dels conceptes clau que la lingüística ha aportat a la IA és el de **gramàtica formal**. Les gramàtiques formals, desenvolupades per lingüistes com **Noam Chomsky**, han permès modelitzar l'estructura del llenguatge de manera que les màquines puguin processar-lo. Aquestes gramàtiques han estat essencials per desenvolupar analitzadors sintàctics i altres eines de processament del llenguatge.

La **semàntica** és una altra àrea de la lingüística que ha tingut un impacte significatiu en la IA. Els models semàntics han ajudat a crear sistemes capaços d'entendre el significat de les paraules i les frases, no només la seva estructura. Això ha estat crucial per desenvolupar aplicacions com els sistemes de pregunta-resposta i els assistents virtuals.

La **pragmàtica**, que estudia com el context afecta el significat del llenguatge, també ha influït en el desenvolupament de sistemes d'IA més sofisticats. Aquesta branca de la lingüística ha ajudat a crear sistemes capaços d'entendre les intencions dels parlants i el significat implícit en les converses.

Els **corpus lingüístics**, grans col·leccions de text anotades lingüísticament, han estat fonamentals per entrenar models d'aprenentatge automàtic en tasques de processament del llenguatge natural. Aquests corpus proporcionen les dades necessàries per què els sistemes d'IA aprenguin patrons lingüístics i millorin la seva comprensió del llenguatge humà.

La **lingüística computacional** ha desenvolupat models i tècniques específiques per al processament automàtic del llenguatge, com els models estadístics del llenguatge i els algoritmes d'anàlisi sintàctica.

4. Història de la IA

Russell i Norvig distingeixen les etapes següents dins la història de la intel·ligència artificial.

- **(1943-1955) Gestació de la intel·ligència artificial**. El primer treball que es considera dins la intel·ligència artificial és el de McCulloch i Pitts el 1943. Es basa en tres fonts: el coneixement de la fisiologia bàsica i la funció de les neurones al cervell, l'anàlisi formal de la lògica proposicional i la teoria de la computació. A partir d'aquí, proposen un model de neurona artificial en què cada cada neurona pot estar activada o no (on/off) com a resposta a una estimulació d'un nombre suficient de neurones veïnes. McCulloch i Pitts demostraren que qualsevol funció es podia aproximar amb una xarxa neuronal prou complexa, i que els connectors lògics (not, or, and) es poden implementar en xarxes senzilles.
- **(1956) El naixement de la intel·ligència artificial**. Aquest any tingué lloc la **trobada d'estiu de Dartmouth** sobre intel·ligència artificial. És aquí on sorgeix el terme **intel·ligència artificial.**
- **(1952-1969) Entusiasme inicial**. Els primers resultats en sistemes d'intel·ligència artificial eren molt prometedors, amb un bon nombre de sistemes que realitzaven tasques sorprenents. Per exemple, **Arthur Samuel** va escriure un programa d'escacs que jugava millor que ell mateix; **John McCarthy** va definir el llenguatge **Lisp**, dominant durant trenta anys; els micromons proposats per **Marvin Minsky** resolien amb èxit tasques en dominis acotats.
- **(1966-1973) Una dosi de realitat.** Després dels èxits inicials, començaren a aparèixer tres tipus de problemes: els sistemes simplement aplicaven regles sintàctiques, sense saber res de la matèria que tractaven; molts de problemes que la IA mirava de resoldre resultaven intractables; a més, aparegueren limitacions estructurals en els sistemes que havien de generar comportament intel·ligent. La neurona simple no era capaç de decidir si les seves dues entrades són iguals o no. Com a conseqüència d'aquestes limitacions, l'interès i el finançament en xarxes neuronals es reduí pràcticament a zero.
- **(1969-1979) Sistemes basats en el coneixement.** En aquesta dècada tingueren èxit un conjunt de sistemes experts com el MYCIN, capaç de diagnosticar infeccions a la sang, mitjançant un conjunt de 450 regles. La indústria va florir, amb centenars de companyies construint sistemes experts, sistemes de visió o robots. Però aviat vingué un període anomenat l'**hivern de la IA**, per la impossibilitat de satisfer promeses exagerades.
- **(1986-avui) Retorn de les xarxes neuronals.** A mitjan dècada dels 1980, quatre grups diferents redescobriren l'algorisme de retropropagació per a entrenar xarxes neuronals, que havia estat proposat per Bryson i Ho el 1969. L'algorisme es va aplicar a molts de problemes en ciències de la computació i psicologia, i la disseminació del seu impacte va causar un gran entusiasme.
- **(1987-avui) Raonament probabilístic i aprenentatge automàtic.** En aquesta etapa destaca la invenció de la xarxa bayesiana com a formalisme per la representació i el raonament amb coneixement incert. Encoratjats pels èxits en la solució de problemes parcials, els investigadors han tornat a encarar el problema de l'agent complet, i un dels entorns més importants per als agents intel·ligents és Internet.
- **(2001-avui) Grans conjunts de dades.** Tradicionalment, l'èmfasi de la intel·ligència artificial ha estat en els algorismes, però això s'ha desplaçat cap a una major importància de les dades, sobretot d'ençà de la disponibilitat de grans fonts de dades.
- **2011 Deep Learning**. L'aprenentatge profund (en anglès *deep learning*) es refereix a l'ús de xarxes neuronals de moltes capes d'elements simples ajustables. Des de la dècada de 1970 s'experimentava amb aquestes xarxes, i en la dècada de 1990 va trobar un cert èxit amb l'ús de xarxes neuronals convolucionals per al reconeixement d'imatges de dígits manuscrits. Però no va ser fins el 2011 que l'aprenentatge profund es va enlairar realment , primer en reconeixement de la parla i després en el reconeixement d'objectes visuals.

5. Aplicacions

Què pot fer avui la intel·ligència artificial? La intel·ligència artificial es considera l'**electricitat del segle XXI**, un servei bàsic sobre el qual s'aniran basant com més va més aplicacions. Qualsevol resum quedarà incomplet, perquè hi ha moltes activitats en molts de subcamps. Aquí n'esmentarem algunes.

- **Vehicles robòtics**. Les primeres demostracions de conducció autònoma per carretera són de la dècada dels 1980. El 2018, l'empresa Waymo passà la fita de 10 milions de milles conduïdes a carreteres públiques sense cap accident seriós. Al medi aeri, s'utilitzen drons autònoms des del 2016.
- **Locomoció**. **BigDog**, un robot quadrúpede, canvià la nostra percepció de com es mou un robot, mostrant la capacitat de recuperar la posició quan se l'empeny o rellisca. **Atlas**, un robot humanoide, no només camina sobre terreny desigual sinó que també és capaç de botar damunt capses o realitzar acrobàcies.
- **Planificació autònoma**. L'any 2000, el programa Remote Agent de la NASA va ser el primer programa de planificació autònoma a bord per controlar una nau espacial. Avui, el toolkit de planificació EUROPA s'usa rutinàriament en els vehicles de Mart de la NASA i el sistema SEXTANT permet la navegació en l'espai profund, fora del sistema GPS global. Cada dia, companyies com Uber o el servei de Google Maps faciliten direccions de conducció a centenars de milions d'usuaris, mostrant ràpidament una ruta òptima que té en compte les condicions del tràfic actuals i futures.
- **Traducció automàtica**. Els sistemes de traducció en línia permeten la lectura de documents en més de 100 llengües. Tot i que encara imperfectes, en general són adequats com a mínim per a la comprensió.
- **Reconeixement de la parla**. Alexa, Siri, Cortana i Google ofereixen assistents que poden respondre preguntes i realitzar tasques per als usuaris.
- **Recomanacions**. Companyies com ara Amazon, Facebook, Netflix, Spotify, Youtube i d'altres usen l'aprenentatge automàtic per recomanar el que pot agradar els usuaris en funció de les experiències passades i de les d'altres persones de perfil similar. El filtratge antispam es pot classificar dins aquesta categoria de sistemes de recomanació.
- **Jocs**. Quan Deep Blue va derrotar el campió mundial d'escacs Garry Kasparov el 1997, els defensors de la supremacia humana tenien posades les seves esperances en el joc de Go, d'una complexitat molt més gran. 20 anys després, AlphaGo superà tots els jugadors humans. AlphaGo aprofitava el coneixement d'un gran nombre de partides i l'expertesa humana. Un programa posterior, AlphaZero, ja no usà cap informació humana llevat de les regles del joc, i en va aprendre exclusivament jugant contra ell mateix, fins aconseguir superar tots els oponents, humans i màquines, als jocs d'escacs, Go i shogi.
- **Comprensió d'imatges**. Hi ha sistemes de conversió d'imatge a text, que donen descripcions verbals a partir de la informació visual.
- **Medicina.** Els algorismes d'IA ja igualen o superen els doctors experts en la diagnosi de malalties, especialment si es basen en proves d'imatge. Els exemples inclouen la síndrome d'Alzheimer, el càncer amb metàstasi, malalties dels ulls i de la pell.
- **Climatologia.** Hi ha treballs que presenten múltiples formes en què l'aprenentatge automàtic es pot usar per fer front al canvi climàtic.

Podem seguir any rere any les novetats sobre intel·ligència artificial a l'informe que publica la Universitat de Stanford.

- https://aiindex.stanford.edu/report/

6. Tècniques de la IA

Cerca i optimització

Molts de problemes en IA es poden resoldre teòricament cercant de manera intel·ligent a través de moltes possibles solucions. El raonament es pot reduir a realitzar una cerca. Per exemple, una prova lògica es pot veure com la recerca d'un camí que condueixi des de les premisses a les conclusions, on cada pas és l'aplicació d'una regla d'inferència. Els algorismes de planificació cerquen a través d'arbres d'objectius i subobjectius, intentant trobar un camí cap a una fita objectiva, un procés anomenat anàlisi de mitjans-fins. Els algorismes de robòtica per moure extremitats i agafar objectes utilitzen cerques locals a l'espai de configuració.

Les cerques exhaustives senzilles poques vegades són suficients per a la majoria dels problemes del món real: l'espai de cerca (el nombre de llocs per cercar) creix ràpidament fins a nombres astronòmics. El resultat és una cerca massa lenta o mai completa. La solució, per a molts problemes, és utilitzar "heurístiques" o "regles empíriques" que prioritzen les eleccions a favor d'aquells camins que tenen una probabilitat més gran d'assolir un objectiu i fer-ho en un nombre més curt de passos. En algunes metodologies de cerca, les heurístiques també poden servir per eliminar algunes opcions que és poc probable que condueixin a un objectiu (poda de l'arbre de cerca). Les heurístiques proporcionen al programa una "suposició òptima" del camí en què es troba la solució. Les heurístiques limiten la cerca de solucions a una mida de mostra més petita.

Un tipus de cerca molt diferent va cobrar protagonisme a la dècada de 1990, basada en la teoria matemàtica de l'optimització. Per a molts problemes, és possible començar la cerca amb alguna forma d'inicialització i després refinar-la de manera incremental fins que no es puguin fer més perfeccionaments.

Lògica

Es fa servir la lògica per a la **representació del coneixement** i per a la **resolució de problemes**, però es pot aplicar també a d'altres problemes. Per exemple, l'algorisme **satplan** usa la lògica per planificar i la programació lògica inductiva com a mètode d'aprenentatge.

La IA fa servir diverses formes de lògica. La **lògica proposicional** implica funcions lògiques com la disjunció (*or*) o la negació (*not*). La **lògica de primer ordre** hi afegeix quantificadors i predicats i pot expressar fets sobre objectes, les seves propietats i les seves relacions. La **lògica difusa** assigna un grau de veritat (entre 0 i 1) a enunciats vagues com ara "N'Alícia és vella" (o rica, o alta), que són massa imprecisos lingüísticament per ser completament certs o falsos.

Mètodes probabilístics per al raonament amb incertesa

En molts de problemes d'IA (per exemple, raonament, planificació, aprenentatge, percepció i robòtica) l'agent ha d'operar amb informació incompleta o incerta. Per aconseguir-ho, s'adopten mètodes de la **teoria de la probabilitat** i de l'**economia**.

De la teoria de la probabilitat, les **xarxes bayesianes** són una eina general que s'aplica a diversos problemes.

- Raonament: inferència bayesiana
- Aprenentatge: algorisme EM
- Planificació: xarxes de decisió
- Percepció: xarxes bayesianes dinàmiques.

De l'economia, un concepte clau per a la IA és la **utilitat**, una mesura de com de valuosa és alguna cosa per a un agent intel·ligent.

Classificadors i mètodes estadístics d'aprenentatge

Les aplicacions d'IA més simples es poden dividir en dos tipus: **classificadors** (si brilla, aleshores és un diamant) i **controladors** (si és un diamant, aleshores agafem-lo). Els controladors, tanmateix, també classifiquen del condicions abans d'inferir accions. Per això, la **classificació** és una part central de molts de sistemes d'IA.

Un classificador es pot entrenar de diverses formes; hi ha moltes variants estadístiques i d'aprenentatge automàtic. L'**arbre de decisió** és l'algorisme més simple i més usat en aprenentatge automàtic simbòlic. L'algorisme **KNN** (***k-nearest neighbours***, k-veïns) va ser l'algorisme més usat fins a mitjans dels 1990. Llavors

passaren a dominar els **mètodes de kernel** com els **SVM** (*Support Vector Machine*). Actualment el mètode més usat és el **classificador naïf bayesià**. Les **xarxes neuronals** també es fan servir en classificació.

Xarxes neuronals artificials i **aprenentatge profund**

Les xarxes neuronals artificials s'inspiraren en el model biològic de les neurones del cervell humà. Una neurona rep com a entrada les sortides d'unes altres neurones, que poden estar activades o no. La contribució de totes les entrades es pondera i passa a través d'una funció d'activació, que pot ser no-lineal. Aquesta sortida serà una de les entrades d'una neurona de la següent capa, fins a la capa final de sortida. L'aprenentatge consisteix en l'ajust dels pesos mitjançant un algosirme de **descens de gradient**. La tècnica més habitual d'entrenament és la **retropropagació** (*backpropagation*).

En l'**aprenentatge profund** (*deep learning*) hi ha més d'una capa intermèdia de neurones entre la capa d'entrada i la de sortida de la xarxa. Això permet modelitzar relacions entrada-sortida més complexes, però també necessita una quantitat de dades d'entrenament més gran. La revifada recent de les xarxes neuronals ve de la mà de l'aprenentatge profund. Això ha estat possible gràcies als avenços al maquinari, que permeten entrenar sistemes més grans en un temps raonable.

Llenguatges especialitzats i **maquinari**

El primer **llenguatge** desenvolupat per a intel·ligència artificial va ser **Lisp**. Inclou característiques per realitzar resolució de problemes general, com ara llistes, associacions, *schemas*, assignació dinàmica de memòria, tipus de dades, recursió, recuperació associativa, generadors i multitasca cooperativa.

Prolog és un llenguatge declaratiu. Els programes s'expressen en termes de relacions, i s'executen peticions (*queries*) sobre aquestes relacions. Prolog és especialment útil per al raonament simbòlic, i aplicacions de bases de dades i d'anàlisi lingüística. Prolog és àmpliament usat en IA avui dia.

R té un gran ús en la IA actual, quan intervenen càlculs estadístics, anàlisi numèrica inferència bayesiana, xarxes neuronals i en general en aprenentatge automàtic (*machine learning*). És un dels llenguatges estàndard principal en camps com finances, biologia, sociologia o medicina. Ofereix diversos paradigmes de computació, com ara **computació vectorial**, **programació funcional** i **programació orientada a l'objecte**.

Python s'usa habitualment en projectes d'IA i aprenentatge automàtic amb l'ajuda de llibreries com ara TensorFlow, Keras, Pytorch i Scikit-learn. S'usa sovint en processament del llenguatge natural (NLP, *natural language processing*), pel fet que és un llenguatge de guions amb una arquitectura modular, sintaxi simple i eines avançades de processament de text. Aquest serà el principal llenguatge que utilitzarem en aquest curs d'especialització.

Pel que fa al **maquinari**, els **acceleradors d'IA** són una classe d'accelerador de maquinari especialitzat o sistema de computació dissenyat per accelerar les aplicacions d'IA i aprenentatge automàtic, incloent-hi les xarxes neuronals artificials i la visió per computador. Les aplicacions típiques són algorismes per a robòtica, IoT (*Internet of Things*) i altres tasques intensives en dades o basades en sensors. Un xip de circuits integrats d'AI contenia milers de milions de transistors MOSFET el 2018. Hi ha una sèrie de termes específics del fabricant per als dispositius d'aquesta categoria, i és una tecnologia emergent sense un disseny dominant encara.

Com a exemple molt usat, les **GPU** (*Graphic Processing Unit*, unitat de procés gràfic) són maquinari especialitzat per a la manipulació d'imatges i el càlcul de propietats locals d'imatges. Les bases matemàtiques de les xarxes neuronals i de la manipulació d'imatges són semblants, tasques altament paral·leles que impliquen matrius, cosa que fa les GPU siguin cada vegada més usades en tasques d'aprenentatge automàtic.

7. Riscos i beneficis de la IA

Ja el 1609, el filòsof Francis Bacon, a qui s'atribueix la creació del mètode científic, notà que "les arts mecàniques són d'ús ambivalent, i serveixen tant per al dany com per al remei". A mesura que la IA va prenent un paper cada cop més important en els camps social, científic, mèdic, financer i militar, hem de considerar aquests *danys i remeis* (modernament riscos i beneficis) que comporta.

Per començar amb els beneficis, la nostra civilització sencera és el producte de la intel·ligència humana. Si tenim accés a una intel·ligència automàtica més gran, el sostre de la nostra ambició també puja. El potencial de la IA per alliberar la humanitat del treball penós i incrementar dramàticament la producció de bens i serveis podria presagiar una era de pau i abundància. La capacitat d'accelerar la recerca científica podria donar com a resultat cures per a malalties i solucions davant el canvi climàtic o l'escassetat de recursos. Com ha dit Demis Hassabis, CEO de Google DeepMind: "primer resoldre la IA, després usar la IA per resoldre tota la resta".

Tanmateix, abans que tenguem l'oportunitat de resoldre la IA, haurem de fer front a riscos derivats del mal ús de la IA, involuntaris o no. Alguns dels següents ja són ben clars, d'altres semblen probables donades les tendències actuals.

- Armes autònomes letals.
- Vigilància i persuasió.
- Presa de decisions esbiaixada.
- Impacte en el desenvolupament.
- Seguretat crítica.
- Ciberseguretat.

A mesura que els sistemes d'IA van tornant més capaços, és previsible que vagin realitzant activitats abans reservades als humans. Els exemples anteriors mostren la necessitat d'una bona governança, i regulació.

I més a llarg termini? I si desenvolupam intel·ligència artificial comparable o superior a la humana?

La gran majoria d'investigadors en IA s'han especialitzat en subcamps, com ara els jocs, la representació del coneixement, la visió per computadora o el processament del llenguatge natural, assumint que el progrés en cada camp contribuiria als objectius generals de la IA.

Però alguns fundadors de la IA, com John McCarthy, Marvin Minstky i Patrick Winston suggerien que en comptes d'enfocar-se en el rendiment d'aplicacions específiques, la IA hauria de retornar a les seves arrels d'intentar construir, en paraules de Herb Simon, "màquines que pensen, que aprenen i que creen". Aquest objectiu rep el nom de **HLAI** (*human-level AI*, IA de nivell humà): una màquina hauria de ser capaç d'aprendre a fer qualsevol cosa que un humà pot fer.

Al mateix temps, es començaren a aixecar veus (Yudkowsky, Omohundro) que alertaven que la superintel·ligència artificial (Artificial Superintelligence, ASI) podria ser una mala idea.

Ja el 1960, Norbert Wiener, quan va veure el programa d'Arthur Samuel guanyar-li als escacs al seu creador, va manifestar que si per aconseguir els nostres objectius cream una agent mecànic que no podem interferir, hem d'estar ben segurs que el propòsit que persegueix sigui el que realment nosaltres volem.

Però la solució a la objecció de Wiener no pot ser introduir un propòsit dins la màquina. En canvi, necessitam màquines que cerquen aconseguir els objectius humans sabent que no saben exactament quins són aquests objectius.

Per exemple, una màquina té un incentiu positiu en deixar-se apagar si i només si té incertesa sobre quin és l'objectiu humà que ha d'aconseguir.

Els mètodes de l'**aprenentatge de reforç invers** (*inverse reinforcement learning*) permeten a les màquines aprendre les preferències humanes a base d'observar les accions de les persones, però això topa amb dos obstacles: les nostres eleccions depenen de les nostres preferències a través d'una arquitectura cognitiva difícil d'invertir; i els humans no és clar que tenguem unes preferències coherents, ni individualment ni com a grup, de forma que no seria clar què ha de fer exactament la IA per a nosaltres.

8. Ètica de la IA

Així com explica la UNESCO respecte de l'ètica de la intel·ligència artificial,

> Avui dia, la intel·ligència artificial té un paper important en la vida de milers de milions de persones. De vegades desapercebut però sovint amb conseqüències profundes, transforma les nostres societats i desafia el que significa ser humà.
>
> La IA pot proporcionar suport a milions d'estudiants per completar l'educació secundària, ocupar 3,3 milions de llocs de treball addicionals i, de manera més urgent, ajudar-nos a fer front a la propagació i les conseqüències de la pandèmia de la COVID-19. Juntament amb múltiples avantatges, aquestes tecnologies també generen riscos i reptes a la baixa, derivats de l'ús maliciós de la tecnologia o de l'aprofundiment de les desigualtats i les divisions.

Per això, fan falta

> Polítiques internacionals i nacionals i marcs reguladors per garantir que aquestes tecnologies emergents beneficiïn la humanitat en el seu conjunt.
>
> Una IA centrada en l'ésser humà. La IA ha de ser per l'interès més gran de la gent, no al revés.

En aquest context, el novembre de 2021, els 193 estats membres de la Conferència General de la UNESCO van adoptar la Recomanació sobre l'ètica de la intel·ligència artificial, el primer instrument mundial d'establiment de normes sobre el tema. Segons indica la institució, no només protegirà sinó també promourà els drets humans i la dignitat humana, i serà una brúixola guia ètica i una base normativa global que permetrà construir un fort respecte per l'estat de dret al món digital.

Dins les recomanacions de la UNESCO hi ha la següent enumeració de valors generals i la seva concreció en principis.

Valors

- Respecte, protecció i promoció dels drets humans i de les llibertats fonamentals i de la dignitat humana
- Desenvolupament del medi ambient i els ecosistemes
- Diversitat i inclusió
- Pau, justícia i interconnexió social

Principis

- Proporcionalitat i no fer mal
- Seguretat
- Equitat i no discriminació
- Sostenibilitat
- Dret a la privacitat i protecció de dades
- Supervisió i determinació humana
- Transparència i explicabilitat
- Responsabilitat i rendició de comptes
- Conscienciació i alfabetització
- Governança i col·laboració multilateral i adaptativa

9. IA confiable a la Unió Europea

El 8 d'abril de 2019, el grup d'experts d'alt nivell en IA (AI HLEG, High Level Expert Group) va presentar les directrius ètiques per a una intel·ligència artificial fiable. Això es va produir després de la publicació del primer esborrany de les directrius el desembre de 2018, sobre el qual es van rebre més de 500 comentaris mitjançant una consulta oberta.

Segons les directrius, la IA fiable ha de ser:

(1) lícita: respectant totes les lleis i regulacions aplicables

(2) ètica - respectant els principis i valors ètics

(3) robusta: tant des d'una perspectiva tècnica, tenint en compte el seu entorn social

Els sistemes d'IA han de millorar el benestar individual i col·lectiu. A continuació s'enumeren quatre principis ètics, arrelats en eldrets fonamentals, que s'han de respectar per garantir que els sistemes d'IA es desenvolupin, despleguen i s'utilitzen de manera fiable. S'especifiquen com a **imperatius ètics**, de manera que els professionals de la IA s'han d'esforçar sempre per complir-los. Sense imposar una jerarquia, enumerem els principis a continuació d'una manera que reflecteixi l'ordre d'aparició dels drets fonamentals en què es basen a la Carta de la UE.

Els quatre principis són els següents.

- Respecte per l'autonomia humana
- Prevenció de danys
- Equitat
- Explicabilitat

Principi de respecte a l'autonomia humana

Els drets fonamentals sobre els quals es fonamenta la UE estan dirigits a garantir el respecte a la llibertat i l'autonomia dels éssers humans. Els humans que interactuen amb els sistemes d'IA han de ser capaços de mantenir una autodeterminació completa i eficaç sobre ells mateixos i de poder participar en el procés democràtic. Els sistemes d'IA no haurien de subordinar, coaccionar, enganyar, manipular, condicionar o agrupar humans de manera injustificada. En canvi, haurien de dissenyar-se per augmentar, complementar i potenciar les habilitats cognitives, socials i culturals humanes. L'assignació de funcions entre humans i sistemes d'IA hauria de seguir els principis de disseny centrats en les persones i deixar una oportunitat significativa per a l'elecció humana. Això significa assegurar la supervisió humana sobre els processos de treball en sistemes d'IA. Els sistemes d'IA també poden canviar fonamentalment l'esfera de treball. Per això han de donar suport als humans en l'entorn de treball i apuntar a la realització de treball significatiu.

Principi de prevenció de danys

Els sistemes d'IA no han de causar ni agreujar danys ni afectar negativament els éssers humans. Això comporta la protecció de la dignitat humana així com de la integritat mental i física. Els sistemes d'IA i els entorns en què operen han de ser segurs. Han de ser tècnicament robusts i s'ha de garantir que no estan oberts a un ús maliciós. Les persones vulnerables haurien de rebre més atenció i s'han d'incloure en el desenvolupament, el desplegament i l'ús dels sistemes d'IA. També s'ha de prestar una atenció especial a les situacions en què els sistemes d'IA poden causar o agreujar els impactes adversos a causa de les asimetries de poder o informació, com ara entre empresaris i empleats, empreses i consumidors o governs i ciutadans. Prevenir els danys també implica tenir en compte el medi natural i tots els éssers vius.

Principi d'equitat

El desenvolupament, el desplegament i l'ús de sistemes d'IA ha de ser just. Tot i que hi ha moltes interpretacions diferents de l'equitat, l'equitat té una dimensió tant substantiva com de procediment.

La dimensió substantiva implica el compromís de: garantir una distribució equitativa i justa tant dels beneficis com dels costos, i garantir que les persones i els grups estiguin lliures de prejudicis injustos, discriminació i estigmatització. Si es poden evitar els biaixos injustos, els sistemes d'IA fins i tot podrien augmentar l'equitat

social. També s'ha de fomentar la igualtat d'oportunitats en termes d'accés a l'educació, els béns, els serveis i la tecnologia. A més, l'ús de sistemes d'IA no hauria de provocar mai que les persones siguin enganyades o perjudicades injustificadament en la seva llibertat d'elecció.

A més, l'equitat implica que els professionals de la IA han de respectar el principi de proporcionalitat entre mitjans i finalitats i considerar acuradament com equilibrar interessos i objectius contraposats.

La dimensió procedimental de l'equitat implica la capacitat de contestar i buscar una compensació efectiva contra les decisions preses pels sistemes d'IA i pels humans que els operen. Per fer-ho, l'entitat responsable de la decisió ha de ser identificable i els processos de presa de decisions han de ser explicables.

Principi d'explicabilitat

L'explicabilitat és crucial per construir i mantenir la confiança dels usuaris en els sistemes d'IA. Això significa que els processos han de ser transparents, les capacitats i el propòsit dels sistemes d'IA s'han de comunicar obertament i les decisions, en la mesura que sigui possible, explicables per als afectats directament i indirectament. Sense aquesta informació, una decisió no es pot impugnar degudament. No sempre és possible una explicació de per què un model ha generat una sortida o decisió determinada (i quina combinació de factors d'entrada hi ha contribuït). Aquests casos s'anomenen algorismes de "caixa negra" i requereixen una atenció especial. En aquestes circumstàncies, es poden requerir altres mesures d'explicabilitat (per exemple, traçabilitat, auditabilitat i comunicació transparent sobre les capacitats del sistema), sempre que el sistema en el seu conjunt respecti els drets fonamentals. El grau en què es necessita explicabilitat depèn molt del context i de la gravetat de les conseqüències si aquesta sortida és errònia o inexacta.

Els principis anteriors s'han de traduir en requisits concrets per aconseguir una IA fiable. Aquests requisits són aplicables a diferents parts interessades que participen en el cicle de vida dels sistemes d'IA: desenvolupadors, desplegadors i usuaris finals, així com a la societat en general. Per desenvolupadors, ens referim a aquells que investiguen, dissenyen i/o desenvolupen sistemes d'IA. Per desplegadors, ens referim a organitzacions públiques o privades que utilitzen sistemes d'IA dins dels seus processos empresarials i per oferir productes i serveis a altres. Els usuaris finals són aquells que participen amb el sistema d'IA, directament o indirectament. Finalment, la societat en general engloba totes les altres que es veuen afectades directament o indirectament pels sistemes d'IA.

Els diferents grups d'interessats tenen diferents papers per garantir que es compleixin els requisits:

a. Els desenvolupadors han d'implementar i aplicar els requisits als processos de disseny i desenvolupament;

b. Els desplegadors han de garantir que els sistemes que utilitzen i els productes i serveis que ofereixen compleixen els requisits;

c. Els usuaris finals i la societat en general han d'estar informats sobre aquests requisits i poden sol·licitar que es compleixin.

La llista de requisits següent no és exhaustiva. Inclou aspectes sistèmics, individuals i socials:

1 Agència humana i supervisió

Inclou drets fonamentals, agència humana i supervisió humana

2 Solidesa tècnica i seguretat

Incloent la resiliència als atacs i la seguretat, el pla de retrocés i la seguretat general, la precisió, la fiabilitat i la reproductibilitat

3 Privadesa i govern de dades

Incloent el respecte a la privadesa, la qualitat i la integritat de les dades i l'accés a les dades

4 Transparència

Inclou traçabilitat, explicabilitat i comunicació

5 Diversitat, no discriminació i equitat

Incloent l'evitació del biaix injust, l'accessibilitat i el disseny universal, i la participació de les parts interessades

6 Benestar social i ambiental

Incloent la sostenibilitat i el respecte al medi ambient, l'impacte social, la societat i la democràcia

7 Responsabilitat

Inclou auditabilitat, minimització i notificació d'impactes negatius, compensacions i reparacions.

- https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

9.1. Agència i supervisió humana

Els sistemes d'IA han de donar suport a l'autonomia humana i la presa de decisions, tal com prescriu el principi de respecte a l'autonomia humana. Això requereix que els sistemes d'IA actuïn com a facilitadors d'una societat democràtica, pròspera i equitativa donant suport a l'agència de l'usuari i fomentant els drets fonamentals i permetre la supervisió humana.

Drets fonamentals. Com moltes tecnologies, els sistemes d'IA poden tant permetre com dificultar el respecte als drets fonamentals. Poden beneficiar les persones, per exemple, ajudant-les a fer un seguiment de les seves dades personals o augmentant l'accessibilitat a l'educació, donant suport per tant al seu dret a l'educació. Tanmateix, donat l'abast i la capacitat dels sistemes d'IA, també poden afectar negativament els drets fonamentals. En situacions en què hi hagi aquests riscos, s'ha de dur a terme una avaluació de l'impacte dels drets fonamentals. Això s'ha de fer abans del desenvolupament del sistema i incloure una avaluació de si aquests riscos es poden reduir o justificar segons sigui necessari en una societat democràtica per tal de respectar els drets i les llibertats dels altres. A més, s'han de posar en marxa mecanismes per rebre comentaris externs sobre sistemes d'IA que puguin infringir els drets fonamentals.

Agència humana. Els usuaris han de poder prendre decisions autònomes informades sobre els sistemes d'IA. Se'ls han de proporcionar els coneixements i les eines per comprendre i interactuar amb els sistemes d'IA en un grau satisfactori i, quan sigui possible, poder-los avaluar o desafiar raonablement el sistema. Els sistemes d'IA han d'ajudar les persones a prendre decisions millors i més informades d'acord amb els seus objectius. De vegades, els sistemes d'IA es poden desplegar per donar forma i influir en el comportament humà mitjançant mecanismes que poden ser difícils de detectar, ja que poden aprofitar processos subconscients, incloses diverses formes de manipulació injusta, engany i condicionament, que poden amenaçar l'autonomia individual. El principi general d'autonomia de l'usuari ha de ser fonamental per a la funcionalitat del sistema. La clau d'això és el dret a no ser objecte d'una decisió basada únicament en un tractament automatitzat quan això produeixi efectes legals sobre els usuaris o els afecti de manera semblant significativament.

Supervisió humana. La supervisió humana ajuda a garantir que un sistema d'IA no soscavi l'autonomia humana ni provoqui altres efectes adversos. La supervisió es pot aconseguir mitjançant mecanismes de governança com ara un enfocament human-in-the-loop (HITL), human-on-the-loop (HOTL) o human-in-command (HIC).

HITL es refereix a la capacitat d'intervenció humana en cada cicle de decisió del sistema, que en molts casos no és ni possible ni desitjable.

HOTL fa referència a la capacitat d'intervenció humana durant el cicle de disseny del sistema i el seguiment del funcionament del sistema.

HIC fa referència a la capacitat de supervisar l'activitat global del sistema d'IA (incloent-hi el seu impacte econòmic, social, legal i ètic més ampli) i la capacitat de decidir quan i com utilitzar el sistema en qualsevol situació particular. Això pot incloure la decisió de no utilitzar un sistema d'IA, establir nivells d'intervenció humana durant l'ús del sistema o garantir la capacitat d'anul·lar una decisió presa per un sistema.

A més, s'ha de garantir que els agents públics tenguin la capacitat d'exercir la supervisió d'acord amb el seu mandat.

Es poden requerir mecanismes de supervisió en diferents graus per donar suport a altres mesures de seguretat i control, segons l'àrea d'aplicació del sistema d'IA i el risc potencial. Com més poca supervisió pugui exercir un ésser humà sobre un sistema d'IA, més exhaustives han de ser les proves.

9.2. Robustesa i seguretat

Un component crucial per aconseguir una IA fiable és la robustesa tècnica, que està estretament lligada al principi de prevenció de danys. La robustesa tècnica requereix que els sistemes d'IA es desenvolupin amb un enfocament preventiu dels riscos i d'una manera que es comportin de manera fiable com es pretén, alhora que es minimitzen els danys no intencionats i inesperats i es prevenen danys inacceptables. Això també s'ha d'aplicar a possibles canvis en el seu entorn operatiu o a la presència d'altres agents (humans i artificials) que puguin interactuar amb el sistema de manera contradictòria. A més, s'ha de garantir la integritat física i mental de les persones.

Resistència davant atacs i seguretat. Els sistemes d'IA, com tots els sistemes de programari, haurien d'estar protegits contra vulnerabilitats que poden permetre que siguin explotats pels adversaris, mitjançant intrusió. Els atacs es poden dirigir contra les dades (enverinament de dades), el model (fuga de models) o la infraestructura subjacent, tant de programari com de maquinari. Si s'ataca un sistema d'IA, per exemple en atacs adversaris, les dades i el comportament del sistema es poden modificar, fent que el sistema prengui decisions diferents o fent que es desactivi per complet. Els sistemes i les dades també es poden corrompre amb intenció maliciosa o per exposició a situacions inesperades. Els processos de seguretat insuficients també poden provocar decisions errònies o fins i tot danys físics. S'han de tenir en compte el sistema d'IA (per exemple, les aplicacions de doble ús) i el possible ús abusiu del sistema per part d'actors maliciosos, i s'han de prendre mesures per prevenir-los i mitigar-los.

Pla de recuperació i seguretat general. Els sistemes d'IA haurien de tenir garanties que permetin un pla de reserva en cas de problemes.

Això pot significar que els sistemes d'IA canviïn d'un procediment estadístic a un procediment basat en regles, o que demanen un operador humà abans de continuar amb la seva acció.

Cal assegurar-se que el sistema farà el que se suposa que ha de fer sense danyar els éssers vius ni el medi ambient. Això inclou la minimització de conseqüències i errors no desitjats. A més, s'han d'establir processos per aclarir i avaluar els riscos potencials associats a l'ús de sistemes d'IA, en diferents àrees d'aplicació. El nivell de mesures de seguretat requerides depèn de la magnitud del risc que suposa un sistema d'IA, que al seu torn depèn de les capacitats del sistema. Quan es pugui preveure que el procés de desenvolupament o el propi sistema suposaran riscos especialment elevats, és fonamental que les mesures de seguretat es desenvolupin i es posin a prova de manera proactiva.

Precisió. La precisió es refereix a la capacitat d'un sistema d'IA per fer judicis correctes, per exemple, per classificar correctament la informació en les categories adequades, o la seva capacitat per fer prediccions, recomanacions o decisions correctes basades en dades o models. Un procés de desenvolupament i avaluació explícit i ben format pot donar suport, mitigar i corregir els riscos no desitjats de prediccions inexactes. Quan no es poden evitar prediccions inexactes ocasionals, és important que el sistema indiqui la probabilitat d'aquests errors. Un alt nivell de precisió és especialment crucial en situacions en què el sistema d'IA afecta directament les vides humanes.

Fiabilitat i reproductibilitat. És fonamental que els resultats dels sistemes d'IA siguin reproductibles i fiables. Un sistema d'IA fiable és aquell que funciona correctament amb una sèrie d'entrades i en diverses situacions. Això és necessari per examinar un sistema d'IA i evitar danys no desitjats. La reproductibilitat descriu si un experiment d'IA mostra el mateix comportament quan es repeteix en les mateixes condicions. Això permet als científics i als responsables polítics descriure amb precisió què fan els sistemes d'IA. Els fitxers de replicació poden facilitar el procés de prova i reproducció de comportaments.

9.3. Privacitat i govern de les dades

Estretament lligat al principi de prevenció de danys hi ha la privadesa, un dret fonamental especialment afectat pels sistemes d'IA. La prevenció del dany a la privadesa també requereix una governança de dades adequada que cobreixi la qualitat i la integritat de les dades utilitzades, la seva rellevància en funció del domini en què es desplegaran els sistemes d'IA, els seus protocols d'accés i la capacitat de processar dades d'una manera que protegeixi la privadesa.

Privadesa i protecció de dades. Els sistemes d'IA han de garantir la privadesa i la protecció de dades durant tot el cicle de vida d'un sistema.

Això inclou la informació proporcionada inicialment per l'usuari, així com la informació generada sobre l'usuari al llarg de la seva interacció amb el sistema (per exemple, les sortides que el sistema d'IA va generar per a usuaris específics o com els usuaris van respondre a recomanacions particulars). Els registres digitals del comportament humà poden permetre que els sistemes d'IA infereixin no només les preferències de les persones, sinó també la seva orientació sexual, edat, gènere, opinions religioses o polítiques. Per permetre que les persones confiïn en el procés de recollida de dades, s'ha de garantir que les dades recollides sobre ells no s'utilitzaran per discriminar-les de manera il·legal o injusta.

Qualitat i integritat de les dades. La qualitat dels conjunts de dades utilitzats és primordial per al rendiment dels sistemes d'IA. Quan es recullen dades, poden contenir biaixos, inexactituds, errors i errors socials. Això s'ha de resoldre abans de la formació amb qualsevol conjunt de dades donat. A més, s'ha de garantir la integritat de les dades.

Introduir dades malicioses a un sistema d'IA pot canviar el seu comportament, especialment amb els sistemes d'autoaprenentatge.

Els processos i conjunts de dades utilitzats s'han de provar i documentar en cada pas, com ara la planificació, la formació, les proves i el desplegament. Això també s'hauria d'aplicar als sistemes d'IA que no s'han desenvolupat internament, sinó que s'han adquirit en un altre lloc.

Accés a les dades. En qualsevol organització que gestioni dades d'individus (tant si algú és usuari del sistema com si no), s'han d'establir protocols de dades que regulin l'accés a les dades. Aquests protocols haurien d'esbrinar qui pot accedir a les dades i en quines circumstàncies. Només el personal degudament qualificat amb la competència i la necessitat d'accedir a les dades de la persona ha de poder fer-ho.

9.4. Transparència

Aquest requisit està estretament lligat al principi d'explicabilitat i inclou la transparència dels elements rellevants per a un sistema d'IA: les dades, el sistema i els models de negoci.

Traçabilitat. Els conjunts de dades i els processos que donen lloc a la decisió del sistema d'IA, inclosos els de recollida de dades i etiquetatge de dades, així com els algorismes utilitzats, s'han de documentar amb el millor estàndard possible per permetre la traçabilitat i un augment de la transparència. Això també s'aplica a les decisions preses pel sistema d'IA. Això permet identificar els motius pels quals una decisió d'IA va ser errònia i, al seu torn, podria ajudar a prevenir errors futurs.

La traçabilitat facilita l'auditabilitat i l'explicabilitat.

Explicabilitat. L'explicabilitat fa referència a la capacitat d'explicar tant els processos tècnics d'un sistema d'IA com les decisions humanes relacionades (per exemple, les àrees d'aplicació d'un sistema). L'explicabilitat tècnica requereix que les decisions preses per un sistema d'IA puguin ser enteses i rastrejades pels éssers humans. A més, és possible que s'hagin de fer compromisos entre millorar l'explicabilitat d'un sistema (que pot reduir la seva precisió) o augmentar-ne la precisió (a costa de l'explicabilitat). Sempre que un sistema d'IA té un impacte significatiu en la vida de les persones, hauria de ser possible exigir una explicació adequada del procés de presa de decisions del sistema d'IA. Aquesta explicació ha de ser oportuna i adaptada a l'experiència de la part interessada en qüestió (per exemple, un profà, un regulador o un investigador). A més, haurien d'estar disponibles explicacions sobre el grau en què un sistema d'IA influeix i configura el procés de presa de decisions de l'organització, les opcions de disseny del sistema i la raó per implementar-lo (per tant, assegurant la transparència del model de negoci).

Comunicació. Els sistemes d'IA no s'han de representar com a humans davant els usuaris; els humans tenen dret a ser informats que estan interactuant amb un sistema d'IA. Això implica que els sistemes d'IA han de ser identificables com a tals. A més, s'ha de proporcionar l'opció de decidir en contra d'aquesta interacció a favor de la interacció humana on sigui necessari per garantir el compliment dels drets fonamentals. Més enllà d'això, les capacitats i limitacions del sistema d'IA s'han de comunicar als professionals de la IA o als usuaris finals d'una manera adequada al cas d'ús en qüestió. Això podria incloure la comunicació del nivell de precisió del sistema d'IA, així com les seves limitacions.

9.5. Diversitat, no-discriminació i equitat

Per aconseguir una IA fiable, hem de permetre la inclusió i la diversitat al llarg de tot el cicle de vida del sistema d'IA. A més de la consideració i la implicació de totes les parts interessades afectades al llarg del procés, això també implica garantir la igualtat d'accés mitjançant processos de disseny inclusius així com la igualtat de tracte. Aquest requisit està estretament lligat al principi d'equitat.

Evitar prejudicis injustos. Els conjunts de dades utilitzats pels sistemes d'IA (tant per a la formació com per a l'operació) poden patir la inclusió de models històrics inadvertits, incompletitud i mal govern. La continuïtat d'aquests biaixos podria conduir a prejudicis i discriminacions directes i indirectes no intencionats contra determinats grups o persones, que pot agreujar els prejudicis i la marginació. El dany també pot derivar-se de l'explotació intencionada de prejudicis (del consumidor) o de la competència deslleial, com ara l'homogeneïtzació dels preus mitjançant la connivència o un mercat no transparent. Els biaixos identificables i discriminatoris s'han d'eliminar en la fase de recollida sempre que sigui possible. La forma en què es desenvolupen els sistemes d'IA (per exemple, la programació d'algoritmes) també pot patir un biaix injust. Això es podria contrarestar posant en marxa processos de supervisió per analitzar i abordar el propòsit, les limitacions, els requisits i les decisions del sistema d'una manera clara i transparent. A més, la contractació de diferents orígens, cultures i disciplines pot garantir la diversitat d'opinions i s'ha de fomentar.

Accessibilitat i disseny universal. En particular, en els dominis d'empresa a consumidor, els sistemes haurien d'estar centrats en l'usuari i dissenyats de manera que permeti a totes les persones utilitzar productes o serveis d'IA, independentment de la seva edat, gènere, habilitats o característiques. L'accessibilitat a aquesta tecnologia per a les persones amb discapacitat, que estan presents en tots els grups de la societat, és d'especial importància. Els sistemes d'IA no haurien de tenir un enfocament únic i haurien de tenir en compte principis de disseny universal que s'adrecen a la gamma més àmplia possible d'usuaris, seguint els estàndards d'accessibilitat rellevants.

Això permetrà l'accés equitatiu i la participació activa de totes les persones en les activitats humanes actuals i emergents mediades per ordinador i pel que fa a les tecnologies d'assistència.

Participació de les parts interessades. Per tal de desenvolupar sistemes d'IA que siguin fiables, és recomanable consultar les parts interessades que puguin veure's afectades directament o indirectament pel sistema al llarg del seu cicle de vida. És beneficiós demanar comentaris periòdics fins i tot després del desplegament i establir mecanismes a llarg termini per a la participació de les parts interessades, per exemple, assegurant la informació, la consulta i la participació dels treballadors durant tot el procés d'implementació de sistemes d'IA a les organitzacions.

9.6. Benestar social i ambiental

D'acord amb els principis d'equitat i prevenció del dany, la societat en general, altres éssers sensibles i el medi ambient també s'han de considerar com a parts interessades al llarg del cicle de vida del sistema d'IA. S'hauria de fomentar la sostenibilitat i la responsabilitat ecològica dels sistemes d'IA, i s'hauria de fomentar la investigació sobre solucions d'IA que abordin àrees de preocupació global, com ara els Objectius de Desenvolupament Sostenible. Idealment, els sistemes d'IA s'han d'utilitzar per beneficiar tots els éssers humans, incloses les generacions futures.

IA sostenible i respectuosa amb el medi ambient. Els sistemes d'IA prometen ajudar a abordar algunes de les preocupacions socials més urgents, però s'ha de garantir que això es produeixi de la manera més respectuosa amb el medi ambient possible. En aquest sentit, s'hauria d'avaluar el procés de desenvolupament, desplegament i ús del sistema, així com tota la seva cadena de subministrament, per exemple, mitjançant un examen crític de l'ús dels recursos i el consum d'energia durant la formació, optant per opcions menys perjudicials. S'han d'encoratjar les mesures que garanteixin la compatibilitat amb el medi ambient de tota la cadena de subministrament dels sistemes d'IA.

Impacte social. L'exposició omnipresent als sistemes d'IA social en tots els àmbits de la nostra vida (ja sigui en l'educació, el treball, l'atenció o l'entreteniment) pot alterar la nostra concepció de l'agència social o afectar les nostres relacions socials i el nostre vincle. Tot i que els sistemes d'IA es poden utilitzar per millorar les habilitats socials, també poden contribuir al seu deteriorament. Això també pot afectar el benestar físic i mental de les persones. Per tant, els efectes d'aquests sistemes s'han de controlar amb cura.

Societat i democràcia. Més enllà d'avaluar l'impacte del desenvolupament, el desplegament i l'ús d'un sistema d'IA sobre les persones, aquest impacte també s'hauria d'avaluar des d'una perspectiva social, tenint en compte el seu efecte sobre les institucions, la democràcia i la societat en general. S'ha de tenir en compte l'ús de sistemes d'IA, especialment en situacions relacionades amb el procés democràtic, que inclou no només la presa de decisions polítiques sinó també contextos electorals.

9.7. Responsabilitat

El requisit de responsabilitat complementa els requisits anteriors i està estretament lligat al principi d'equitat. Cal que s'estableixin mecanismes per garantir la responsabilitat i la rendició de comptes dels sistemes d'IA i els seus resultats, tant abans com després del seu desenvolupament, desplegament i ús.

Auditabilitat. L'auditabilitat implica l'habilitació de l'avaluació d'algorismes, dades i processos de disseny. Això no implica necessàriament que la informació sobre models de negoci i propietat intel·lectual relacionada amb el sistema d'IA hagi d'estar sempre disponible de manera oberta. L'avaluació per part d'auditors interns i externs, i la disponibilitat d'aquests informes d'avaluació, poden contribuir a la fiabilitat de la tecnologia. En aplicacions que afecten els drets fonamentals, incloses les aplicacions crítiques per a la seguretat, els sistemes d'IA haurien de poder ser auditats de manera independent.

10. Declaracions

Declaració de Barcelona

La Declaració de Barcelona, impulsada per experts en intel·ligència artificial, estableix les bases per al desenvolupament i ús adequat de la IA a Europa. Aquesta iniciativa busca sensibilitzar la societat sobre els beneficis i riscos de la IA, alhora que compromet els dissenyadors, implementadors i usuaris amb principis de prudència, transparència, responsabilitat i fiabilitat. La declaració emfatitza la necessitat d'establir límits ètics i legals clars per a aquesta tecnologia, reconeixent la seva creixent influència en diversos aspectes de la vida quotidiana.

A més, la Declaració de Barcelona advoca per una forta aposta europea en IA, proposant la creació d'una xarxa de laboratoris d'alt nivell, finançament adequat, i la promoció de la formació d'enginyers especialitzats. També destaca la importància de desenvolupar plataformes de recursos oberts per a la investigació. La declaració subratlla la necessitat de regles clares per restringir el comportament dels sistemes d'IA i establir responsabilitats en cas d'errors o fracassos.

https://www.iiia.csic.es/barcelonadeclaration/

Pausa a la investigació d'IA avançada

El març de 2023, desenes de milers d'investigadors, professors i ciutadans, encapçalats pels investigadors **Yoshua Bengio** i **Stuart Russell**, han signat una carta oberta que demana una pausa en l'entrenament de sistemes d'intel·ligència avançada més avançats que GPT-4.

https://futureoflife.org/open-letter/pause-giant-ai-experiments/

La motivació d'aquesta carta, explicada al document, és que els sistemes d'IA amb intel·ligència competitiva amb la humana poden suposar riscos profunds per a la societat i la humanitat. La IA avançada podria representar un canvi profund en la història de la vida a la Terra i s'hauria de planificar i gestionar amb una cura i recursos proporcionats. Malauradament, aquest nivell de planificació i gestió no s'està produint, tot i que els últims mesos s'han vist els laboratoris d'IA bloquejats en una carrera fora de control per desenvolupar i desplegar ments digitals cada cop més potents, i que ningú, ni tan sols els seus creadors, pot entendre, predir o controlar de manera fiable.

A continuació la carta planteja els quatre interrogants següents.

- Hem de deixar que les màquines inundin els nostres canals d'informació amb propaganda i falsedat?
- Hem d'automatitzar totes les feines, fins i tot aquelles que ens siguin satisfactòries?
- Hauríem de desenvolupar ments no humanes que eventualment puguin superar-nos en nombre, ser més intel·ligents, obsoletes i substituir-nos?
- Ens hem d'arriscar a perdre el control de la nostra civilització?

Declaració breu sobre risc existencial de la IA

El maig del 2023, poc després de la carta oberta demanant una pausa en la investigació de sistemes d'IA avançada, es va fer pública la següent declaració sobre el risc de la IA.

> Mitigar el risc d'extinció a causa de la IA hauria de ser una prioritat global al costat d'altres riscos a escala social, com les pandèmies i la guerra nuclear

https://www.safe.ai/work/statement-on-ai-risk

Cimera de Bletchley

La Declaració de Bletchley, signada per diversos països del bloc occidental, entre ells els Estats Units, el Regne Unit i la Unió Europea, en una cimera sobre seguretat de la IA, reconeix tant el potencial transformador com els riscos de la IA. Aquest acord internacional emfatitza la necessitat de col·laboració global per garantir que el desenvolupament de la IA sigui segur i responsable. La declaració destaca la importància de protegir els drets humans i les llibertats fonamentals en l'era de la IA, alhora que promou la innovació i el progrés tecnològic de manera ètica.

https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration

11. EU AI Act

Les directrius europees sobre IA es concreten en la Llei d'Intel·ligència Artificial de la UE (en anglès, **EU AI Act**), que ha entrat en vigor dia 21 de maig del 2024.

La llei d'IA recull una sèrie de pràctiques prohibides, classifica els sistemes d'IA en tres nivells de risc, estableix condicions per als de més risc i contempla mesures d'innovació i governança, a més de proposar un codi de conducta d'aplicació al sector. Les noves normes estableixen obligacions per als proveïdors i usuaris en funció del nivell de risc de la intel·ligència artificial. Tot i que molts sistemes d'IA representen un risc mínim, cal avaluar-los.

Risc inacceptable

Els sistemes d'IA de risc inacceptable són sistemes considerats una amenaça per a les persones i seran prohibits. Inclouen:

Manipulació cognitiva conductual de persones o grups vulnerables específics: per exemple, joguines activades per veu que fomenten comportaments perillosos en els nens.

Scoring social: classificació de persones en funció del comportament, l'estatus socioeconòmic o les característiques personals.

Sistemes d'identificació biomètrica en temps real i remots, com ara el reconeixement facial.

Es poden permetre algunes excepcions: per exemple, els sistemes d'identificació biomètrica remota "post" on la identificació es produeix després d'un retard significatiu es permetran per a perseguir delictes greus, però només després de l'aprovació judicial.

Alt risc

Els sistemes d'IA que afectin negativament la seguretat o els drets fonamentals es consideraran d'alt risc i es dividiran en dues categories:

1) Sistemes d'IA que s'utilitzen en productes inclosos en la legislació de seguretat de productes de la UE. Això inclou joguines, aviació, cotxes, dispositius mèdics i ascensors.

2) Sistemes d'IA en vuit àrees específiques que s'hauran de registrar en una base de dades de la UE:

- Identificació biomètrica i categorització de persones físiques
- Gestió i explotació d'infraestructures crítiques
- Educació i formació professional
- Ocupació, gestió dels treballadors i accés a l'autoocupació
- Accés i gaudi dels serveis privats essencials i dels serveis i prestacions públiques
- Aplicació de la llei
- Gestió de migracions, asil i control de fronteres
- Assistència en la interpretació i aplicació jurídica de la llei

Tots els sistemes d'IA d'alt risc seran avaluats abans de sortir al mercat i també al llarg del seu cicle de vida.

IA generativa

La IA generativa, com per exemple ChatGPT, hauria de complir els requisits de transparència:

- Revelar que el contingut ha estat generat per IA
- Dissenyar el model per evitar que generi contingut il·legal
- Publicació de resums de dades amb drets d'autor utilitzades per a la formació

Risc limitat

Els sistemes d'IA de risc limitat haurien de complir els requisits mínims de transparència que permetrien als usuaris prendre decisions informades. Després d'interaccionar amb les aplicacions, l'usuari pot decidir si vol continuar utilitzant-les. Els usuaris han de ser conscients quan estan interactuant amb la IA. Això inclou sistemes d'IA que generen o manipulen contingut d'imatge, àudio o vídeo, per exemple, *deepfakes*.

12. AESIA

El juny del 2024, poc després de l'entrada en vigor de la Llei d'Intel·ligència Artificial, es va posar en funcionament l'**Agència Espanyola de Supervisió de la Intel·ligència Artificial** (AESIA).

L'AESIA és un organisme públic espanyol responsable de la supervisió, l'assessorament, la conscienciació i la formació dirigides a entitats públiques i privades per a la implementació adequada de la normativa estatal i europea en l'ús i desenvolupament de sistemes d'intel·ligència artificial. L'objectiu darrer d'aquest organisme és minimitzar els riscos que pot suposar l'ús d'aquesta tecnologia, assegurant el desenvolupament i potenciació dels sistemes d'IA de manera que no afectin negativament la integritat, la intimitat, la igualtat de tracte i la no discriminació, especialment entre dones i homes, així com altres drets fonamentals que poden veure's afectats pel mal ús dels sistemes.

La AESIA té la seu a La Corunya (Galícia), i es va establir com a resposta a la necessitat de supervisió i regulació de la IA a Espanya, anticipant-se a l'entrada en vigor del futur reglament europeu que estableix la obligació per als Estats membres de comptar amb una autoritat supervisora en aquesta matèria.

Imatge: *Edifici La Terraza de La Corunya, seu de l'AESIA. Font: wikimedia*

13. Conveni Marc del Consell d'Europa

El Conveni Marc del Consell d'Europa sobre Intel·ligència Artificial i Drets Humans, Democràcia i Estat de Dret, signat a Vilnius el 5 de setembre de 2024, estableix un marc legal comú per regular l'ús de la IA. El seu objectiu principal és garantir que les activitats relacionades amb sistemes d'IA siguin coherents amb els drets humans, la democràcia i l'estat de dret.

El conveni defineix la IA com un sistema basat en màquines que genera prediccions, continguts, recomanacions o decisions que poden influir en entorns físics o virtuals. S'aplica a activitats d'IA realitzades per autoritats públiques, actors privats en nom d'autoritats públiques i, en certa mesura, actors privats. No obstant això, exclou activitats relacionades amb la seguretat nacional o la defensa.

Els principis generals del conveni inclouen la protecció dels drets humans, la integritat dels processos democràtics, el respecte a l'estat de dret, la dignitat humana i l'autonomia individual. També emfatitza la importància de la transparència, la supervisió, la responsabilitat, la igualtat i la no discriminació, així com la privacitat i la protecció de dades personals.

Les parts signants es comprometen a adoptar mesures legislatives, administratives o d'altre tipus per implementar el conveni. Això inclou l'establiment de mecanismes de supervisió independents i la garantia de recursos accessibles i efectius per a violacions de drets humans. També es requereix la implementació d'un marc de gestió de riscos i impactes.

El conveni promou la transparència i la supervisió en les activitats relacionades amb la IA, incloent la identificació de contingut generat per sistemes d'IA. També estableix mesures per garantir la responsabilitat i la rendició de comptes pels impactes adversos en els drets humans, la democràcia i l'estat de dret.

La igualtat i la no discriminació són aspectes fonamentals del conveni. Les parts es comprometen a adoptar mesures per superar les desigualtats i aconseguir resultats justos i equitatius en relació amb les activitats dels sistemes d'IA. També es fa èmfasi en la protecció de la privacitat i les dades personals.

El conveni estableix l'obligació de proporcionar recursos efectius per a les violacions dels drets humans resultants de les activitats dels sistemes d'IA. Això inclou mesures per garantir que la informació rellevant sobre els sistemes d'IA estigui documentada i disponible per a les persones afectades.

Un aspecte important del conveni és el marc de gestió de riscos i impactes. Les parts han d'adoptar mesures per identificar, avaluar, prevenir i mitigar els riscos que plantegen els sistemes d'IA, considerant els impactes reals i potencials sobre els drets humans, la democràcia i l'estat de dret.

El conveni estableix un mecanisme de seguiment a través de la Conferència de les Parts, que supervisa l'aplicació del conveni. Les parts han de presentar informes periòdics sobre les mesures adoptades i es fomenta la cooperació internacional per prevenir i mitigar riscos.

Finalment, el conveni inclou disposicions sobre la seva entrada en vigor, la possibilitat d'esmenes, la resolució de disputes i la possibilitat d'adhesió per a estats no membres del Consell d'Europa. També conté una clàusula federal per a estats federals i la possibilitat de denúncia del conveni.

El newsleetter The Batch en dona notícia a la secció Western Powers Sign AI Treaty.

14. Superintel·ligència i consciència

Hi ha un conjunt de conceptes que es debaten al voltant de la IA i les seves implicacions ètiques. Alguns dels més importants són les següents.

La **superintel·ligència**, expressada pel filòsof **Nick Bostrom**, és una intel·ligència superior a la humana. Presenta el perill d'un creixement cada vegada més ràpid i que, una vegada superada la nostra, planteja problemes difícils pel que fa al control i alineament.

David Chalmers és un filòsof de la ment que ha teoritzat sobre el **problema difícil de la consciència**. Aquest problema fa referència a la dificultat que té l'explicació científica per donar compte de la naturalesa subjectiva de la consciència, és a dir, per què tenim experiències conscients i com sorgeixen aquestes experiències en el nostre cervell.

Segons Chalmers, el problema de la consciència és difícil perquè no es pot reduir a explicacions científiques de caràcter objectiu, sinó que implica una dimensió subjectiva que no es pot reduir a una explicació física. Això és conegut com el "problema difícil" de la consciència.

Degut a la dificultat en la caracterització de la consciència, no és clar que no pugui aparèixer consciència sobre substrats diferents que el cervell humà. Si hi ha una explosió de consciència no humana, apareix el problema de l'experiència del **sofriment artificial** per part d'aquestes noves consciències. Per això, el filòsof **Thomas Metzinger** va proposar una **moratòria** en el desenvolupament de sistemes en què pugui aparèixer consciència fins a tenir una comprensió més clara de la qüestió.

L'historiador **Yuval Noah Harari** clou el seu llibre **Homo Deus** amb aquestes tres qüestions finals.

1. Són els organismes només algorismes, i la vida just processament de dades?
2. Què té més valor: la **intel·ligència** o la **consciència**?
3. Què passarà a la societat, la política i la vida quotidiana quan algorismes no conscients però altament intel·ligents ens coneguin més bé que nosaltres mateixos?

El seu darrer llibre, **Nexus**, es presenta com una història de les xarxes d'informació, des de l'edat de pedra fins a la IA. Acaba amb la reflexió que "hem convocat una intel·ligència inorgànica aliena que podria escapar del nostre control i posar en perill no només la nostra espècie sinó moltes altres formes de vida. Les decisions que tots prenguem en els propers anys determinaran si convocar aquesta intel·ligència aliena és un error terminal o l'inici d'un nou capítol esperançador en l'evolució de la vida".

14.1. Selecció d'opinions

Actualment s'estan definint molts aspectes ètics i legals que determinaran la influència de la intel·ligència artificial en el nostre futur més proper i més llunyà.

Acabam aquest capítol amb una llista dels tecnòlegs i humanistes que estan defensant posicions destacades, amb enllaços a diversos vídeos i textos recents on podeu escoltar i llegir les seves opinions.

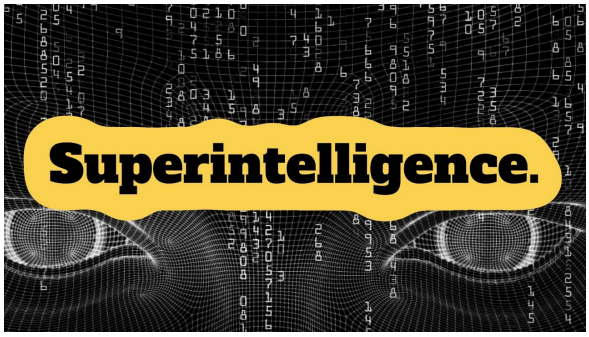
- Stuart Russell

- Steven Pinker

[Will ChatGPT supplant us as writers, thinkers?]

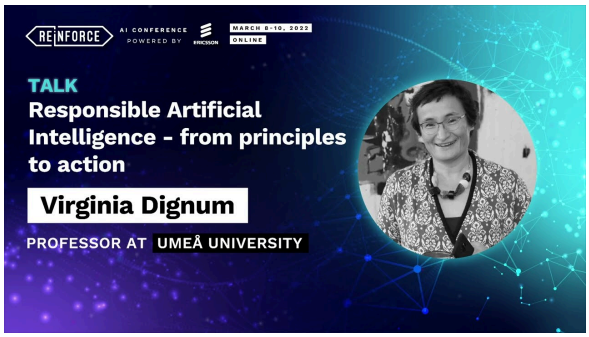
- Yuval Noah Harari

- David Chalmers



- Max Tegmark

- Virginia Dignum



- Thomas Metzinger

- Timnit Gebru

- Gary Marcus

15. Casos d'ètica i regulació

Ramon López de Mántaras té els següents apartats sobre ètica i regulació al seu llibre "100 coses que cal saber sobre intel·ligència artificial".

1. El problema és el Dr. Frankenstein, en què critica tant l'aproximació de dalt a baix com de baix a dalt en l'intent de dotar les màquines de criteri moral.
2. L'efecte retrovisor, en què es mostra com la IA sembla sempre molt més a prop del que realment és; explica la necessitat de dotar de sentit comú els sistemes d'IA; defensa la combinació entre humans i màquines com el millor equip i la necessitat de regular els riscos ja presents de la IA (impacte en l'ocupació, vigilància massiva amb reconeixement facial, manipulació d'opinions, privacitat i armes letals autònomes).
3. Algorismes esbiaixats i gens transparents, discuteix la problemàtica dels conjunts de dades esbiaixats que transmeten la seva manca de neutralitat als sistemes que s'hi entrenen.
4. Dades sintètiques, que explica que els intents d'evitar el biaix generant dades artificials encara pot tenir el biaix de les dades que han entrenat els generadors.
5. Robots amb llicència per matar, en què es fa ressò de la campanya Stop Killer Robots, fins a l'ús de sistemes intel·ligents autònoms en els actuals conflictes d'Ucraïna i Gaza.
6. El complex de Frankenstein, en què repassa les referències literàries i cinematogràfiques de la IA, que també podem consultar al diàleg https://revistaidees.cat/media/dialeg-limaginari-simbolic-de-la-intelligencia-artificial/
7. El cor té raons que la raó desconeix planteja la pregunta si "encara que la TRE (tecnologia de reconeixement d'emocions) pogués detectar amb precisió les nostres emocions, volem una vigilància íntima i permanent de les nostres vides?"
8. IA a la gran pantalla: plantejant qüestions ètiques, sobre la possibilitat qeu les màquines experimentin sentiments humans i que els humans projectem afectivitat cap a les màquines i n'esperem reciprocitat.
9. El dret de desconnectar, que planteja la pregunta de si en un futur es poden desconnectar intel·ligències artificials fortes, és a dir, conscients i sensibles.
10. Consciència i responsabilitat, en què explica el cas de l'exenginyer de Google Blake Lemoine, que va atribuir consciència al xatbot amb què havia estat treballant molt de temps.

16. Bibliografia

Per elaborar aquest capítol s'han emprat els recursos següents.

Llibres

Artificial Intelligence, a Modern Approach, de Peter Norvig i Stuart Russell. Especialment els capítols d'aprenentatge a partir d'exemples

100 coses que cal saber d'intel·ligència artificial, de Ramon López de Mántaras